A PREPROCESSING MODEL FOR HAND-WRITTEN ARABIC TEXTS BASED ON VORONOI DIAGRAMS

Atallah M. Al-Shatnawi

Department of Information Systems, Al al-Bayt University, Mafraq, Jordan

ABSTRACT

In this paper, a preprocessing model for hand-written Arabic text on the basis of the Voronoi Diagrams (VDs) is presented and discussed. The proposed VD-based pre-processing model consists of five stages: a preparatory stage, page segmentation, thinning, baseline estimation, and slanting correction. In the preparatory stage, the text image is converted via VDs into a group of geometrical forms that consist of edges and vertices that are used to create the other stages of the proposed model. This stage consists of four main processes: binarization, edge extraction and contour tracking, sampling, and point-VD construction. The second stage is the page segmentation stage based on the VD area. In the third stage, an efficient method for text structuring (that is, thinning) is presented. In the fourth stage, a novel baseline based VD method is presented. In the fifth stage, an efficient technique for slanting detection and correction is proposed and discussed.

KEYWORDS

Preprocessing; Arabic Text Recognition; Voronoi Diagram; Page Segmentation; Thinning; Baseline Detection; Slanting Correction.

1. INTRODUCTION

The eventual goal of any Arabic Character Recognition (ACR) scheme is to emulate the human understanding abilities so that the computer machine will be able to read, revise, understand and perform similar activities to those which the human mind performs with the Arabic written text. In the pattern recognition area, language recognition is regarded as one of the important sophisticated problems in Artificial Intelligence (AI). ACR can be performed in either of two methods. One handles image of the text after it has been input to the computer machine by scanning, for example. This approach is referred to as offline recognition. The second method has a different input approach where the writer directly types to the system, for instance using a light pen as the device of input. This recognition method is described as online recognition. This is quite often easier to deal with than the previous problem owing to that more information is available in the former than the latter case. For instance, motion of the pen can be utilized as a feature of the text [1][5][7][28].

In comparison with the Latin, Japanese, and Chinese character recognition systems, development of the ACR systems did not receive adequate attention of researchers [5]. Whereas Latin character recognition started in 1940 [11], it was only in 1975 when the first attempt was made to recognize the Arabic characters [23]. The Arabic language is a universal language and is the formal language of 25 countries and more than 300 million capita worldwide. Moreover, numerous Arabic characters are employed in many languages like the Ardu, Iranian, Kordi, and Jawi languages [1][2][7][35]. The typical ACR system comprises five major processes: image acquisition, preprocessing, segmentation, classification (i.e., recognition), and feature extraction.

DOI:10.5121/ijcsit.2015.7601

Successive implementation of these five processes results in fulfilment of the objective of the recognition process and in enhanced performance of the recognition scheme. The ACR pre-processing phase influences effectiveness, dependability and reliability of the processes of classification, feature and extraction segmentation. To enhance the performance of the ACR system, the pre-processing process should include the processes of binarization; image segmentation and decomposition; smoothing and noise removal; slanting correction; skew extraction; thinning; and baseline detection [1][4][5][7].

The objective of this study is to suggest a pre-processing model for recognition of hand-written Arabic text so as to ensure highly reliable text recognition and make better use of the data in the hand-written text. This study, hence, proposes a working technique depending on geometrical rules, specifically, the Voronoi Diagram (VD), and is designed in view of exploited components of the VD. This paper is organized as follows. Section 2 introduces the writing characteristics of the Arabic text. Section 3 provides a general background on the pre-processing methods of the hand-written Arabic text. Section 4 presents a general background on the definitions and properties of the VD. Section 5 provides the proposed pre-processing model for the hand-written Arabic text. Afterwards, conclusions and future directions are provided.

2. CHARACTERISTICS OF THE ARABIC WRITING

It is widely conceived that recognition of Arabic text is more challenging than recognition of others like the Chinese and Latin characters [1][7][35]. The difficulties in recognition of the Arabic characters may be mainly attributed to the following seven reasons:

1- The Arabic language has 28 characters and the Arabic characters are written cursively. Each character is written in either of 4 shapes, depending on its place in the word.

2- The Arabic text is usually written from the right direction to left as shown in Figure 1 for the hand-written sentence: **(جامعة آل البيت تحييكم وترحب بكم أجمل ترحيب),** which means 'Al Al-Bayt University greets, and warmly welcomes, you'.

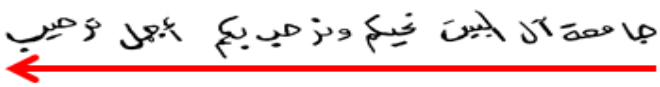


Figure 1: Direction of writing of the Arabic text.

3- The Arabic word may comprise two or more sub-words and the particular word is divided into sub-words if any of the characters (أ, د, ذ, ر, ز, و) exists mid of the word. Figure 2 presents a hand-written Arabic word consisting of four sub-words.

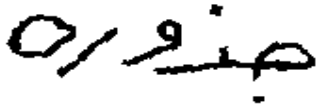
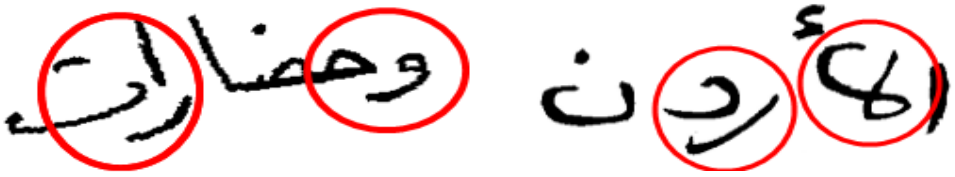


Figure 2: An example of a hand-written Arabic word (جذوره), that is, 'Its roots', consisting of four sub-words.

4- The Arabic words may have overlapping and ligatures. Overlapping exists when two or more strokes crosscut one the other (see Figure 3 (i)). Ligatures, on the other hand, occur when two or more strokes are connected to one the other (see Figure 3(ii)).



(i) (ii)

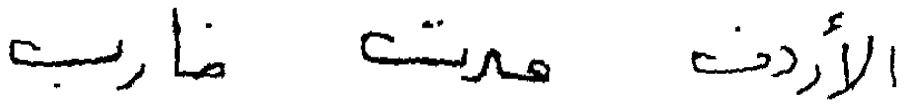
Figure 3: Hand-written Arabic words: (i) Overlapping and (ii) Ligatures.

5- The Arabic words may include diacritics (called in Arabic 'Tashkeel'), which are regarded as short vowels. These Tashkeels are 'Dhammah', 'Fathah', 'Sukun', 'Kasrah', 'Tanween', 'Shaddah' and 'Maddah'. Figure 4 displays positions of these diacritics in association with some characters.

Fathah Dhammah Kasrah Sukun Maddah Shaddah Tanween

Figure 4: Positions of diacritics in association with some characters.

