Analysis of Handwritten Arabic Letters Using Selected Feature Extraction Techniques

GHEITH A. ABANDAH AND MOHAMMED Z. KHEDHER

Faculty of Engineering and Technology, The University of Jordan
Amman 11942, Jordan
abandah@ju.edu.jo, khedher@ju.edu.jo

The Arabic letters are used in many writing languages. However, little work has been done to analyze and characterize handwritten Arabic letters comprehensively. Such characterization is important for the active research in computer processing of Arabic written scripts. We extract carefully selected features from a large database of handwritten Arabic letters. We extract features from the letter's secondary components, main body, skeleton, and boundary. These features are studied and statistically analyzed to reach the targeted characterization. Observations about the important writing style variations are presented and statistically specified. The Arabic letters have multiple forms depending on the letter's position in the word. Comparisons among the four main letter forms (isolated, initial, medial, and final) are also presented.

Keywords: Pattern analysis; pattern characterization; handwritten Arabic letters; feature extraction.

1. Introduction

Arabic letters are used in about 27 writing languages including Arabic, Persian, Kurdish, Urdu, and Jawi [1]. The Arabic writing system flows from right-to-left and is always cursive; both when printed and handwritten. Computer processing of handwritten Arabic scripts includes several fields such as online recognition, offline word recognition, offline character recognition, writer identification and verification, and signature recognition and verification. These fields are active research areas. Example research in these fields are [2, 3, 4, 5, 6], respectively. Researchers dealing with processing of unconstrained handwritten Arabic cursive scripts must overcome many difficulties such as unlimited variation in human handwriting, similarities of distinct character shapes, character overlaps, and interconnections of neighboring characters.

Research in these fields would benefit from thorough description of handwritten Arabic letters, survey of their writing variations, and analysis of their characteristics. Many research papers in these fields have short introductory

sections about the general characteristics of the printed and handwritten Arabic scripts [7, 8, 9, 10]. In the online recognition field, El-Wakil and Shoukry studied the structure of Arabic letters and noticed that every Arabic letter has a main stroke and some letters have dots and secondary stroke [11]. Mezghani *et al.* studied the variations in written Arabic letters [12]. Biadsy *et al.* characterized some aspects of the Arabic script [13].

In the offline character recognition field, Sari *et al.* described the general characteristics of Arabic text and used morphological features of the Arabic letters such as turning points, holes, ascenders, descenders, and dots for segmentation and recognition [14]. Menasri *et al*. identified letter body alphabet for handwritten Arabic letters; they classified Arabic letters into root shapes and optional tails. Multiple Arabic letters that only differ in the existence and number of dots are mapped to the same root shape. This alphabet also includes common vertical ligatures of joined letters [4].

Pechwitz *et al.* have collected a database of handwritten Arabic names for Tunisian towns and published statistics about the size of this database in words, parts of Arabic words (PAWs), and characters [15]. Khedher and Abandah described the main characteristics of the Arabic writing and provided statistics for PAWs and letter forms [16]. Malas *et al.* provided statistics about frequencies of Arabic letters and letter pairs [17].

Some analyses have been done for handwritten scripts of other languages. Nakagawa and Matsumoto analyzed databases of online handwritten Japanese character patterns concentrating on variations in stroke count [18]. Chang and Yan have also analyzed and extracted stroke structures of optically scanned Chinese characters [19]. Deshpande *et al.* described the general features of the Devnagari, the script of the Hindi language; they extracted directional features of Devnagari characters, and represented them in regular expressions for recognition [20].

In this paper, we present a comprehensive analysis and characterization of handwritten Arabic letters. We also describe some important variations encountered in these letters, stressing out those variations that present problems for computer applications. We hope that this characterization would be useful to researchers involved in the various fields of computer processing of these letters. As far as we know, this paper is the only paper dedicated to this subject.

For this characterization, we rely on extracting carefully selected features from a database of 104 handwritten Arabic letter forms. These features are often extracted in Arabic character recognition [21, 9, 22, 10, 23]. The extracted letter features are analyzed to find the characteristics of handwritten Arabic letters.

This paper is organized in 6 sections. Section 2 is an introduction on the Arabic letters. Section 3 describes our experimental setup including the used database of Arabic letter samples and feature extraction and analysis tools. Section 4 describes the feature extraction techniques used in this research. These techniques include extracting features from the secondary components of the letter, the main body of the letter, the letter skeleton, and the letter boundary. Section 5 uses the features extracted from this database to characterize handwritten Arabic letters. Finally, Section 6 states the main conclusions.

2. Overview of Arabic Letters

In this paper, we characterize the Arabic letters that are commonly used in the Arabic language. There are 28 basic letters in the Arabic alphabet. However, in order to accommodate the needs of other languages, additional letters and symbols were added to this alphabet. Table 1 shows this basic alphabet. We have added in this table the **Hamza** character (ء) because this character is often found in the Arabic writing. **Hamza** has several shapes; its shape changes according to its position in the Arabic word and types of short vowels (harakat) present around it [24].

As shown in this table, each letter has multiple forms depending on its position in the word. Each letter is drawn in an isolated form when it is written alone, and is drawn in up to three other forms when it is written connected to other letters in the word. For example, the letter **Ain** has four forms: *isolated* (ع), *initial* (عـ), *medial* (ـعـ), and *final* (ـع). Moreover, letters **Alef**, **Teh**, and **Hamza** have other forms as shown in Table 1. These four forms have similar frequencies in Arabic text: isolated 23.4%, initial 27.8%, medial 21.0%, and final 27.8% [16].

Within a word, every letter can connect from the right with the previous letter. However, there are six letters that do not connect from the left with the next letter (see Table 1). These letters have only the isolated and final forms. When one of these six letters is present in a word, the word is broken into *sub-words*, often called *parts of Arabic word* (PAWs). For example, the word "Arabic" (عربية) has two PAWs: the first PAW consists of initial **Ain** (عـ) and final, left-disconnecting, **Reh** (ـر); and the second PAW consists of initial **Beh** (بـ), medial **Yeh** (ـيـ), and final **Teh Marbuta** (ـة). Note that the letters are usually connected at a certain horizontal level called the *baseline* [25].

The average Arabic word has 4.3 letters and 2.2 PAWs [16]. Figure 1 shows the frequencies of PAWs comprising 1 to 8 letters. The percentage of PAWs consisting of one letter (isolated form) is 45.8%. The PAWs are relatively short; about 90% of PAWs have one to three letters.

