Expert Systems with Applications 39 (2012) 5470–5477

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

A novel statistical feature extraction method for textual images: Optical font recognition

Bilal Bataineh *, Siti Norul Huda Sheikh Abdullah, Khairuddin Omar

Center for Artificial Intelligence Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia

ARTICLE INFO

Keywords:
Font recognition
Classification
Global analysis feature extraction
Gray level co-occurrence matrix
Statistical feature extraction

ABSTRACT

The binary image is essential to image formats where the textual image is the best example of the binary image representation. Feature extraction is a fundamental process in pattern recognition. In this regard, pattern recognition studies involve document analysis techniques. Optical font recognition is among the pattern recognition techniques that are becoming popular today. In this paper, we propose an enhanced global feature extraction method based on the on statistical analysis of the behavior of edge pixels in binary images. A novel method in feature extraction for binary images has been proposed whereby the behavior of the edge pixels between a white background and a black pattern in a binary image captures information about the properties of the pattern. The proposed method is tested on an Arabic calligraphic script image for an optical font recognition application. To evaluate the performance of our proposed method, we compared it with a gray-level co occurrence matrix (GLCM). We classified the features using a multilayer artificial immune system, a Bayesian network, decision table rules, a decision tree, and a multilayer network to identify which approach is most suitable for our proposed method. The results of the experiments show that the proposed method with a decision tree classifier can boost the overall performance of optical font recognition.

1. Introduction

Pattern recognition in images is an essential aspect of computer visioning in artificial intelligence. The aim is to recognize patterns depending on available knowledge, information or features. Generally, pattern recognition applications consist of three main stages: preprocessing, feature extraction and recognition. In the preprocessing stage, the input image is prepared by removing the noise, correcting the skew, converting the image into a scoped format, and segmenting it into a sub-image pattern or normalizing it into an image block. During the feature extraction stage, an input image is transformed into feature representations containing discrete information, which is used during the recognition stage for classifying known or unknown patterns (Jiang, 2009).

Based on the literature and our experience, feature extraction is one of the most critical issues in pattern recognition applications. Jiang (2009) claimed that the effectiveness of each feature extraction technique is highly associated with distinguishing the similarity of patterns that belong to each identical class from other patterns or noise. Accordingly, we can consider two standard approaches to the feature extraction phase: local (Abdullah, Khalid, & Yusof, 2006) and global (Busch, Boles, & Sridharan, 2005). The local feature extraction approach concerns disconnected parts of an image such as lines, edges, corners, shapes and sub-image regions. This approach interprets the image features that have been manipulated after the segmentation process during the preprocessing stage. Therefore, we can conclude that the performance of local features depends highly on the accuracy of the segmentation phase. Joshi, Garg, and Sivaswamy (2007) also claimed that this approach can cause a lack of class generalization. Conversely, a global approach concerns an overall or sub-region analysis of a natural image. The global feature extraction approach is accomplished with the use of global (or texture) analysis techniques, which extract the global properties of an input image. Those global properties are used as general features in the recognition process. Hence, the global approach can perform better in terms of accuracy during classification of more complex datasets.

The available global feature extraction techniques are normally implemented on grayscale or color images such as medical images, satellite images, iris images, facial images and most natural texture images (Chen, Chen, & Chien, 2009; Mohamed & Abdelsamea, 2008; Najab, Khan, Arshad, & Ahmad, 2010; Sellahewa & Jassim, 2010). However, the binary image is also the standard format for image pattern recognition. A binary image consists of two levels: the white level represents the background and the black level represents the foreground. The binary image format can consist of various texture types (Petrou & García Sevilla, 2006) and textual images (Abdullah et al., 2006; Joshi et al. 2007). Feature extraction

* Corresponding author.
E-mail addresses: b.btnh@yahoo.com (B. Bataineh), mimi@ftsm.ukm.my (Siti Norul Huda Sheikh Abdullah), ko@ftsm.ukm.my (K. Omar).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.11.078

in a binary image is used in applications related to documents, scripts and character recognition (Abdullah et al., 2006; Joshi et al., 2007; Journet, Ramel, Mullot, & Eglin, 2008; Pal, Wakabayashi, & Kimura, 2009), including optical font recognition (OFR). Several previous studies in the field of OFR have focused on East Asian (Ding, Chen, & Wu, 2007; Sun, 2006) or Romance languages (Ma & Doermann, 2005; Zramdini & Ingold, 1998). Until now, the benefits of defining Arabic calligraphy have been neglected. These benefits include classifying the layout of the document, the purposes of the document layout, the document library, the classification of the history of the document, and improving the process of preparing and reprinting the document as the original input image format.