6- Of the 28 basic Arabic characters, fifteen characters have dots ranging in number from one to three. The dots differentiate a character from other(s) of the same shape. These characters are (ب، ت، ث، ج، خ، ذ، ز، ش، ض، ظ، غ، ف، ق، ن، ي). On the other hand, some characters have a zigzag appended character called in Arabic 'Hamzah'. Examples of these characters include (أ، ك). Figure 5 gives examples on some Arabic words with hamzah and dots.



(i) (ii) (iii)

Figure 5: Examples of Arabic words with (i) one dot, (ii) two dots, and (iii) hamzah.

7- The Arabic word is often made up of more than two characters that are linked through horizontal imaginary baseline (see Figure 6).

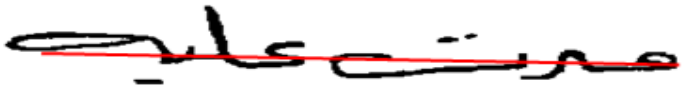


Figure 6: Baseline of the hand-written Arabic text.

3. METHODS OF PREPROCESSING FOR ARABIC TEXT RECOGNITION: GENERAL BACKGROUND

The purpose of image preprocessing is to lower the noise coefficients and raise readability of the input text by the recognition system. As well, the preprocessing stage is necessary to improve

uniformity in texts, which is essential for the recognition system [1] [7]. This stage is quite a crucial phase in the ACR phasess. It does directly affect the effectiveness, dependability and reliability of the processes of segmentation, feature extraction, and classification. So as to enhance performance of the ACR system, the preprocessing stage should generally include binarization; image segmentation and decomposition; smoothing and noise removal; slanting correction; skew extraction; thinning; and baseline detection[1][3][7][8]. The general framework for the methods of preprocessing of the hand-written Arabic text is displayed by Figure 7 and an explanation of these methods is given in the subsequent sub-sections.

3.1 Binarization

Usually, the ACR systems accept the inputs in a bi-level format or, more specifically, a binary format. In general, the input text is presented in grayscale image format. Thus, a preprocessing process referred to as 'Binarization' is needed. This process converts the image format from the grayscale to a bi-level, or binary, image format taking into account a cutoff pixel value for comparison purposes. This value may be calculated using a histogram of the image's gray values [3][7].

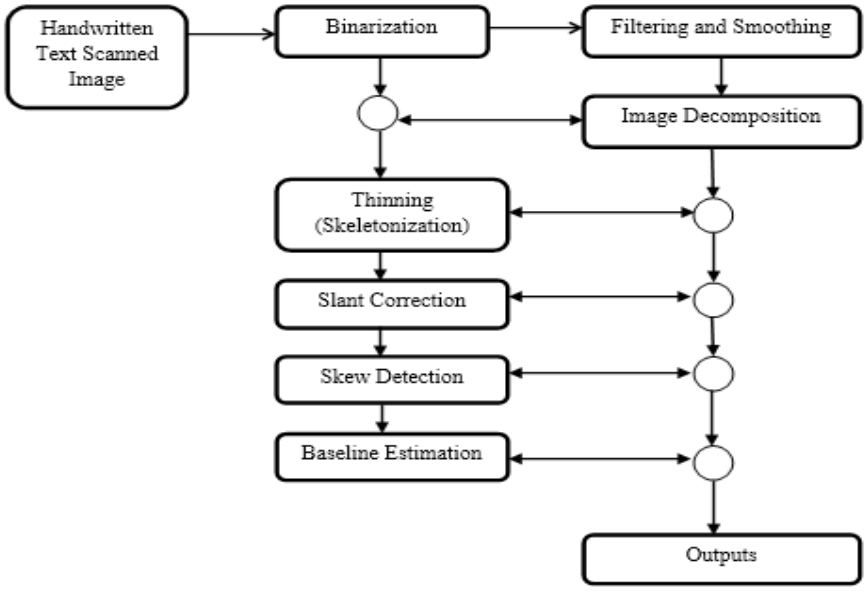


Figure 7: The general framework for the methods of preprocessing of the hand-written Arabic text [1] [7].

3.2 Filtering and Smoothing

Noise may develop in an image after binarization or scanning. It is fundamental to eliminate the noise and smoothen the text's input image in order to formulate the data for subsequent processing. In general, the ACR systems are highly sensitive to noise as it negatively affects system's performance. Therefore, usually Gaussian or median filters are used a procedure in image processing to eliminate the noise [7][16][17].

3.3 Segmentation of Pages by VDs

The segmentation process is one of the main stages in document preprocessing for analysis and the VD is used extensively for segmentation purposes. Ittner and Baird [18] suggested a VD-based system for detection of text line orientation and segmentation of document images into

lines, blocks, symbols, and words. Xiao and Yan [34] suggested the delaunay tessellation for the purpose of extracting text regions from document images. Wang et al. [33] suggested the area Voronoi tessellation as a means of segmenting connected elements (that is, overlapping characters) into individual characters or graphics. This method detects overlapping characters contours and edges then separates the overlapping characters on the basis of the Area-VD that has been drawn from the characters contour [1].

Particularly in the extraction of text lines from binary images, mainly non-rectangular pages, the physical structure and orientation of pages are maintained when using the Area-VD, which was drawn from the sub-graphs, as it preserves connectivity [19]. The same VD extraction technique was employed by Viska [32] for extraction of the Jawi sub-words and words from binary Jawi document images. So as to accelerate VD page segmentation, Lu and Tan [20] suggested the chain code technique instead of the neighbor graph technique for Area-VD construction [1].

Lu et al. [21] suggested extraction of words from binary text images on the basis of Area-VD. This suggested method was applied to various types of document images, e.g., journals, books, and theses. Zaki et al. [36] too proposed the Area-VD method, which was actually developed from the Point-VD technique, for fragmenting Arabic text into sub-words and fragmenting secondary characters.

4.3 Thinning

Thinning (Skeletionization) is a very vital and important process in the ACR systems. It simplifies the shapes of the Arabic characters for the purposes of segmentation, feature extraction, and classification. This brings about reduction in the volume of data needing handling [1][5].

On the other hand, numerous ACR systems have been proposed and established on the basis of text skeleton [5] which has been used extensively to support the feature extraction and the classification processes [14]. In addition, it has been utilized as the foundations for a number of Arabic text segmentation techniques [35] and for drawing the baseline for the hand-written Arabic texts [29].

Al-Shatnawi and Omar [5] conducted a survey of the skeletonization methods particularly designed for the Arabic text or which have been developed for other purposes but have been employed for Arabic text processing. The different methods have been divided into non-iterative and iterative methods. For further details on these methods, the interested reader is referred to Al-Shatnawi and Omar [5].

Al-Shatnawi et al. [6] supported that the effective thinning procedure must be robust to noise and must conserve each of the dots, shape, and text connectivity. Furthermore, it should generate a skeleton that has one-pixel width, should deal with the necking problem, and should not yield spurious tails [5][6]. Details on the challenges facing thinning of the Arabic text can be found in Al-Shatnawi et al. [6].