Table 1. Arabic Letters and Their Four Forms

No	Letter Name [a]	Isolated Form	Initial Form	Medial Form	Final Form	No	Letter Name	Isolated Form	Initial Form	Medial Form	Final Form
1	Alef [b, c]	ا ى	-	-	ا ى	16	Tah	ط	ط	ـطـ	ـط
2	Beh	ب	بـ	ـبـ	ـب	17	Zah	ظ	ظ	ـظـ	ـظ
3	Teh [d]	ت ة	تـ	ـتـ	ـت ـة	18	Ain	ع	عـ	ـعـ	ـع
4	Theh	ث	ثـ	ـثـ	ـث	19	Ghain	غ	غـ	ـغـ	ـغ
5	Jeem	ج	جـ	ـجـ	ـج	20	Feh	ف	فـ	ـفـ	ـف
6	Hah	ح	حـ	ـحـ	ـح	21	Qaf	ق	قـ	ـقـ	ـق
7	Khah	خ	خـ	ـخـ	ـخ	22	Kaf	ك	كـ	ـكـ	ـك
8	Dal [b]	د	-	-	ـد	23	Lam	ل	لـ	ـلـ	ـل
9	Thal [b]	ذ	-	-	ـذ	24	Meem	م	مـ	ـمـ	ـم
10	Reh [b]	ر	-	-	ـر	25	Noon	ن	نـ	ـنـ	ـن
11	Zain [b]	ز	-	-	ـز	26	Heh	ه	هـ	ـهـ	ـه
12	Seen	س	سـ	ـسـ	ـس	27	Waw [b]	و	-	-	ـو
13	Sheen	ش	شـ	ـشـ	ـش	28	Yeh	ي	يـ	ـيـ	ـي
14	Sad	ص	صـ	ـصـ	ـص	29	Hamza [e]	ء	ئـ	ـئـ	أ
15	Dad	ض	ضـ	ـضـ	ـض						

[a] Letter names are as in the Unicode Standard [25].

[b] Letters that do not connect from the left.

[c] **Alef** has straight forms (ا) and curly forms (ى).

[d] **Teh** has open forms (ت) and closed forms (ة) named **Teh Marbuta**.

[e] In addition to these forms, the **Hamza** has the isolated forms (أ إ ؤ آ) and the final forms (ؤ ئ آ).

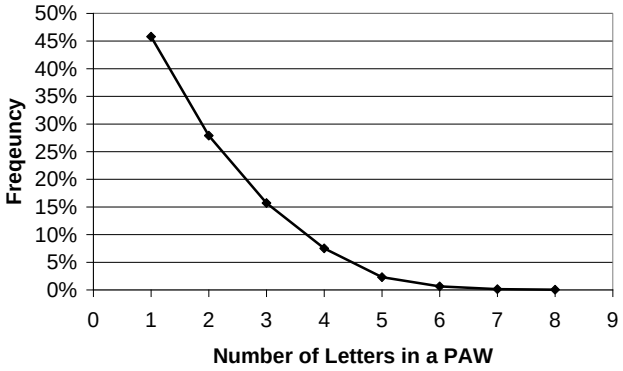


Fig. 1. Frequencies of PAWs as Function of the Number of Comprising Letters

In a typical Arabic text, the frequencies of Arabic letters widely vary. Figure 2 shows the frequencies of 29 Arabic letters [17]. The most frequent three

letters are **Alef** (ا), **Lam** (ل), and **Yeh** (ي) with frequencies of 15.7%, 11.4%, and 8.0%, respectively. The least frequent three letters are **Zah** (ظ), **Ghain** (غ), and **Dad** (ض), with frequencies of 0.2%, 0.5%, and 0.6%, respectively.

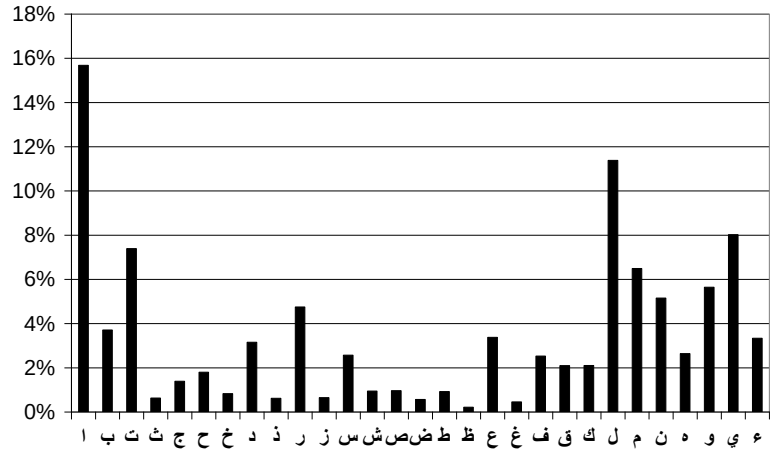


Fig. 2. Frequencies of 29 Arabic Letters

Some letter sequences have special composite *ligatures* when they come in one word. For example, initial **Lam** (ل) followed by final **Alef** (ا) is usually drawn (لا) not (ل ا), and initial **Meem** (م) followed by medial **Hah** (ح) is often drawn (مـحـ) rather than (مح).

In printed Arabic text, the four letter forms usually have fixed shapes irrespective of the surrounding letters. However, in Arabic handwriting, there are slight shape variations for the four letter forms according to the surrounding letters. These variations are usually smaller than the variations present in the written forms between one writer and another.

The Arabic language has some *diacritics* that are used in the holly book Qur'an and sometimes in teaching material and poetry. These diacritics are small markings used above or below the letters of a word to specify the exact pronunciation of the word. They are not commonly used in the daily, scientific, and business uses, and are not discussed further in this paper.

3. Experimental Setup

Our experimental setup comprises a database of handwritten Arabic samples and feature extraction and analysis tools.

3.1. Database of handwritten Arabic samples

Our database of handwritten Arabic samples was collected from 48 persons [16, 27]. These persons were selected to represent various age, gender, and educational background groups. The samples were collected by asking the participants to write, as they normally do write, on a blank paper a one page of cursive Arabic text. This text was carefully selected so that it contains all the letter forms of the 28 Arabic letters. The sample pages were optically scanned with a resolution of 300 dpi.

Although the IFN/ENIT database of handwritten town names is widely used in Arabic OCR research [15], it is not as suitable to our purposes as this database. The IFN/ENIT database does not include some letter forms (e.g. isolated **Ghain**), it has on average about 28 samples per town name (versus 48), and it does not include full sentences and paragraphs.

We have extracted from the 48 page samples about 440 collections of individual words, PAWs, and letter forms. Each collection comprises 48 samples from 48 different persons. Figure 3 shows the collection of 48 samples of the isolated **Ain** form.

Fig. 3. A Collection of 48 Samples of the Isolated **Ain** Form

The collections for initial, medial, and final letter forms were extracted after manually segmenting their cursive PAWs into individual letters. Manual segmentation is used to avoid errors that may come from an automatic letter segmentation process. Automatic segmentation often suffers from over

segmentation, under segmentation, or imprecise segmentation points positioning [14, 28, 29]. We use in this research 104 collections of letter forms: 30 isolated forms, 22 initial forms, 22 medial forms, and 30 final forms. These collections contain all the 28 basic Arabic letters.

3.2. Feature extraction tools

To allow easy extraction of many features from this database of handwritten Arabic samples, we developed a desktop application using Microsoft Visual Studio C++. This application is an expandable tool that allows developers to easily add various preprocessing and feature extraction routines. It enables the user to select the order of the routines to be applied on the sample collections. This application allows the user to visualize the results of preprocessing routines and obtain the results of the feature extraction routines. Figure 4 shows this application with its dialog box for selecting what routines to apply on the collection of the isolated **Ain** samples.