The objective of this study is to propose an enhanced method for global feature extraction. Based on statistical analysis of the behavior of edge pixels in binary images, we have developed a novel method for feature extraction. In this study, we apply this method in an OFR application for an Arabic calligraphic script database. The proposed method is evaluated by comparing it with the gray level co-occurrence matrix (GLCM) suggested by Haralick, Shanmugam, and Dinstein (1973). Then artificial immune system (AIS), Bayesian network, decision table rule, decision tree, and multilayer network classifiers (Castro & Timmis, 2002; Holmes & Lakhmi, 2005; Rokach & Maimon, 2005; Witten & Frank, 2005; Daniel et al., 2008) are applied to obtain the suitable classifier technique and the best performance for the proposed method. This paper is organized as follows. Section 2 reviews the state of the art in previous studies of OFR, approaches to global feature extraction, and a dataset of Arabic calligraphic script. Section 3 explains the OFR system stages and the proposed method, and Section 4 presents and analyzes the experimental results. Finally, conclusions are presented in Section 5.

2. State of the art

OFR is a document analysis technique based on a feature extraction process. Feature extraction is a process between the preprocessing and classification phases and is a fundamental process in pattern recognition to represent features of patterns. As mentioned earlier, the feature extraction process has two categories: local feature extraction and global feature extraction. However, this section describes only previous studies of OFR and a global feature extraction technique called gray level co-occurrence (GLCM).

2.1. Previous studies of OFR

Two categories of feature extraction have been used in OFR applications. For the local approach, Abuhaiba (2005) has presented a framework for printed Arabic OFR based on a general Arabic font type with geometrical features and a decision tree classification technique. These features included geometrical attributes, horizontal projections, invariant moments and Walsh coefficients as the feature value. However, the main disadvantage to this approach is it failed to recognize whether any of the common Arabic words did not exist in the text. Sun (2006) presented a stroke template technique to input features into a decision tree for Chinese and English fonts. However, the stroke method was unable to generalize to different languages; thus, it required modification when any new font was inserted. Additionally, the stroke method is based on a local analysis approach which requires a privet segmentation process, preprocessing and template matching before the font recognition process is executed. Ding et al. (2007) presented another global analysis approach, the wavelet transform for Chinese OFR application. They applied Box–Cox transformation and linear discriminate analysis (LDA) to discriminate wavelet features on a predefined segment to obtain non-overlapping pattern blocks. Finally, the modified quadratic distance function classifier was used to recognize seven font types. However, that step led to slow processing and increased the possibility of risks if other types of fonts were included.

Zramdini and Ingold (1998) presented another geometrical feature method that involved identifying the weight, size, typeface, and slope of a printed English text image block. Basically, they employed a statistical approach based on global typographical features using vertical and horizontal projection profiles, main strokes of characters and connected components. Nevertheless, some font types are difficult to recognize with only a short text sample or a single character. Moreover, this method is based on geometrical properties such as baseline and slope, which does not reveal enough information to apply it to other languages with different structural properties such as Asian or Arabic.

Regardless, the Gabor filter has also become a favorable feature-extraction method (Ding et al., 2007; Ma & Doermann, 2005; Ramanathan et al., 2009a; Ramanathan et al., 2009b; Zhu, Tan, & Wang, 2001) in OFR applications. Zhu et al. (2001) and Ramanathan et al. (2009) used 2-D Gabor filters to extract the texture features for Chinese and English, and Tamil fonts, respectively. Similarly, this method cannot recognize separate characters for each font. These researchers concentrated on machine documents with high quality and ignored low-quality samples such as scanned documents. On the other hand, Ding et al. (2007) claimed that the correlation between the accuracy rate of the Gabor filter and the number of fonts or classes is inversely proportionate. Furthermore, Tuceryan and Jain (1999, chap. 2.1) and Petrou & García Sevilla, 2006 claimed that the Gabor filter consumes more processing time compared to other feature extraction techniques. This method is also sensitive to image noise and ignores the structural properties of the image.

Ma and Doermann (2005) proposed a feature extraction method called the grating cell operator to compensate for isotropic Gabor filter problems. Extracting the orientation texture features of the text images and testing the proposed method using a weighted Euclidean distance classifier and a back-propagation neural network classifier, this approach outperformed the Gabor filter.