3.5 Slanting Estimation and Correction

Slanting is a popular problem in the images of hand-written Arabic texts that is a reflection of the differences between writers in their hand-writing styles. The slanting problem is encountered when the vertical elements of the text are vertical in a slant form on the baseline of the cursive text. The slanting elements are supposed to lie perpendicular to the text's baseline. These vertical characters are referred to as Ascenders (see Figure 8). Leaving the slanting characters with no

modification produces wrong results in the sequent stages of the recognition process like feature extraction and baseline detection [7].

(a) (b)

Figure 8: The hand-written Arabic word 'Laian' (ليان): (a) does not need performing slanting correction and (b) needs performing slanting correction.

In the process of slanting character correction, slopes of the 'Ascenders' must be identified prior to correction of the slanting characters. These slopes can be determined by calculating the center of gravity for each slanting stroke. Then, they can be rotated on the basis of the determined skew angle [12]. A histogram of the gradient's orientation can too be employed for correcting the slants in the images of the hand-written texts [25].

3.6 Skew Detection and Correction

Skew often takes place when using image-acquisition tool or when scanning the document of interest by the computer. The process of skew detection and correction is in general regarded as a critical step in the analysis and understanding of documents in general and is an integral phase for the understanding of Arabic characters in particular. Skew correction may be performed on text lines, or paragraphs, or even sub-words and words. It has a substantial impact on the efficiency and dependability of later character recognition stages and, therefore, it influences the ultimate recognition level and accuracy. Furthermore, the process of skew detection and correction has direct effects on the processes of segmentation and feature extraction. This process has effects that reflect at the whole page level. For further details on skew detection and correction, the interested reader is referred to Al-Shatnawi [3] and Al-Shatnawi and Omar [4].

3.7 Baseline Estimation

As was mentioned earlier in this article, the Arabic text is most often cursively written and the text baseline is clarified as an imaginary line that links all characters of the word [2][8][10]. In view of this fact and the definition of baseline, the shape and characters of the Arabic text are classified as 'Descenders', 'Ascenders', and 'Diacritics' (Figure 9). Descenders and Ascenders lie below and above the baseline, respectively. The diacritics are critical in that different diacritics mark differences in meanings between words or even the same word with different diacritics [2]. The major uniqueness facets of ACR begin with baseline estimation, which can be either utilized in Arabic text segmentation or dependent-feature extraction and skew normalization [15][22].

Figure 9. Descenders, Ascenders, and diacritics in the hand-written Arabic text.

Over the last three decades, researchers have paid much attention to the methods of detection of the baseline in the hand-written Arabic texts and the first such attempt was carried out in 1981 by Parhami and Taraghi [27] and used horizontal projection. This method was later enhanced by Timsari and Fahimi in 1996 [31]. Thereafter, the detection methods grew more and more

sophisticated and researchers started using various techniques like the word skeleton (e.g., Pechwitz and Maergner [29]), principal component analysis (e.g., Burrow [13]), contour representation (e.g., Farooq et al. [15]), and VDs (e.g., Al-Shatnawi [10]). Additional details on the various baseline detection methods have been provided by Al-Shatnawi and Omar [9], Al-Shatnawi and Omar [2], and Al-Shatnawi [10].

4. THE VORONOI DIAGRAMS (VDS)

The VD theory was first formulated in 1644. Later, Dirchlet and Voronoi reinvented this concept and extended it to the 3D space. The VD starts with subpartition of the plane of the image into a group of points – also known as sites – and the faces comply with the regions of a closed site. The VD's have been given varied names, generalized, and extensively studied before they found applications in numerous fields that include "physics, marketing, chemistry, astronomy, medicine, image processing, networks, microbiology, telecommunications, geography, and imagery" [1][24]. They are especially involved in cases where a space needs to be divided into irregular lattices named spheres of influence. Consequently, besides geometrical construction, VDs have widely-varying applications that are not at all limited to microstructure modeling, three-dimensional (3D) space and surface planning, and material disposition in the space and in the environment.

Once the central point and the perpendicular edge bisectors are linked together, a VD is obtained. In Figure 10, consider that P is a plane. A number of different points, n, which are termed Voronoi generators, is obtained. The VD is designed when P is sub-divided into geometric objects, or cells, corresponding to one point [1][9][16]. A point, q, will be positioned in the cell px if, and only if, its distance from px (that is, q) is not higher than its distance of any other point py. In this case, the Voronoi region is actually the cell Px [1][10]. In mathematical terms, this as [1][9][10]:

$$q \in P_x \text{ iff } |q - P_x| < |q - P_y|;\ x \neq y \ldots \quad (1)$$

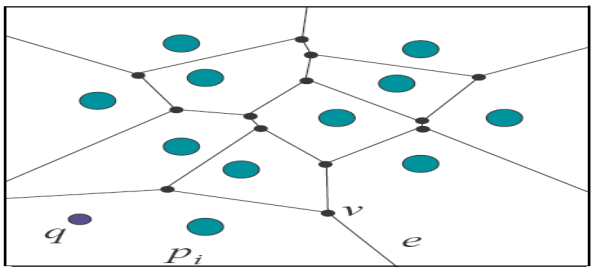


Figure 10. A VD plane of points Px. In this figure, q is a free point, e is the Voronoi edge, and v is the Voronoi vertex [1][10].

The VD has three main properties:

1. "The Voronoi vertex is the center of the sphere of influence" [1][10].
2. "The Voronoi regions are infinite and finite convex polygons" [1][10].
3. The Voronoi edges are infinite lines, line segments, or half lines that are the boundaries of Voronoi regions. The Voronoi edges are shaped by connecting the perpendicular bisectors with the segment of the line which links two points in the plane. Meanwhile, the Voronoi neighbors are the sides of the two points which form the Voronoi edge [1][10].

5. PROPOSED VD-BASED MODEL FOR PREPROCESSING HAND-WRITTEN ARABIC TEXTS

Due to impact and importance of the preprocessing step on the segmentation, feature extraction, and classification stages, this paper proposes a thorough VD-based model for preprocessing of hand-written Arabic texts. This model consists of four main stages: a preparatory stage, page segmentation, thinning (Skeletonization), and baseline estimation and slanting correction. In the first, i.e., preparatory, stage, the text image is converted via VDs into a group of geometrical forms that consist of edges and vertices and which are used to construct the other stages of the proposed model. These edges and vertices are generated from sampling points that are determined on the basis of text contour. This stage consists of four main processes: binarization, edge extraction and contour tracking, sampling, and Point-VD construction.

The second stage in the proposed model is the page segmentation depending on the Area-VD which was built on the basis of Point-VD. This process can divide the text into text-lines or sub-words or diacritics. In the third stage, an efficient method for text structuring (that is, thinning) is presented. This method can efficiently extract the text by means of vertices whose coordinates fall entirely within the text coordinates, with some restrictions and provisions through which potential baseline points can be determined. This method can find the baseline in straight or curved manner. In the fourth stage, an efficient technique for slanting detection and correction is proposed that first specifies a slanting then corrects it. This technique depends on slanting specification depending on each of the potential baseline points, text thickness, and vertices for each stroke in the text. Figure 11 displays the framework and stages of the suggested VD-based preprocessing model.