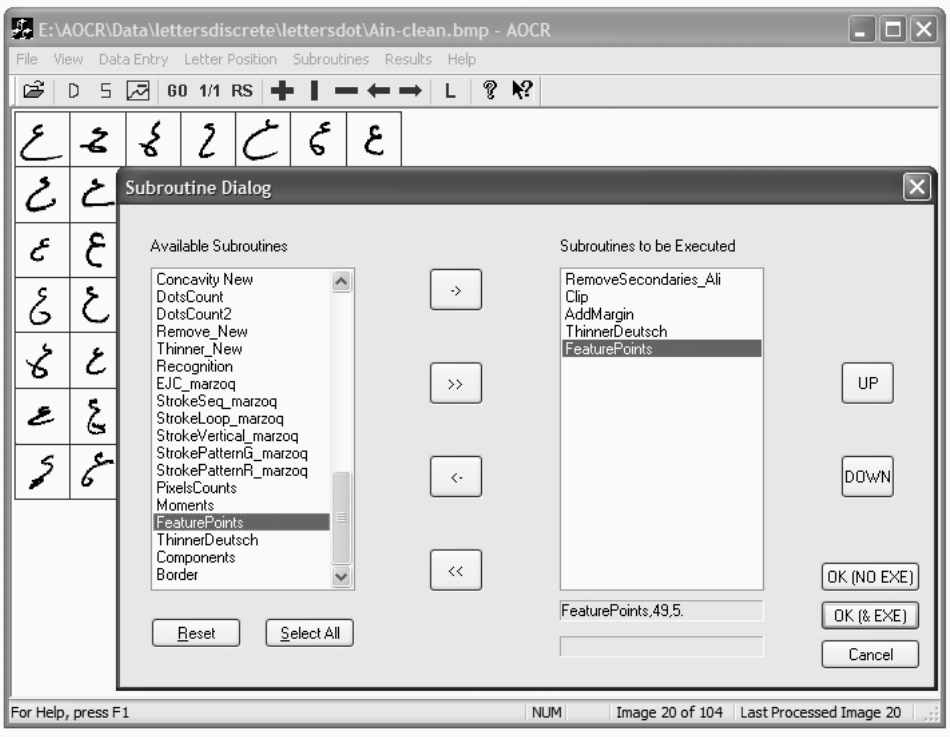


Fig. 4. The Feature Extraction Application and Its Routine Selection Dialog Box

The preprocessing routines implemented in this application include binarization, noise removal, thinning, and boundary finding. This application also features batch processing where the selected routines can be applied on multiple sample collections. The results of the feature extraction routines can be exported from this application into an Excel spreadsheet.

We have implemented in this application feature extraction routines for many features including the selected features described in Section 4. These routines were applied on the 104 collections of letter forms and the results where exported for further analysis as described below.

The feature value x_{ijk}; $i = 1, 2, \ldots, L$; $j = 1, 2, \ldots, M$; and $k = 1, 2, \ldots, N$ is the *kth* feature of the *ith* sample of the *jth* letter form. There are $L = 48$ samples, $M = 104$ different letter forms, and $N = 27$ features. Therefore, the average of the *kth* feature for letter form ω_j is

$$\bar{x}_{jk} = \frac{1}{L}\sum_{i=1}^{L} x_{ijk} \,. \tag{1}$$

The averages shown in some of Section 5's tables are averages of these averages over the four Arabic letter forms. The variance of the *kth* feature for letter form ω_j is

$$s_{jk}^2 = \frac{1}{L-1}\sum_{i=1}^{L} (x_{ijk} - \bar{x}_{jk})^2 \,. \tag{2}$$

In order to characterize the average dispersion of the *kth* feature within every letter form, we calculate the *average coefficient of variance* (C.O.V.) by

$$\text{Average C.O.V.}_{\cdot k} = \frac{1}{M}\sum_{j=1}^{M} \frac{s_{jk}}{\bar{x}_{jk}} \,. \tag{3}$$

For some structural features, e.g., loop existence, we estimate the *hit ratio* of the feature. For a letter form *j* that is normally written with this feature, the hit ratio is the number of samples that *does* have this feature to the number of samples.

$$\text{Hit Ratio}_{jk} = \frac{1}{L}\sum_{i=1}^{L} h_k(x_{ijk}) \,, \tag{4}$$

$$\text{where} \quad h_k(x_{ijk}) = \begin{cases} 1 & x_{ijk} \text{ has feature } k \\ 0 & x_{ijk} \text{ doesn't have feature } k \end{cases} . \tag{5}$$

4. Feature Extraction

The following subsections describe the techniques and algorithms used to extract an assortment of features used to characterize handwritten Arabic letters. We start by detecting the secondary components of the Arabic letters and extracting features from these components. Then we remove the secondary components and extract additional features from the main body, the main body's skeleton, and the main body's boundary.

4.1. Secondary components detection and removal

More than half the Arabic letters are composed of *main body* and *secondary components*. The secondary components are letter components that are disconnected from the main body. For example, **Beh** (ب) has a dot under its main body, **Teh** (ت) has two dots above its main body, and **Kaf** (ك) has a zigzag enclosed within the main body.

Detecting the secondary components can be done after segmenting the binary image of the letter into its disconnected components using the *connected component labeling* techniques [30]. Then the main body is easily identified as it is usually the largest component and is closer to the letter's center than the secondary components. The *secondary position* is then easily found as the position of the secondary components relative to the main body. Finally, the number and position of the secondary components play important role in finding the *secondary type*. However, our approach in classifying the secondary components also utilizes other features extracted from the secondary components such as size, orientation, roundness, and spatial distribution (see Section 4.2).

After detecting and classifying the secondary components, we remove them from the letter image and pass the main body to the other feature extraction stages described below.

4.2. Main body features

Main body features are mainly statistical features. They are found from the letter image after removing the secondary components. Note that the 104 letter forms have only 55 distinct main body shapes: 17 isolated, 11 initial, 11 medial, and 16 final main body shapes. For example, the letter form sets: (خ ح ج), (ئـ يـ نـ ثـ تـ بـ), (ـغـ ـعـ), and (ـض ـص) have same main bodies. The following paragraphs define some main body features: area, width, height, pixel distribution, orientation, roundness, and number of loops.

Size. We use a threshold function to convert the 2-dimentional image into a binary image $B(x,y) \in (0,1)$; black pixels are the foreground pixels and take the value 1 [31]. A low threshold is used to maintain connectivity of light pen strokes. The *area A* of the letter body is found by

$$A = \sum_x \sum_y B(x,y) . \tag{6}$$

To find the main body's *width W* and *height H*, the image is clipped into a rectangular shape such that all four borders have at least one black pixel. We also derive a scale-invariant feature; the *width to height ratio W/H* [32].

Distribution. We partition the clipped image into four equal quadrants and find the fraction of black pixels in each quadrant relative to the area *A*. The resulting four fractions are: upper-right *UR/A*, lower-right *LR/A*, lower-left *LL/A*, and upper-left *UL/A*. We also find the fractions of the four halves relative to *A*: upper *U/A*, right *R/A*, lower *Lo/A*, and left *Lt/A*.