Based on the previous related studies, we can conclude that most research conducted with the local feature extraction approach on a single type of language led to high accuracy in classification. Most existing methods are prone to noise interference and highly segmented dependence. The global feature extraction approach, on the other hand, is easy to develop and generalize to other databases. The global approach does not require any basic modification for changes or when other fonts are included; furthermore, it does not require much preprocessing and probability is not affected by noise. In the global feature extraction approach, the accuracy rate relates to the ability of the feature extraction technique that was used. Generally, results from local feature extraction are more accurate than those from global feature extraction. Additionally, there are skipping on using an effected global feature extraction techniques have used in other applications similar to OFR behavior (Busch et al., 2005; Peake & Tan, 1997). Furthermore, only a few studies have focused on the global feature extraction approach.

2.2. Global feature-extraction approaches

The global feature extraction approach can be classified into four categories: statistical methods, structural or syntactic methods, model-based methods and signal-processing techniques. The statistical methods apply statistical descriptions to the observed elements that are related in the image. GLCM is the most common technique among the statistical methods (Haralick et al., 1973; Petrou & García Sevilla, 2006). On the other hand, the syntactic method presents the features by rule replacement to define primitives such as arrangement of textures and description of curves. Syntactic methods used much less often than other methods. In syntactic methods, the grammar is a strictly defined. Additionally, the rule placement is based on regular primitives that have spatial relationships. Hence, syntactic methods are more sensitive to noise and less sensitive when implemented on irregular or variant structural images.

We define the binary image texture as black values that have spatial distribution properties on a white background. Therefore, statistical methods are useful because they are based on the spatial distribution of the values in image. This method describes the distribution of the pixels and the pixels' relationships in the image. A statistical method is simple to implement, is not affected by noise and is easily generalized to different images.

The GLCM is one of the most common techniques that apply a statistical approach to global feature extraction. It consists of a matrix that assigns values that explain the distribution of occurrence in the image. The GLCM values present the number of occurrences that have grayscale value and that are related by specific relationships. The occurrence presents a pair of two pixels or two sets of pixels. If G is a GLCM matrix, I is a $n \times m$ image and $(\Delta x, \Delta y)$ denotes the value of the pairs of pixels that have a gray level value of i and j, as the following formula:

$$G_{\Delta x,\Delta y}(i,j) = \sum_{p-1}^{n}\sum_{q=1}^{m}\begin{cases}1, & \text{if } I(p,q)=i \quad \text{and} \quad I(p+\Delta x, q+\Delta y),\\ 0, & \text{otherwise.}\end{cases} \tag{1}$$

The co-occurrence matrix size is $2^{\text{bit}} \times 2^{\text{bit}}$. The matrix size is $2^1 \times 2^1$ in binary images and $2^{32} \times 2^{32}$ in 32-bit color images. Then the G matrix is converted into a symmetrical form designated as $S_{i,j}$ in formula (2). Finally, the $S_{i,j}$ matrix values will normalize and the P matrix results:

$$S(i,j) = C(i,j) + C(j,i). \tag{2}$$

Haralick et al. (1973) proposed several statistics equations to compute texture features from the GLCMs. Some of these features are the following:

$$\text{Contrast} = \sum_{i,j=0}^{N-1}\mathbf{P}_{i,j}(i-j)^2, \tag{3}$$

$$\text{Homogenetity} = \sum_{i,j=0}^{n-1}\left(\frac{\mathbf{P}_{i,j}}{1+(i-j)^2}\right), \tag{4}$$

$$\text{Angular Second Moment} = \sum_{i,j=0}^{N-1}\mathbf{P}_{i,j}^2, \tag{5}$$

$$\text{Entropy} = \sum_{i,j=0}^{N-1}\mathbf{P}_{i,j}(-\ln \mathbf{P}_{i,j}), \tag{6}$$

$$\text{Variance} \quad \sigma_i^2 = \sum_{i,j=0}^{N-1}P_{i,j}(i-\mu_i)^2, \quad \text{Variance} \quad \sigma_j^2 = \sum_{i,j=0}^{N-1}P_{i,j}(j-\mu_j)^2, \tag{7}$$

$$\text{Correlation} = \sum_{i,j=0}^{N-1}P_{i,j}\left[\frac{(i-\mu_i)(j-\mu_j)}{\sqrt{(\sigma_i^2)(\sigma_j^2)}}\right], \tag{8}$$

where P is the normalized matrix and N is the number of gray levels.