5.1 The Preparatory Stage:

This stage formulates the data for subsequent stages. It consists of four operations: binarization, edge extraction and contour tracking, sampling processing, and Point-VD construction. In this stage, edges of the lines of the input text are determined and both the outer and inner contours are tracked. Afterwards, the contours are converted into a group of sampling points that are then used to generate Voronoi frameworks which contain numerous features that can be used in the stages later to this particular stage.

5.1.1 Binarization

Otsu's [26] method was followed in this paper to transform the input images into a binary format by means of thresholding and clustering. This method employs a threshold to reduce variance between white and black pixels. In consequence, the pixels which do not fall on either the background or foreground are computed based on the level of pixel spread at each side from the threshold. The aim of this approach is to discover the lowest threshold value and the lowest value of the sum of background and foreground spreads. The algorithm of this method recommends that a text image of threshold consists of two categorizes of pixels; background and foreground pixels. These two categorizes are then isolated so as to compute the optimal threshold with the lowest intra class variance (i.e., lowest combined spread) that can be summed up. The researcher employed this method owing to that (i) it is a worldwide binarization method, and (ii) it has a short running time [3][30].

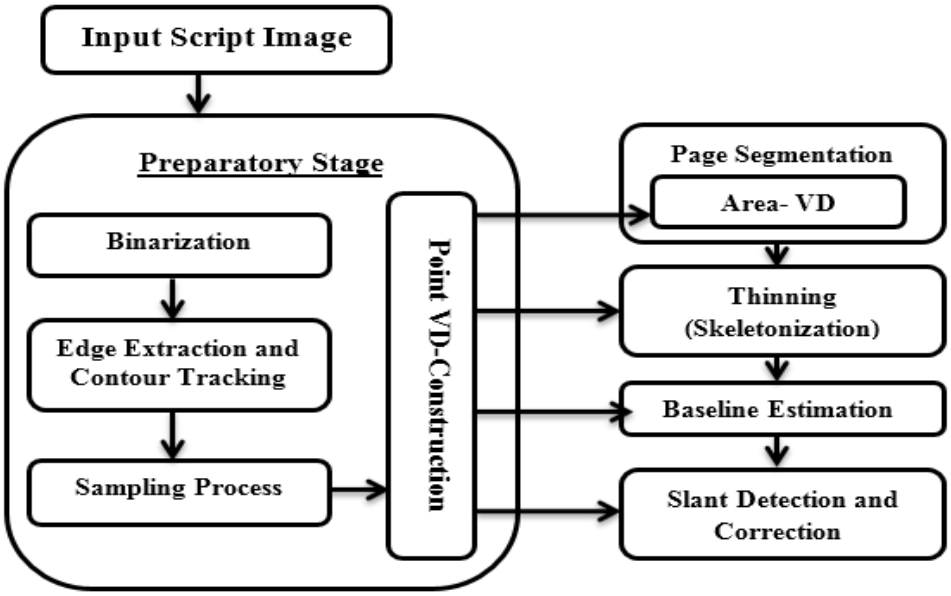


Figure 11. Framework and stages of the suggested VD-based preprocessing model.

5.1.2 Edge extraction and Contour Tracking

In the process of edge extraction and contour tracking, the edges are extracted by finding a 1-0 or 0-1 transition in the pixel values. This is generally performed using window dimension of 3 x 3. For whichever given pixel, neighbours in the eight potential directions (south-west, south, north-west, west, north-east, north, south-east, and east) are to be compared as shown in Figure 12 (a).

(i) (ii)

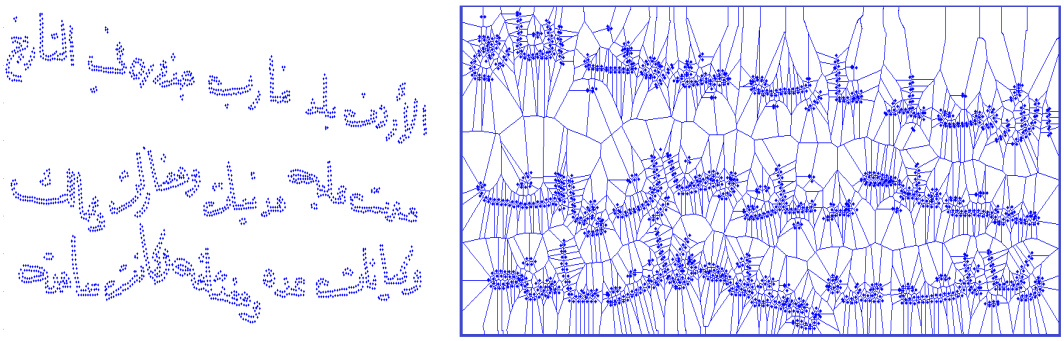
Figure 12. Hand-written Arabic script: (i) edge extraction, and (ii) contour tracking.

The method of contour representation is a mere extension of this method. It attempts to mark the edges and remove other image contents by converting the 1's into 0's for the core pixel [1][9][10]. The text of concern is traced so as to be then converted into a group of sampling points that can be used as the VD generators which will be employed in baseline identification in the subsequent step. The contour of the image of the text presented in Figure 12 (a) is portrayed by Figure 12 (b).

5.1.3 The Sampling Process

This process demands down sampling of the pixels produced by contour tracking in the preceding step. The head pixel is selected using a special function that handles the image column wise from left to right. Once a starting point is specified, the next step corresponds to down selection points

(sampling) of the contour by selecting every Rth pixel retained while eliminating all pixels in between, where R is a selection argument and is higher than 1.0. This function moves clockwise. When R = 1, all the contour pixels are recommended. The sampling interval, R, is selected on the basis of VD construction [10]. The sampling points are presented in Figure 13 (a).



(i) (ii)

Figure 13. Hand-written Arabic script: (i) The sampling process, and (ii) VD construction.

5.1.4 Voronoi Diagram (VD) Construction

At this stage, the VD is created from the sampling points of the text contour (that is, from the VD generators) that have been produced in the first stage. The VD construction process transforms the text, which contains edges and vertices, into the geometric (mathmatical) forms of polygons and convex holes. In this paper, construction of the Point-VD will be clarified by example. The hand-written Arabic script displayed in Figure 12 (a) is read first. Then, edges are detected and the contour is tracked via the 8-nigborhood contour representation scheme as illustrated by Figure 12 (b). Thereafter, the samples are taken along the contours based on R = 6 (Figure 13 (a)). Afterwards, the VD is created by using all the samples as VD generators (Figure 13 (b)).