Orientation. The *orientation* θ of an elongated object is the orientation of the elongation axis [31]. The axis of least inertia is the elongation axis. The inertia of the elongation axis is found by

$$\chi^2 = \sum_x \sum_y r^2 B(x,y) , \tag{7}$$

where *r* is the perpendicular distance from point (x, y) to the elongation axis. Using polar coordinates and utilizing the fact that the elongation axis passes through the center of mass, the inertia is found from the second-order central moments by

$$\chi^2 = \frac{1}{2}(\mu_{20} + \mu_{02}) - \frac{1}{2}(\mu_{20} - \mu_{02})\cos 2\theta - \mu_{11} \sin 2\theta . \tag{8}$$

The orientation of the elongation axis can be found by solving the minimization problem of Eq. (8) with respect to θ. The orientation θ then can be found by solving

$$\sin 2\theta = \pm \frac{2\mu_{11}}{\sqrt{4\mu_{11}^2 + (\mu_{20} - \mu_{02})^2}} \text{ and} \tag{9}$$

$$\cos 2\theta = \pm \frac{(\mu_{20} - \mu_{02})}{\sqrt{4\mu_{11}^2 + (\mu_{20} - \mu_{02})^2}} . \tag{10}$$

Roundness. The positive and negative values for sine and cosine of 2θ in Eqs. (9) and (10) can be plugged in Eq. (8) to find the minimum and maximum inertia values, respectively. The object *roundness R*, defined using Eq. (11), is a ratio between 0 for a straight line and 1 for a circle.

$$R = \frac{\chi^2_{\min}}{\chi^2_{\max}} \tag{11}$$

Loops. The number of main body loops is a structural feature. There are many techniques to find the number of loops in an image. We use the connected component labeling algorithm to find the number of loops. The number of background components (white components) minus one is the number of loops. For example, **Sad** (ص) has one loop because it has two background components; the large background component surrounding the letter (always present) and the small component enclosed within the loop in the right.

4.3. Skeleton features

Thinning is usually a pre-processing stage in character recognition where the character image is reduced to a simplified one-pixel wide skeleton. We use Deutsch's thinning algorithm which gives good skeletons for our samples [33]. We use the skeleton of the main letter's body to extract five features: vertical and horizontal crossings and three feature points.

Vertical and horizontal crossings are found by counting the number of white-black-white transfers when scanning the image's pixels on a vertical line and a horizontal line, respectively. These lines are the two lines that pass through the center of mass of the main body's skeleton.

Feature points. Three important feature points can be easily found from the skeleton by examining the eight immediate neighbors of every black pixel: *end point* is a point with one black neighbor, *branch point* has three black neighbors, and *cross point* has four black neighbors.

4.4. Boundary features

Boundary finding is another pre-processing stage in character recognition where the character outer contour is found [34]. We find the boundary of the main letter's body and use it to extract five features: number of boundary pixels, perimeter length, perimeter to diagonal ratio, compactness ratio, and bending energy.

Boundary pixels. The number of *boundary pixels m* is directly found by counting the boundary pixels $(x_i, y_i), i = 1, 2, \ldots, m$. Then Freeman chain code is used to compactly encode the boundary pixels [35]. The direction from every boundary pixel to the next boundary pixel is put in the chain. The direction from the last pixel to the first pixel is the last code in the chain. The direction codes $f_i \in [0, 7]$ are used such that right is 0, up-right is 1, up is 2, *etc*.

Perimeter length. The *perimeter length T* is found by summing the distances from one pixel to the next. Formally, it is found from the chain code using

$$T = \sum_{i=1}^{m} \mathrm{L}(f_i), \quad \text{where} \quad \mathrm{L}(f_i) = \begin{cases} 1 & f_i \text{ is even} \\ \sqrt{2} & f_i \text{ is odd} \end{cases}. \tag{12}$$

Perimeter to diagonal ratio. We also use a scale-invariant feature which is the ratio of half the perimeter length to the diagonal of the clipped main body rectangle *T/2D*. For simple shapes like **Alef** (ا), this ratio is 1, and this ratio is larger than 1 for more complex shapes.

$$T/2D = \frac{T/2}{\sqrt{W^2 + H^2}} \tag{13}$$

Compactness ratio. Another derived feature from the perimeter length and the area is the *compactness ratio* or roundness ratio which is found by Eq. (14) [36].

$$\gamma = \frac{T^2}{4\pi A} \tag{14}$$

This ratio is 1 for a filled circle and is larger than 1 for distributed complex shapes.

Bending energy. The *bending energy E* is a measure of the curvature of the boundary [36]. It can be found from the chain code by summing the squares of the direction changes from one boundary pixel to the next.

$$E = \frac{1}{T} \sum_{i=1}^{m} \left(\frac{\pi}{4} \times \mathrm{IF}(k_i > 4, 8 - k_i, k_i) \right)^2, \tag{15}$$

$$\text{where} \quad k_i = \begin{cases} \mathrm{mod}(f_{i+1} - f_i, 8) & i < m \\ \mathrm{mod}(f_1 - f_m, 8) & i = m \end{cases}. \tag{16}$$

5. Characteristics of Handwritten Arabic Letters

The following subsections present characteristics of Arabic letters and some observations. These characteristics are found by analyzing the features extracted from the 104 collections of letter forms. We concentrate on the characteristics differences among the four letter forms.

5.1. Secondary components characteristics

Tables 2 and 3 list the secondary components types and positions that we encountered in the written Arabic samples.

Table 2. Types of the Secondary Components

No	Secondary Type	Examples	Average Hit Ratio
1	No Secondary	ء و ه م ل ع ط ص س ر د ح ى ا	99.7%
2	One Dot	ن ف غ ظ ض ز ذ خ ج ب	87.2%
3	Two Dots	ي ق ة ت	90.4%
4	Three Dots	ش ث	87.0%
5	Zigzag	ك	40.6%
6	Vertical Bar [a]	ط	35.4%
7	Vertical bar and a dot [a]	ظ	18.1%
8	Long Stroke [b]	ک	25.0%

[a] This secondary is encountered when the upper vertical stroke is drawn disconnected from the loop of **Tah** and **Zah**.

[b] This secondary is encountered when the upper stroke is drawn disconnected from the lower part of initial **Kaf**.

Table 3. Possible Positions of Secondary Components

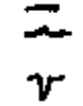
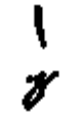
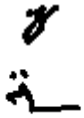
No	Secondary Position	Examples	Average Hit Ratio
1	No Secondary	ء و ه م ل ع ط ص س ر د ح ى ا	99.7%
2	Above	ن ق ف غ ض ش ز ذ خ ث ة ت	96.0%
3	Within	ك ظ ج	84.4%
4	Below	ي ب	97.1%

The type and position of the secondary components are very important features of Arabic letters. For example, recognizing two dots below the main body are sufficient to recognize the letter **Yeh** (ي) because **Yeh** is the only letter that has two dots below the main body. Furthermore, some letters can only be distinguished by their secondary components. For example, **Teh** (ت) and

Theh (ﺙ) differ only by the number of dots above the main body, and medial **Teh** (ﺗ) and medial **Yeh** (ﻳ) differ only by the position of the two dots.