The GLCM has been used for research in documents analysis that depend on a global analysis approach, such as identification of scripts and languages (Busch et al., 2005; Peake & Tan, 1997), characterization of documents (Journet et al., 2008) and writer identification (Shahabi & Rahmati, 2006). The approach has been used in other areas such as iris recognition (Azizi & Pourreza, 2009), fingerprint classification (Yazdi, Yazdi, & Gheysari, 2008), Chinese Sign Language recognition (Quan, Jinye, & Yulong, 2009) and in other studies.

2.3. Dataset of Arabic calligraphic script

Arabic language is an international language that is the primary language spoken in twenty-two countries. Other countries, such as Iran, Nigeria, Senegal and Turkey, use Arabic as their second language. The Arabic alphabet has also been widely adopted in other languages, such as Jawi, Persian, Kurdish, Pashto, Urdu, and Hausa. There are three types of written forms of Arabic: printed, handwritten and calligraphic. The printed version is a machine produced form of a type of Naskh calligraphy. The handwritten form is usually produced by average people in their daily dealings. Finally, Arabic calligraphy, which is the oldest form, is of many types, has fixed standards of accuracy and is written mostly by calligraphy specialists. Arabic calligraphy is also known as an art expressing Islamic culture. For example, museums and libraries frequently keep historical scientific or artistic documents that were written in different types of calligraphy.

Understanding Arabic scripts first requires understanding that the standard Arabic alphabet used in a word or a sentence normally consists of a main baseline and an ascending line and a descending line extending above or below the baseline. Additional characteristics, which differ from Arabic calligraphy, are special slanting, straightness, length and thickness. There are eight major types of Arabic calligraphy: Old Kufi, Kufi, Thuluth, Naskh, Roqaa, Diwani, Persian and Maghrebi (Andalusi). Fig. 1(a) through (h) show examples of a sentence (al-khat lesan al-arab) with different structures or natures based on the type of calligraphy.

3. The proposed method

We defined a method for extraction of statistical features based on the behavior of pixels in the boundaries between black and white in binary images. We found that the relationships between edge pixels can be associated with pattern structures in binary images. Textual images provide the best presentation of binary images. For that, we tested this technique with document applications. Fig. 2 shows the flowchart of the steps in the OFR approach, that is, the preprocessing stage, the edge-detection stage, the feature extraction stage and the recognition stage.

3.1. Preprocessing

The input images usually have a variety of properties, and we apply processes to overcome this problem. At this stage, we prepare the image using binarization, skew correction, and text normalization. We perform binarization using the Otsu threshold value (Otsu, 1979). Then we use the Hough Transform (Singh, Bhatia, & Kaur, 2008) to deduct and correct the skew angle in the skew correction sub-phase. Finally, we remove the spaces between words and lines, fill the incomplete lines and prepare the text size and the number of lines to generate the full text blocks consecutively in the text normalization sub-phase. The final result of preprocessing is a 512×512 text block image with 96 DPI, which is suitable for global analysis approaches. Fig. 3 shows a selected input image and the final image results after applying the above methods in the preprocessing stage.

3.2. Edge detection

The proposed method analyzes the edge images. Many edge-detection techniques are available, including those of Canny, Kirsch, Sobel and others. However, in this study, we applied a

Fig. 1. The sentence. (al-khat lesan al-yad) "the calligraphy is a tongue of hand" written in each of the major types of Arabic calligraphy: (a) Diwani, (b) Kufi, (c) Thuluth, (d) Persian, (e) Roqaa, (f) Naskh, (g) Maghrebi, and (h) Old Kufi.

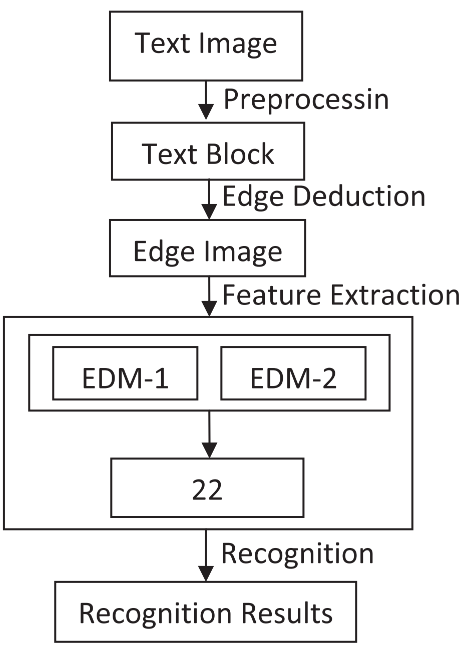


Fig. 2. The flowchart for font recognition of Arabic calligraphy.