As has been mentioned in the previous sections, the VD has three major characteristics: Voronoi edges, Voronoi area (i.e., Voronoi regions), and Voronoi vertex. The Voronoi edges of the Point-VD have been identified into the following four kinds in view of their positions in the base text. The edges are characterized in the zoomed image of the word (ساحته) that is displayed in Figure 14 (a) accordingly [1][9][10]:

1- Edges completely falling inside the object (that is, sub-words or joined components).
2- Edges completely falling outside the object. Four such objects are known:

    2a. Edges created from twice nearby Voronoi polygons, each arriving of a generator that falls on different linked component.
    2b. Edges created from twice close Voronoi polygons that both come from one linked component. The generators are adjacent but are not next to one the next on the contour.
    2c. Edges created from two nearby Voronoi polygons that both derive from one joined component but the generators are far away from one the other; one often falls on other tail of the linked component.
    2d. Edges located in a hole.
3- Edges partially falling inside the text body.
4- Edges extending relatively infinitely.

On the other hand, the Voronoi vertices are identified into the following categories in light of their positions relative to the text object. The vertices are labeled in the zoomed image of the word (بلد) in Figure 14 (b) according to this classification [1][10]:

a) Coordinates of vertices located in the object,
b) Coordinates of vertices located outside of the object, and
c) Vertices with unknown coordinates. Virtually, these vertices are considered as having coordinates extending to infinity.

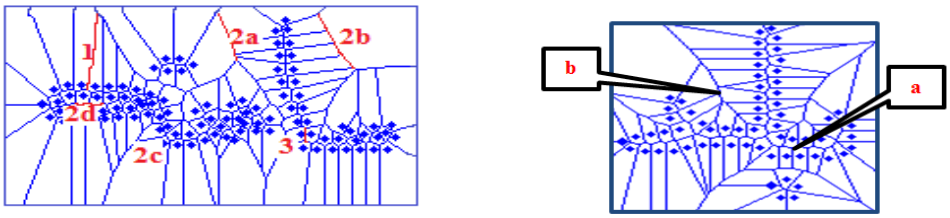


Figure 14. (a) Kinds of Voronoi edges (b) Kinds of Voronoi vertices.

The above-recorded sorts of Voronoi edges and vertices will be exploited as a tool for developing the method for pre-processing of the hand-written Arabic text that is suggested by this paper.

5.2 Page segmentation of hand-written Arabic scripts using VDs

At this stage, an efficient method for page segmentation of hand-written Arabic script based on Area-VD is presented. The proposed VD-based page segmentation method segments the Arabic script into a set of text-lines or sub-words or diacritics and the Area-VD is produced from the Point-VD. In this study, page segmentation of the hand-written Arabic script will be achieved through the following steps:

1- Script of the image of the hand-written Arabic text is first read.
2- The image is then binarized following Otsu's [26] method.
3- The edges are then determined and the outer and inner contours are tracked following the 8-nigborhood method of contour representation.
4- The samples are selected along the contours of the text to serve as the VD generators.
5- The VD is created using the VD generators.
6- The edges of the 2a type (refer to previous section) are maintained and all other edges are deleted.
7- The text shown in Figure 12 has been divided into words and diacritics or into sub-words or even lines (Figure 15).

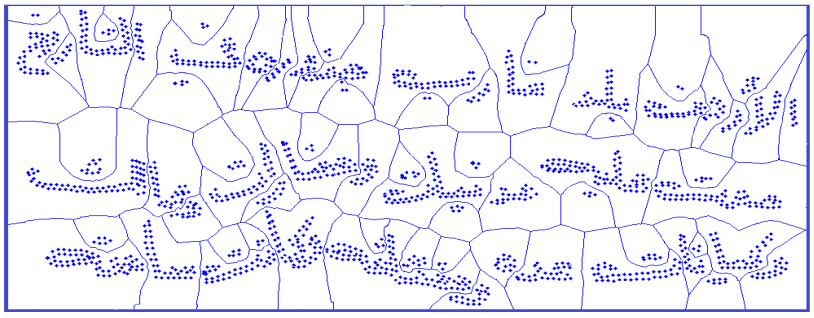


Figure 15: Segmentation of sub-words and diacritics using Area-VD.

5.3 Thinning of the hand-written Arabic text using VDs

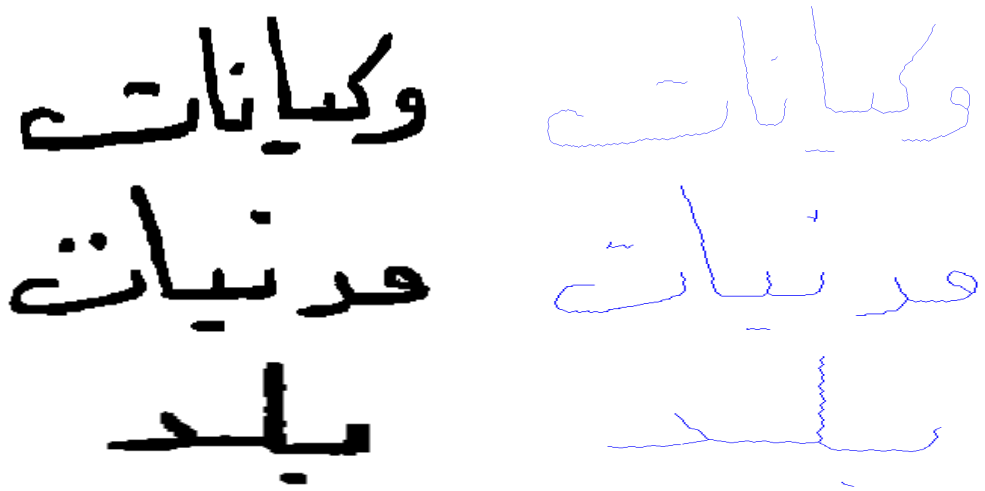
Al-Shatnawi [1] suggested efficient non-iterative thinning technique for the hand-written Arabic text on the basis of the exploited vertices of the VD. This technique detects the skeleton of the hand-written text using the samples chosen along the text contour. Then, a Point-VD is built from these sampling points. Only the VD vertices that lie within the boundaries of the text are maintained and connected. In addition, Al-Shatnawi [1] suggested a thinning-based VD technique for the hand-written Arabic text by choosing vertices of Type A (Algorithm 1). Outcomes of this proposed technique are shown in Figure 16.

Algorithm 1: Al-Shatnawi [1] suggested thinning method that is depending on the exploited VD vertices.

{

- Each pixel in the text image is identified using a number that refers to the joined component to which it belongs. All background pixels are labelled and identified with a zero values.
- Both the outer and the inner contours are tracked.
- Samples are chosen along the contour by using a constant sampling interval, *R*.
- A point-VD is built by considering the all chosen samples as generators.
- Type 'A' voronoi vertices are kept and all other VD components are deleted.
- If two vertices possess two or more specified VD cells, then they are nearby vertices.
- Each vertex is connected with its neighbouring vertices.