There are important variations in drawing the secondary components; mostly in drawing two dots and three dots. As shown in Table 4—Samples A1, A2, and A3, the two dots come in three variations: two disconnected dots, two connected dots, and horizontal dash. Samples A5, A6, and A7 show three variations in drawing the three dots: three disconnected dots, one dot above horizontal dash, and hat shape "^". Any secondary components classification process should take these variations into consideration [37].

Table 4. Samples Showing Variations in Handwritten Letters

	1	2	3	4	5	6	7	8	9	10
A										
B										
C										
D										
E										
F										

It is important to note that some writers use styles that replace the secondary components of isolated and final forms with main body curves. Table 4 shows some examples: Samples A9 and A10 show how the two dots of isolated **Qaf** are replaced, Samples B1 and B2 show how the one dot of isolated **Noon** is replaced, and Samples B4 and B5 show how the zigzag of final **Kaf** is replaced.

One difficulty in recognizing the secondary components comes when hasty writers draw them connected to the main body. For example, Sample B7 shows the zigzag connected to **Kaf**'s body, Sample B8 shows the two dots connected to **Teh**'s body, Sample B9 shows the three dots connected to **Theh**'s body, and Sample B10 shows the dot connected to **Jeem**'s body.

Tables 2 and 3 also show the average hit ratios for every secondary type and secondary position. These averages are taken over all letter forms that have the corresponding secondary type or position. The high hit ratio for the type "No Secondary" (99.7%) indicates that this feature is stable against writing variations. However, the hit ratios of the dots features are lower due to writing style

variations (varying sizes of dots and dots replaced by body curves) and bad writing (secondaries touching the main body). The hit ratio of the zigzag feature is only 40.6% because most writers draw it as in Sample B5. The secondary strokes: vertical bars in letters **Tah** (ط) and **Zah** (ظ) and long upper stroke in initial and medial **Kaf** (ک) are found disconnected from the main body in 35.4%, 18.1%, and 25.0% of the relevant samples, respectively.

The high hit ratios in Table 3 compared with the hit ratios of Table 2 indicates that the secondary position features are more stable against writing variations. Once a secondary feature is present, there is little variation in its position. This observation is also supported by the C.O.V. averages. The secondary type features have an average C.O.V. of 0.63 and the secondary position features have an average of 0.08. In other words, the dispersion of the secondary type features within each letter form is larger than the dispersion of the secondary position features.

5.2. Main body characteristics

Table 5 shows the averages of the statistical main body features. These averages are found for the features extracted from the four letter forms. The averages in the first three rows indicate that final and isolated forms are larger than initial and medial forms. Samples C1 and C2 of Table 4 show two extremes; the final **Kaf** is much larger than the initial **Feh**. Moreover, Samples C2 and C3 show that the initial and final forms of **Feh** have totally different sizes.

Table 5. Average Values of Some Statistical Features for the Four Letter Forms

No	Feature	Isolated	Initial	Medial	Final	Avg. C.O.V.
1	Area *A* (in pixels)	731	494	556	764	0.23
2	Width *W* (in pixels)	52	40	49	59	0.22
3	Height *H* (in pixels)	42	30	29	39	0.22
4	Ratio *W/H*	1.40	1.51	2.09	1.75	0.29
5	*UR/A*	0.28	0.31	0.22	0.23	0.42
6	*LR/A*	0.24	0.26	0.29	0.24	0.41
7	*LL/A*	0.33	0.32	0.32	0.34	0.19
8	*UL/A*	0.15	0.11	0.17	0.19	0.75
9	*U/A*	0.43	0.43	0.39	0.42	0.21
10	*R/A*	0.52	0.57	0.51	0.47	0.16
11	*Lo/A*	0.57	0.57	0.61	0.58	0.15
12	*Lt/A*	0.48	0.43	0.49	0.53	0.16
13	Orientation θ	37°	34°	22°	27°	0.17
14	Roundness *R*	0.24	0.23	0.25	0.22	0.62

From the width and height averages, we can conclude that Arabic letters are generally elongated in the horizontal direction. Also note that the ratio *W/H* of medial and final forms is larger than that of isolated and initial forms. Sample C5 shows isolated **Alef**, which has small *W/H* ratio. And Sample C6 shows the medial **Seen**, which has large *W/H* ratio. Figure 5 shows the scatter chart of the average widths and heights for the 104 letter forms.

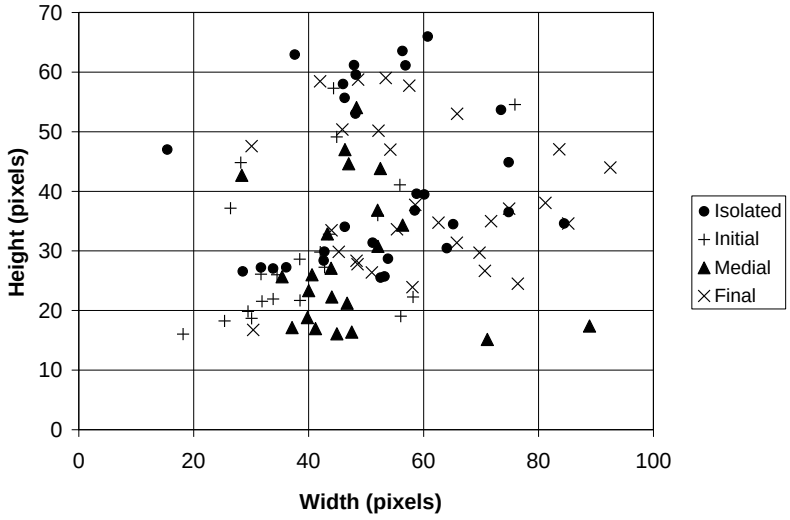


Fig. 5. Scatter Chart of the Letter Forms Sizes

By studying the averages of pixel distribution fractions, we can reach some interesting conclusions about the characteristics of handwritten Arabic letters. In general, Arabic letters have more mass in the lower half of the clipped letter image. However, on average initial forms have more mass in the right half, and final forms have more mass in the left half. Sample C8 shows initial **Yeh** that demonstrates an example of large relative mass in the right half, and Sample C9 shows final **Alef** that demonstrates an example of large relative mass in the left half. Both these samples have most of their respective masses in the lower half. Moreover, the C.O.V. averages indicate that the dispersion within every letter form of the four quadrants is larger than that of the four halves.

In general, the Arabic letters go from right to left and up to down. The average orientation is 30°. However, the four forms have different orientation averages. The medial form's average is the closest to the horizontal direction and the isolated form's average is the farthest. Sample D1 shows medial **Teh** which has a small orientation angle and Sample D2 shows isolated **Alef** which has a

large orientation angle. Figure 6 shows the distribution curves for the orientation and other selected features.