Laplacian filter with a 3×3 kernel matrix with one iteration number (Fig. 4(a)), which is a powerful technique to detect edges in all directions and is also effective to solve salt-and-pepper noise. The image result of the Laplacian filter consists of a black background and white edges, as depicted in Fig. 4(c). Because we were interested in the black pixels, we inverted Fig. 4(c) using an inversion filter, which produced the white background and black edges as shown in Fig. 4(d). Fig. 4 shows (a) the 3×3 kernel matrix filter values, (b) an input image, (c) an image after applying the Laplacian filter and (d) a final image result after applying the inversion filter.

3.3. Features extraction

We proposed a novel feature extraction method based on a statistical analysis of the relationships among pattern boundary edge pixels in binary images. Each pixel in the image associates with a 3×3 kernel matrix of eight pixels that neighbor the center pixel, as depicted in Fig. 5(a). The position of each of the neighboring pixels represents a relationship direction with the center pixel, as depicted in Fig. 5(b). We set a standard rule to represent each pixel position as diagonal up left or up right or down left or downright, vertical up or down, and horizontal left or right, as shown in Table 1. We used the eight position values as occurrences where the number of occurrences is stored in the related cell in the edge direction matrix (*EDM*) as depicted in Fig. 5(c). Regarding the above representations, the *EDM* values are presented in two ways: (1) the first-order relationship and (2) the second-order relationship.

In the first-order relationship, we created a 3×3 first-order edge direction matrix (EDM_1). Each cell in EDM_1 is located at a position from 0° to 315° degrees, as shown in Fig. 5(c). Next, we found the relationship directions of the scoped pixel, Iedge(x,y) by calculating the number of occurrences for each value in EDM_1. The algorithm is as follows:

For each pixel in *Iedge*(x,y)
  If *Iedge*(x,y)) is black pixel at center, **then**
  Increase number of occurrence at EDM_1 (2,2) by 1
  If *Iedge*$(x+1,y)$ is black pixel at 0°, **then**
  Increase number of occurrence at EDM_1 (2,3) by 1
  If *Iedge*$(x+1,y-1)$ is black pixel at 45°
  Increase number of occurrence at EDM_1 (1,3) by 1
  If *Iedge*$(x,y-1)$ is black pixel at 90°, **then**
  Increase number of occurrence at EDM_1 (1,2) by 1
  If *Iedge*$(x-1,y-1)$ is black pixel at 135°, **then**
  Increase number of occurrence at EDM_1 (1,1) by 1

Fig. 3. An input image (left) and the image results after the preprocessing stage (right).

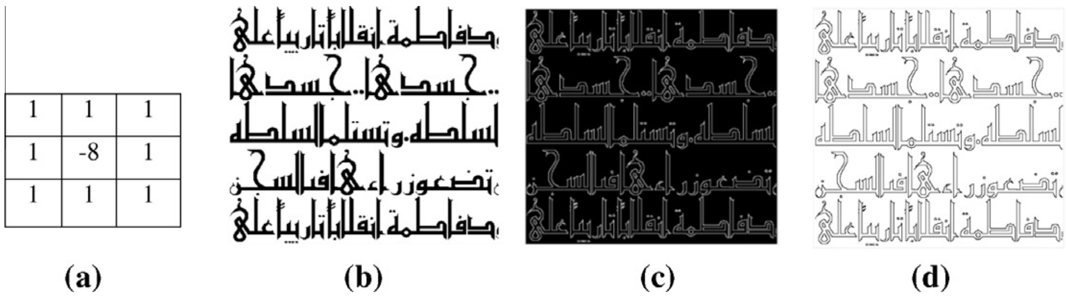


Fig. 4. (a) Laplacian filter value, (b) the original image, (c) filtered image by the Laplacian process and (d) after applying the inversion filter.