}



(a): Original image **(b): Skeleton**

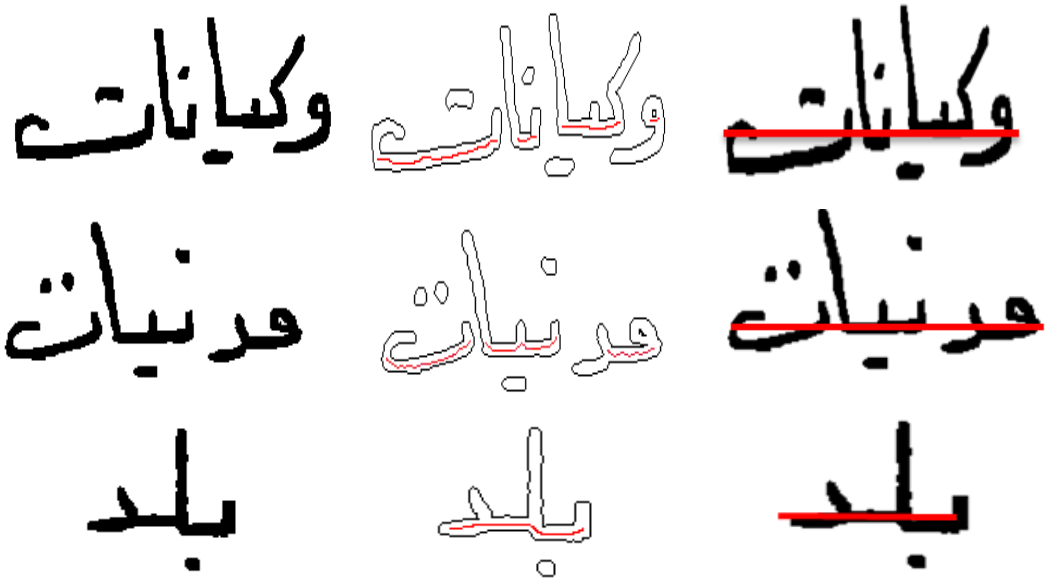
Figure 16. Examples of Arabic text skeletons obtained from Al-Shatnawi [1] thinning-based VD extraction method.

5.4 Estimation of the baseline of hand-written Arabic text using VD

Al-Shatnawi [10] suggested effective method for estimation of the baseline of hand-written Arabic text that is based on exploited components of the VD. This method determines the baseline in two stages. In the first stage, the candidate (potential) points of the baseline are chosen on the basis of the Voronoi edges and vertices located within the text boundaries (namely, Type 'A' and '1' (see Algorithm 2) of the Voronoi edges and vertices). Any vertex/vertices satisfying one or more of the next conditions is a potential baseline point:

1. Three or more edges linked with one the other in the one vertex.
2. Angle between two nearby edges that are tied to the same vertex lies between 45 and 90 degrees.
3. A single potential point in any linked component point will be overlooked.

While in the second stage the baseline is detected. The method proposed by Al-Shatnawi [10] estimates the baseline of the hand-written Arabic text in curved or straight appearances. Outcomes of an example application of Al-Shatnawi [10] proposed VD-based method for estimation of the baseline of hand-written Arabic text are shown in the Figure 17.



(i): Original image (ii): Curved baseline (iii): Straight baseline

Figure 17. Examples of Arabic text baselines gained using Al-Shatnawi [10] method for baseline estimation in straight and curved lines.

Algorithm 2: Method for estimation of the baselines of hand-written Arabic texts depending on the exploited components of VD [10].

{

- A number that identifies the linked components is assigned to each pixel in the text's image. Meantime, the background pixels are assigned a value of zero.
- Both the outer and the inner contours are fully traced.
- The Voronoi generators are specified during contour tracing by using a constant interval, *R*.
- Point-VD is built by using all of the produced samples.
- Type '1' of the Voronoi edges and type 'A' of the Voronoi vertices are studied and all other VD components are ignored.
- If greater than or three edges are tied with one the other in the one vertex or if the angle between couple nearby edges lies between 45 and 90 degrees, then these vertices are considered as potential baseline points.
- All individual probable points (that is, single probable points in a linked component) are deleted.

- The chosen vertices are connected by straight or kept connected in curved line.

}

5.5 Slanting detection and correction in hand-written Arabic texts using VDs

In this paper, an efficient and effective slanting estimation and correction technique for hand-written Arabic text is proposed. This technique is developed by exploiting the Voronoi vertices and edges that are located above the text's baseline as specified using Algorithm 2. This algorithm proceeds in two main steps: (i) slanting detection for each part of the text (stroke) through computation of text thickness and determination of the potential baseline points and their relations with the vertices associated with the stroke; and (ii) slanting correction. This is achieved by considering the potential baseline point for each stroke needing correction as the focal point based on which the letter is corrected. Figure 18 shows a hand-written Arabic text that requires slanting detection and correction.

Figure 18: A hand-written Arabic text requiring slanting correction

In this paper, segmentation of hand-written Arabic script will be achieved according to the following steps:

1- Image of the hand-written Arabic text is first read.
2- The image is binarized according to Otsu's [26] technique (see Figure 11).
3- All linked components (diacritics) having less than 50 pixels are removed and their position coordinates and associated linked components in the array are saved (see Figure 19(i)).
4- The edges are determined and the outer and inner contours are traced by using the 8-nigborhood contour representation method (see Figure 19(ii)).
5- Text thickness is estimated by using the method presented in Al-Shatnawi [1].
6- The samples are chosen along the contours of the text to serve as the VD generators (see Figure 19(iii)).
7- A VD is built using the VD generator (see Figure 19(iv)).
8- Estimate the text's baseline using Algorithm 2, and keep the coordinates of the potential baseline points (see Figure 19(vi)).
9- Keep all the VD components (Types '1' and 'A' of the Voronoi edges and vertices, respectively) and their associated stroke (segmented component) that is located above the baseline in an array.
10- To detect the strokes that require slanting correction, the following steps are needed:
    a) Let each potential point's *y* value have a range (*y* plus or minus the estimated text thickness), and
    b) If the *y* values of all vertices of every stroke are located within the range of its potential baseline point, then these *y* values require slanting correction, (e.g., أ ,ل , ب). Otherwise, no slanting correction is needed (e.g., م , ك ,ط)
11- For strokes requiring slanting correction, let the *y* values of their vertices be equivalent to the *y* value of their potential baseline points (see Figure 19 (vii)).
12- The diacritics are returned into their position in the slanted image (see Figure 19 (viii)).