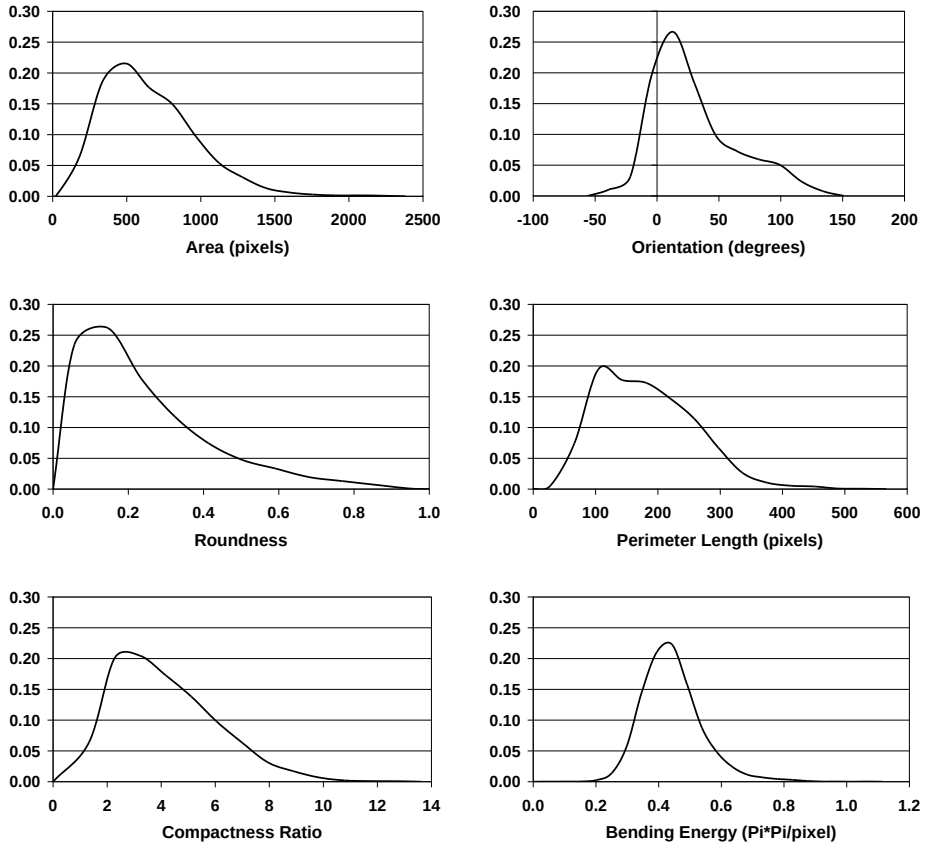


Fig. 6. Distributions of Selected Features

The last row of Table 5 indicates that the average Arabic letter is far from the rounded shape. However, Sample D4 shows the isolated **Teh Marbuta** (closed form) which is the closest form to perfect circle. On the other hand, Sample D5 shows isolated **Reh** which is almost a straight line.

While more than half the Arabic letters are usually written without loops (see Table 6), ten other letters are usually written with one loop in all four forms, two letters are written with one loop in the medial and final forms only, and three letters are written with or without a loop according to the writing style. For example, isolated **Jeem** (ج) is written without a loop and with a loop as shown in Samples D7 and D8, respectively. The hit ratio of finding a loop in the samples of the three letters is 36.3%.

Table 6. Existence of Loops in Arabic Letters

No	Loop Existence	Examples	Average Hit Ratio
1	No loops	ء ي ن ل ك ش س ز ر ذ د ث ت ب ى ا	95.8%
2	One loop in all forms	و ه م ق ف ظ ط ض ص ة	62.5%
3	One loop in some forms	غ ع	41.7%
4	One loop in some styles	خ ح ج	36.3%

Medial **Heh** has large style variation; Samples E1, E2, and E3 show that this form has styles with no loops, one loop, and two loops, respectively. Moreover, some writing styles introduce additional loops to the isolated and final forms by extending the curve of the letter's end. Examples are Letters **Beh** (Sample E5), **Teh** (ت), **Theh** (E6), **Ain** (E7), **Ghain** (غ), **Feh** (E8), **Qaf** (ق), **Kaf** (B5), **Noon** (ن), curly **Alef** (ى), and **Yeh** (E9). Some writers don't close the loop of the final forms of closed **Teh** (ة) and **Heh** (ه), as illustrated in Samples F1–F3.

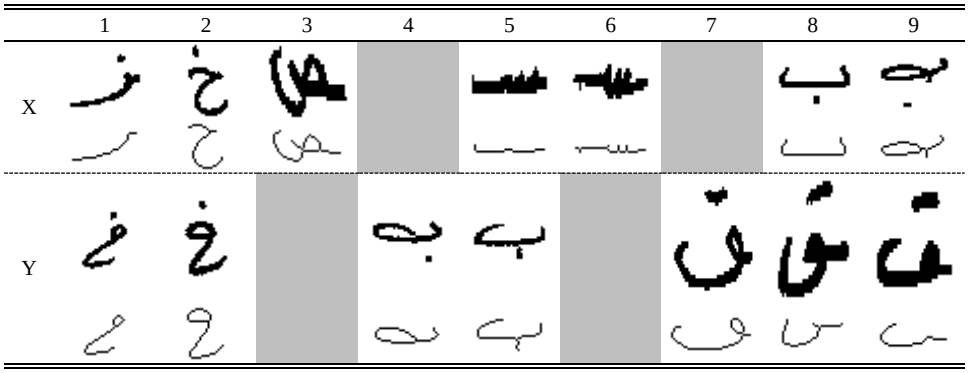
We have noticed that some samples of the isolated and final forms of the letters that have a rounded cusp have unexpected loops when the cusp is drawn completely closed. We have noticed this observation with some samples of Letters **Seen** (س), **Sheen** (ش), **Sad** (ص), **Dad** (see Sample F5), and **Noon** (ن). Also we have noticed that many samples of letters that have a small loop are drawn with a filled loop that is hard to discover. This was frequently noticed with samples of Letters **Feh** (ف), **Qaf** (ق), **Meem** (م), and **Waw** (و). Samples F7 and F8 show how the **Waw** loop is drawn punctured and filled, respectively. Note also that Sample E8 shows final **Feh** drawn with a filled loop.

All these style variations give relatively low loop feature hit ratio as shown in Table 6. Also the average C.O.V. of the loop feature is high (1.84).

5.3. Skeleton characteristics

Table 7 shows some sample letters and the respective main body skeletons. The vertical and horizontal crossings are measures of the letter's complexity. For example, Samples X1, X2, and X3 in Table 7 show the simple final **Zain** that has one vertical and one horizontal crossing, isolated **Khah** that has three vertical crossings and one horizontal crossing, and the complex final **Sad** that has two vertical crossings and four horizontal crossings.

Table 7. Letter Samples and Respective Skeletons



Elongated letters have large variance in the number of crossings in the elongation direction. For example, Samples X5 and X6 of the medial **Seen**, which is horizontally elongated, have one vertical crossing and two and five horizontal crossings, respectively. These two samples illustrate another problem; **Seen** has three small teeth that are often lost through the thinning process.

Decorative loops in the isolated and final forms increase the number of crossings. Samples X8 and X9 illustrate that the vertical crossings of isolated **Beh** increase from one to two when this letter is written with a decorative loop. Also handwriting variations introduce variance in the number of crossings. Samples Y1 and Y2 show two more samples of the isolated **Khah**; Sample Y1 has two vertical crossings because it is written with the loop shifted to the back, and Sample Y2 has four vertical crossings because it is written with the loop hanging to the front.

The number of feature points is affected when decorative loops are added to the isolated and final forms. Although isolated **Beh** has only two end points as illustrated by Sample X8, adding a decorative loop adds a cross point, or eliminates an end point and adds a branch point as illustrated by Samples X9 and Y4, respectively.

The number of feature points is also affected when the secondary objects touch the main body. Sample Y5 shows an isolated **Beh** with its dot touching the main body. As a result, the main body of isolated **Beh** gets one more end point and one branch point.

Variations in drawing loops also affect the number of feature points. Samples Y7 and Y8 show two final **Qaf** letters with punctured and filled loops, respectively. The punctured loop feature gives one cross point, whereas the filled loop gives one branch point and one end point. However, the thinning process

may dissolve the filled loop completely and end up with no feature points as illustrated in Sample Y9.

Moreover, the thinning process may remove the teeth of **Seen** (س), **Sheen** (ش), **Sad** (ص), and **Dad** (ض), as illustrated in Sample X5. The removal of every tooth eliminates one branch point and one end point.

Table 8 shows the averages of the features extracted from the skeleton for the four letter forms. The averages of the medial and final forms are larger than the averages of the isolated and initial forms, which is an indication that medial and final forms are more complex. Note that the averages of the number of end points is around two or larger. Simple letters have two ends unless one end is a loop as in isolated **Waw** (و). The complex forms have more end points, branch points, and cross points.

Table 8. Average Values of the Skeleton Features for the Four Letter Forms

No	Feature	Isolated	Initial	Medial	Final	Avg. C.O.V.
1	Vertical Crossings	1.66	1.55	1.68	1.58	0.33
2	Horizontal Crossings	1.75	1.59	1.84	1.91	0.29
3	End Points	1.96	2.00	2.47	2.41	0.26
4	Branch Points	0.71	0.88	1.20	0.99	1.50
5	Cross Points	0.06	0.05	0.08	0.07	2.82

We noticed that the number of cross points is smaller than the expected number. For example, we expected that the cross point feature would be found in 6 medial forms out of 23 (averaging 0.26). But the extracted average was only 0.08. The reason is that cross points are often lost through the thinning process and are converted to pairs of neighboring branch points as Fig. 7 illustrates. Here the main body of medial **Ain** has one perceptible cross point at the base of the loop. But the thinning process converts the cross into two pixels that are two adjacent branch points.

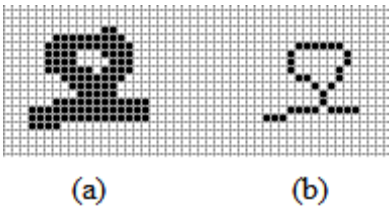


Fig. 7. Medial **Ain**: (a) Main Body, (b) Skeleton after Thinning

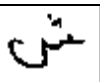
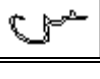
These writing variations and thinning process conversions yield high dispersion for the cross and branch points features within the various letter forms.

This is confirmed by the high the C.O.V. averages for branch and cross points shown in Table 8.

5.4. Boundary characteristics

Table 9 shows some sample letters and the respective main body boundary. Sample Z1 shows isolated **Reh** that has small *T/2D* ratio and Sample Z2 shows final **Khah** that has large *T/2D* ratio. Samples Z4 and Z5 show isolated **Teh** and final **Sheen**, which are two extreme examples of small and large compactness ratios, respectively.

Table 9. Letter Samples and Respective Boundaries

	1	2	3	4	5	6	7	8	9
Z									
									

Small rounded shapes tend to have large bending energy factor. One example is the initial **Feh** shown in Sample Z7. As the isolated **Ain** shown in Sample Z8 has rounded and coarse boundary, it also has a relatively large bending energy. The isolated **Lam** shown in Sample Z9 is an example large letter that has smooth boundary and low pending energy.

Table 10 shows the averages of the five features extracted from the boundary for the four letter forms. The averages of the number of boundary pixels and the perimeter length indicate that the final and isolated forms are larger than medial and initial forms.

Table 10. Average Values of the Boundary Features for the Four Letter Forms

No	Feature	Isolated	Initial	Medial	Final	Avg. C.O.V.
1	Boundary Pixels	177	115	135	194	0.22
2	Perimeter Length	203	130	152	221	0.21
3	Perimeter to Diagonal Ratio	1.5	1.2	1.3	1.5	0.10
4	Compactness Ratio	4.6	2.9	3.4	5.2	0.26
5	Bending Energy	0.41	0.47	0.49	0.42	0.17

The averages of perimeter to diagonal ratio and compactness ratio indicate that the final and isolated forms are more complex and spread than the medial and initial forms. Finally, the averages of the bending energy indicate that the

medial and initial forms have slightly more curly boundaries than the final and isolated forms.

As indicated by the low C.O.V. averages shown in Table 10, the boundary features have small dispersions within the 104 letter forms. The perimeter to diagonal ratio has the smallest average dispersion among these features.

6. Conclusions

This paper uses selected feature extraction techniques to characterize handwritten Arabic letters. The Arabic letters have up to four forms depending on the letter's position in the word: isolated, initial, medial, and final. More than half of these letters have secondary components. The type and position of these components are important features. However, there are variations in drawing some secondary components and some writers often replace them in isolated and final forms with main body curves, or hastily draw them connected to the main body.

Final and isolated forms are generally larger and less compact than initial and medial forms, whereas medial and final forms are the most complex. Arabic letters in general have more mass in the lower half, and initial forms have more mass in the right half, while final forms have more mass in the left half. The average letter orientation is 30° where the medial form's average is the closest to the horizontal direction and the isolated form's average is the farthest. Although several letters are formally written with loops, some small loops are hard to discover when drawn filled, and some writers add decorative loops to the isolated and final forms.

There are high dispersions within the samples of each letter form in the features extracted from the main body's skeleton. This dispersion is due to variations in writing styles and shape conversions done by the thinning process. On the other hand, features extracted from the boundary have low dispersions.

Acknowledgments

The authors wish to thank the students who have built some of the tools used in this work, especially Thamer Abu-Yasin, Nijad Anabtawi, Rola Alsa'feen, Maha Nasser, Ali Alrefaee, Ahmed M. Alghoul, and Yousef Qasim.

References

[1] R. Gordon, editor, *Ethnologue: Languages of the World*, 15th ed. (SIL International, Dallas, 2005).

[2] N. Mezghani, A. Mitiche, and M. Cheriet, Classification of online Arabic characters by Gibbs modeling of class conditional densities, *IEEE Trans. on Pattern Analysis and Machine Intelligence* **30**(7) (2008) 1121–1131.

[3] R. Al-Hajj, C. Mokbel, and L. Likforman-Sulem, Combination of HMM-based classifiers for the recognition of Arabic handwritten words, in *Proc. 9th Int'l Conf. on Document Analysis and Recognition (ICDAR 2007)*, vol. 2, 2007, pp. 959–963.