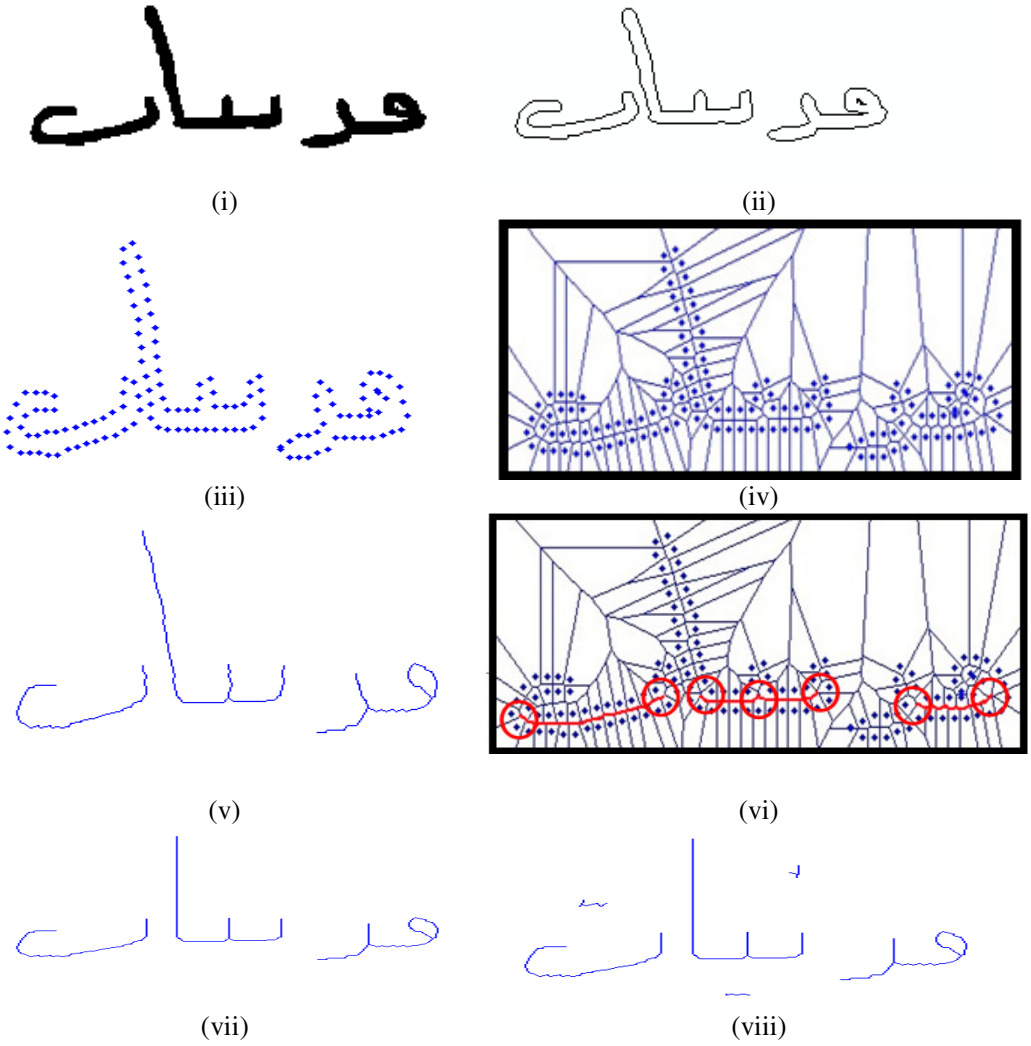


Figure 19: The hand-written Arabic word 'civilizations' (مدنيات): (i) without diacritics, (ii) contour tracking, (iii) the sampling process, (iv) VD construction, (v) Skeleton, (vi) baseline estimation and baseline potential points, (vii) slanting correction, (vii) with diacritics.

6. Conclusions and Future Directions

The preprocessing phase is the single most important phase in ACR. It directly affects the dependability, efficiency and effectiveness of the segmentation, feature extraction, and classification processes. In this paper, a model for preprocessing of hand-written Arabic text based on VDs was presented and discussed. In the proposed VD-based model, four steps for preprocessing of the handwritten Arabic text were developed. These steps are page segmentation, thinning (Skeletonization), baseline estimation, and slanting detection and correction. For development of these steps using VDs, an initial stage, referred to as the preparatory stage, has been established. This stage consists of four main processes: binarization, edge extraction and contour tracking, sampling, and point VD-construction. In this stage, the text image is converted by VDs into a group of geometrical forms made up of edges and vertices that are used for

constructing subsequent stages of the proposed model. These edges and vertices are generated based on the sampling points that are determined on the basis of the text contour.

In the process of page segmentation based on VDs, the page is partitioned based on Area-VD, which is established on the basis of Point-VD. One of the outcomes of this process is splitting the text into text lines, sub-words, or diacritics. For thinning purposes, this study employed the method proposed by Al-Shatnawi [1] for text skeletonization, which is regarded as a non-iterative thinning approach for hand-written Arabic text depending on the exploited vertices of VDs. This method functions by extracting the skeleton from the VD vertices that lie within the text boundaries and has a variety of advantages including maintaining shape connectivity, avoiding spurious tails, preserving one-pixel width skeleton, and working accurately with skewed text images. On the other hand, for baseline estimation, this study employed the method proposed by Al-Shatnawi [16], which operates on defining the baseline by utilizing the edges and vertices whose coordinates fall entirely within the coordinates of the text, with some restrictions and provisions through which potential baseline points can be determined. This method can determine the baseline, be it straight or curved track.

As well, a slanting detection and correction method depending on VDs has been proposed in this study. The method operates in two major steps: (i) slanting detection for each part of the text (stroke) through computation of text thickness and determination of the potential baseline points and their relations with the vertices associated with the stroke; and (ii) slanting correction. This is achieved by considering the potential baseline point for each stroke needing correction as the focal point based on which the letter is corrected. In the proposed model, the VDs have been constructed one time only and by means of these diagrams the four proposed processes have been executed. By so doing, the time and cost of performing ACR are reduced. As well, this makes use of the geometric frameworks (that is, VDs) an acceptable and easy process. In view of this study and its findings, future research is directed to construct a comprehensive, integrated system of ACR using VDs based upon careful reading of the outcomes of the model proposed by the current study and utilizing this model in the other stages such as feature extraction and classification.

REFERENCES

[1] AL-Shatnawi Atallah. A Non-Iterative Thinning Method Based on Exploited Vertices of Voronoi Diagrams, Phd thesis, Universiti Kebangsaan Malaysia, Malysia, 2010.

[2] AL-Shatnawi, Atallah and Khairuddin Omar, Methods of Arabic Language Baseline Detection -The State of Art, International Journal of Computer Science and Network Security, 2008, vol. 8 (10), pp. 137-142.

[3] Al-Shatnawi, Atallah M. "A skew detection and correction technique for Arabic script text-line based on subwords bounding." In Computational Intelligence and Computing Research (ICCIC), 2014 IEEE International Conference on, pp. 1-5. IEEE, 2014.

[4] Al-Shatnawi, Atallah M., and Khairuddin Omar. "Skew detection and correction technique for arabic document images based on centre of gravity." Journal of Computer Science 5, no. 5 (2009): 363.

[5] Al-shatnawi, Atallah M., and Khairuddin Omar. "The Thinning Problem in Arabic Text Recognition " A Comprehensive Review." International Journal of Computer Applications 103, no. 3 (2014).

[6] AL-Shatnawi, Atallah M., Khairuddin Omar, and Ahmed M. Zeki. "Challenges in thinning of arabic text." In ICGST International Conference on Artificial Intelligence and Machine Learning (AIML-11), Dubai. United Arab Emiratis, pp. 127-133. 2011.