[4] F. Menasri, N. Vincent, M. Cheriet, and E. Augustin, Shape-based alphabet for off-line Arabic handwriting recognition, in *Proc. 9th Int'l Conf. on Document Analysis and Recognition (ICDAR 2007)*, vol. 2, 2007, pp. 969–973.

[5] M. Bulacu, L. Schomaker, and A. Brink, Text-independent writer identification and verification on offline Arabic handwriting, in *Proc. 9th Int'l Conf. on Document Analysis and Recognition (ICDAR 2007)*, vol. 2, 2007, pp. 769–773.

[6] M. A. Ismail and S. Gad, Off-line Arabic signature recognition and verification, *Pattern Recognition* **33**(10) (2000) 1727–1740.

[7] A. Amin, H. Al-Sadoun, and S. Fischer, Hand-printed Arabic character recognition system using an artificial network, *Pattern Recognition* **29**(4) (1996) 663–675.

[8] M. Khorsheed, Off-line Arabic character recognition: A review, *Pattern Analysis & Applications* **5**(1) (2002) 31–45.

[9] R. Safabakhsh and P. Adibi, Nastaaligh handwritten word recognition using a continuous-density variable-duration HMM, *The Arabian J. Science and Eng.* **30**(1B) (2005) 95–118.

[10] L. Lorigo and V. Govindaraju, Offline Arabic handwriting recognition: A survey, *IEEE Trans. Pattern Analysis & Machine Intelligence* **28**(5) (2006) 712–724.

[11] M. El-Wakil and A. Shoukry, On-line recognition of handwritten isolated Arabic characters, *Pattern Recognition* **22**(2) (1989) 97–105.

[12] N. Mezghani, A. Mitiche, and M. Cheriet, On-line recognition of hand-written Arabic characters using a Kohonen neural network, in *Proc. 8th Int'l Workshop on Frontiers in Handwriting Recognition*, 2002, pp. 490–495.

[13] F. Biadsy, J. El-Sana, and N. Habash, Online Arabic handwriting recognition using hidden Markov models, in *Proc. 10th Workshop on Frontiers of Handwriting Recognition*, 2006.

[14] T. Sari, L. Souici, and M. Sellami, Off-Line handwritten Arabic character segmentation algorithm: ACSA, in *Proc. 8th Int'l Workshop Frontiers in Handwriting Recognition*, 2002, pp. 452–457.

[15] M. Pechwitz, S. Snoussi Maddouri, V. Märgner, N. Ellouze, and H. Amiri, IFN/ENIT–Database of handwritten Arabic words, in *Proc. 7th Collque Int'l Francophone sur l'Ecrit et le Document (CIFED 2002)*, 2002, pp. 129–136.

[16] M. Khedher and G. Abandah, Arabic character recognition using approximate stroke sequence, in Proc. Workshop Arabic Language Resources and Evaluation: Status and Prospects at 3rd Int'l Conf. on Language Resources and Evaluation (LREC 2002), 2002.

[17] T. Malas, S. Taifour, and G. Abandah, Toward optimal Arabic keyboard layout using genetic algorithm, in *Proc. 9th Int'l Middle Eastern Multiconf. on Simulation and Modeling (MESM 2008)*, 2008, pp. 50–54.

[18] M. Nakagawa and K. Matsumoto, Collection of on-line handwritten Japanese character pattern databases and their analyses, *Int'l J. on Document Analysis and Recognition (IJDAR)* **7**(1) (2004) 69–81.

[19] H.-H. Chang and H. Yan, Analysis of stroke structures of handwritten Chinese characters, *IEEE Trans. on Systems, Man, and Cybernetics*, Part B **29**(1) (1999) 47–61.

[20] P. Deshpande, L. Malik, and S. Arora, Fine classification & recognition of hand written Devnagari characters with regular expressions & minimum edit distance method, *J. of Computers* **3**(5) (2008) 11–17.

[21] A. Amin, Arabic character recognition, in *Handbook of Character Recognition and Document Image Analysis*, ed. H. Bunke and P. Wang, (World Scientific, 1997), pp. 397–420.

[22] R. El-Hajj, L. Likforman-Sulem, and C. Mokbel, Arabic handwriting recognition using baseline dependant features and hidden Markov modeling, in *Proc. Int'l Conf. Document Analysis and Recognition (ICDAR'05)*, 2005, pp. 893–897.

[23] H. El-Abed and V. Märgner, Comparison of different preprocessing and feature extraction methods for offline recognition of handwritten Arabic words, in *Proc. Int'l Conf. Document Analysis and Recognition (ICDAR'07)*, 2007, pp. 974–978.

[24] G. Abandah and F. Khundakjie, Issues concerning code system for Arabic letters, *Dirasat Engineering Sciences J.* **31**(1) (2004) 165–177.

[25] M. Pechwitz and V. Märgner, Baseline estimation for Arabic handwritten words, in *Proc. 8th Int'l Workshop Frontiers in Handwriting Recognition*, 2002, pp. 479–484.

[26] The Unicode Consortium, *The Unicode 5.0 Standard*, 5th ed. (Addison-Wesley, Reading, MA, 2006).

[27] G. Abandah and M. Khedher, Printed and Handwritten Arabic Optical Character Recognition: Initial Study, A report on research supported by the Higher Council of Science and Technology, Jordan, Aug. 2004.

[28] L. Lorigo and V. Govindaraju, Segmentation and pre-recognition of Arabic handwriting, in *Proc. Int'l Conf. Document Analysis and Recognition (ICDAR 2005)*, 2005, pp. 605–609.

[29] R. Bentrcia and A. Elnagar, Handwriting segmentation of Arabic text, in *Proc. 5th IASTED Int'l Conf. on Signal Process, Pattern Recognition & Applications (SPPRA 2008)*, 2008, pp. 122–127.

[30] A. Rosenfeld and A. Kak, *Digital Picture Processing* (Academic Press, New York, 1976).

[31] R. Jain, R. Kasturi, and B. Schunck, *Machine Vision* (MacGraw-Hill, Inc., New York, 1995).

[32] O. Trier, A. Jain, and T. Taxt, Feature extraction methods for character recognition: A survey, *Pattern Recognition* **29**(4) (1996) 641–662.

[33] E. Deutsch, Thinning algorithms on rectangular, hexagonal, and triangular arrays, *Comm. of the ACM* **15**(9) (1972) 827–837.

[34] T. Ha and H. Bunke, Image processing methods for document image analysis, in *Handbook of Character Recognition and Document Image Analysis*, ed. H. Bunke and P. Wang, (World Scientific, 1997), pp. 1–47.

[35] H. Freeman, On the encoding of arbitrary geometric configurations, *IRE Trans. Electronic Computers* **10**(2) (1961) 260–268.

[36] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, 3rd ed. (Academic Press, 2006).

[37] M. Khedher and G. Al-Talib, Recognition of secondary characters in handwritten Arabic using fuzzy logic, in *Proc. Int'l Conf. on Machine Intelligence (ICMI'05)*, 2005.