[7] AL-Shatnawi, Atallah Mahmoud, Safwan AL-Salaimeh, Farah Hanna AL-Zawaideh, and Khairuddin Omar. "Offline arabic text recognition–an overview."World of Computer Science and Information Technology Journal (WCSIT) 1, no. 5 (2011): 184-192.

[8] AL-Shatnawi, Atallah, and Khairuddin Omar. "A comparative study between methods of arabic baseline detection." In Electrical Engineering and Informatics, 2009. ICEEI'09. International Conference on, vol. 1, pp. 73-77. IEEE, 2009.

[9] Al-Shatnawi, Atallah, and Khairuddin Omar. "Detecting arabic handwritten word baseline using voronoi diagram." In Electrical Engineering and Informatics, 2009. ICEEI'09. International Conference on, vol. 1, pp. 18-22. IEEE, 2009.
[10] Al-Shatnawi, Atallah. Published Online. A Novel Baseline Estimation Method for Arabic handwritten Text Based on Exploited Components of Voronoi Diagrams. The International Arab Journal of Information Technology. Vol 13 (3). May 2016.
[11] ALshebeili, Saleh A., Asim A-F. Nabawi, and Sabri A. Mahmoud. "Arabic character recognition using 1-D slices of the character spectrum." Signal Processing 56, no. 1 (1997): 59-75.
[12] Bozinovic, Radmilo M., and Sargur N. Srihari. "Off-line cursive script word recognition." Pattern Analysis and Machine Intelligence, IEEE Transactions on11, no. 1 (1989): 68-83.
[13] Burrow, Peter. "Arabic handwriting recognition." Report of Master of Science School of Informatics, University of Edinburgh (2004).
[14] Cowell, John, and Fiaz Hussain. "Thinning Arabic characters for feature extraction." In Information Visualisation, 2001. Proceedings. Fifth International Conference on, pp. 181-185. IEEE, 2001.
[15] Farooq, Faisal, Venu Govindaraju, and Michael Perrone. "Pre-processing methods for handwritten Arabic documents." In Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on, pp. 267-271. IEEE, 2005.
[16] Gonzalez, R.C. and Woods, R.E. 2002. Digital Image Processing. Prentice Hall. 2nd edation.
[17] Gross, Ari, and Longin Jan Latecki. "Digital geometric methods in document image analysis." Pattern Recognition 32, no. 3 (1999): 407-424.
[18] Ittner, David J., and Henry S. Baird. "Language-free layout analysis." InDocument Analysis and Recognition, 1993., Proceedings of the Second International Conference on, pp. 336-340. IEEE, 1993.
[19] Kise, Koichi, Akinori Sato, and Motoi Iwata. "Segmentation of page images using the area Voronoi diagram." Computer Vision and Image Understanding70, no. 3 (1998): 370-382.
[20] Lu, Yue, and Chew Lim Tan. "Constructing area Voronoi diagram in document images." In Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on, pp. 342-346. IEEE, 2005.
[21] Lu, Yue, Zhe Wang, and Chew Lim Tan. "Word grouping in document images based on Voronoi tessellation." In Document Analysis Systems VI, pp. 147-157. Springer Berlin Heidelberg, 2004.
[22] Nawaz, S. N., M. Sarfraz, A. Zidouri, and W. G. Al-Khatib. "An approach to offline Arabic character recognition using neural networks." In Electronics, Circuits and Systems, 2003. ICECS 2003. Proceedings of the 2003 10th IEEE International Conference on, vol. 3, pp. 1328-1331. IEEE, 2003.
[23] Nazif. Ahmad. 1975. A system for the recognition of the printed Arabic characters. M.Sc. Thesis. Cairo University.
[24] Okabe, Atsuyuki, Barry Boots, Kokichi Sugihara, and Sung Nok Chiu. Spatial tessellations: concepts and applications of Voronoi diagrams. Vol. 501. John Wiley & Sons, 2009.
[25] Omar, Khairuddin, Abd Rahman Ramli, Ramlan Mahmod, and Md Nasir Sulaiman. "Skew detection and correction of Jawi images using gradient direction." Jurnal Teknologi 37, no. 1 (2002): 117-126.
[26] Otsu, Nobuyuki. "A threshold selection method from gray-level histograms."Automatica 11, no. 285-296 (1975): 23-27.
[27] Parhami, Behrooz, and M. Taraghi. "Automatic recognition of printed Farsi texts." Pattern Recognition 14, no. 1 (1981): 395-403.
[28] Parvez, Mohammad Tanvir, and Sabri A. Mahmoud. "Offline Arabic handwritten text recognition: a survey." ACM Computing Surveys (CSUR) 45, no. 2 (2013): 23.
[29] Pechwitz, Mario, and Volker Märgner. "Baseline estimation for Arabic handwritten words." In Frontiers in Handwriting Recognition, 2002. Proceedings. Eighth International Workshop on, pp. 479-484. IEEE, 2002.
[30] Sahlol, Ahmed T., Cheng Y. Suen, Mohammed R. Basyouni, and Abdelhay A. Sallam. "A Proposed OCR Algorithm for the Recognition of Handwritten Arabic Characters." (2014): 90-104.
[31] Timsari, Bijan, and Hamid Fahimi. "Morphological approach to character recognition in machine-printed Persian words." In Electronic Imaging: Science & Technology, pp. 184-191. International Society for Optics and Photonics, 1996.
[32] Viska, M. 2007. Segmention of Jawi Text Using Voronoi Diagram (in Malay). M.Sc. Thesis. University Kebangsaan Malaysia.
[33] Wang, Yalin, Ihsin T. Phillips, and Robert Haralick. "Using area Voronoi tessellation to segment characters connected to graphics." In Proceedings of Fourth IAPR International Workshop on Graphics Recognition (GREC2001), Kingston, Ontario, Canada, pp. 147-153. 2001.

[34] Xiao, Yi, and Hong Yan. "Text region extraction in a document image based on the Delaunay tessellation." Pattern Recognition 36, no. 3 (2003): 799-809.
[35] Zeki, Ahmed M. "The segmentation problem in arabic character recognition the state of the art." In Information and Communication Technologies, 2005. ICICT 2005. First International Conference on, pp. 11-26. IEEE, 2005.
[36] Zeki, Ahmed, Mohamad S Zakaria, and Choong Yeun Liong. "Isolation of dots for arabic ocr using voronoi diagrams." Proceedings of the International Conference on Electrical Engineering and Informatics, 2007.

AUTHOR

Atallah Mahmoud AL-Shatnawi has received his BSc in Computer Science from Yarmouk University (Jordan) in 2005, MSc in Computer Science from the University Science Malaysia (USM) and PhD in System Sciences and Management/Computer Sciences from the National University of Malaysia (UKM) in 2007 and 2010 respectively. Currently, he is an Assistant Professor at the Department of Information Systems, Prince Hussein Bin Abdullah College for Information Technology, Al Al-Bayt University (Jordan). His research interests include: Expert Systems, Pattern Recognition, Image Analysis and Processing as well as Embedded Systems. He has published numerous papers related to these areas.