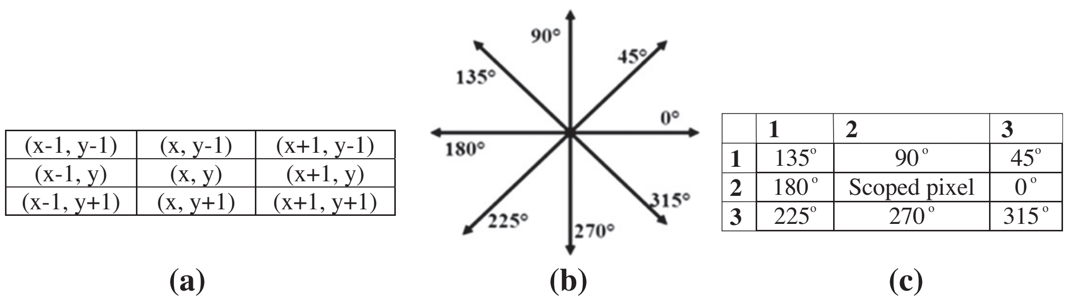


Fig. 5. (a) The eight neighboring pixels, (b) the direction of angles in the image, and (c) the edge direction matrix (*EDM*) of the values of the cells.

Table 1
Representations of each pixel behavior, its coordinates and the related Storage cell in the EDM.

Type	Positions (θ)	Coordinates	*EDM* (x,y)
Center	Scoped pixel	(x,y)	(2,2)
Horizontal right	0°	$(x+1,y)$	(3,2)
Upright diagonal	45°	$(x+1,y+1)$	(3,1)
Vertical up	90°	$(x,y-1)$	(2,1)
Upleft diagonal	135°	$(x-1,y-1)$	(1,1)
Horizontal left	180°	$(x-1,y)$	(1,2)
Downleft diagonal	225°	$(x-1,y+1)$	(1,3)
Vertical down	270°	$(x,y+1)$	(2,3)
Downright diagonal	315°	$(x-1,y+1)$	(3,3)

In the first-order relationship, each pixel in the edge image relates to two or more pixels. For example, in Fig. 6(a) the scoped pixel represents 180° for X_1 and 45° for X_2. That means each pixel represents two relationships in (EDM_1). That finding led to a symmetrical EDM_1, as shown in Fig. 6(b and c). Hence, we address the values of the angles from 0° to 135° in the EDM_1.

In the second-order relationship, each pixel must represent one relationship only. First, we created a 3×3 s-order edge direction matrix (EDM_2). Next, we determined the importance of the relationship for $Iedge(x,y)$ by sorting the values in EDM_1 in descending order, as shown in Fig. 6(c) and Table 2, respectively. We determined the most important relationship of the scoped pixel in $Iedge(x,y)$ by calculating the number of occurrences for each value in EDM_2. The relationship orders were as follows:

(i) If there is more than one angle with the same number of occurrences, then the smaller angle is selected first.
(ii) Next, the reversal angle is selected.

The algorithm of the second order of the EDM_2 relationship is as follows:

Step 1: Sort in descending order the relationships from $EDM_1(x,y)$,
Step 2: For each pixel in $Iedge(x,y)$,
Step 3: **If** $Iedge(x,y)$ is a black pixel, **then**
Step 4: Find the available relationships between neigboring pixels,
Step 5: Compare the relationship values between the available relationships,
Step 6: Increase the number of occurrence at the related cell in $EDM_2(x,y)$.

The results of the first-order relationship and the second-order relationships presented in EDM_1 and EDM_2 are shown in Fig. 7.

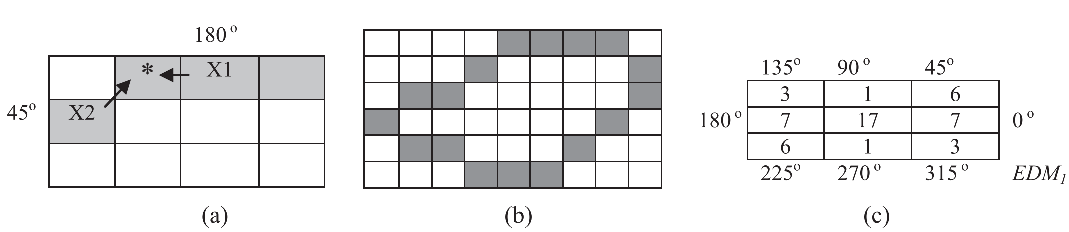


Fig. 6. (a) The two neighboring pixels and (b) the edge image and (c) is its EDM_1.

Table 2
The order of the angle's importance.

Order	Angle	Value
1	0°	7
2	180°	7
3	45°	6
4	225°	6
5	135°	3
6	315°	3
7	90°	1
8	270°	1

Note that the sum of the values of all angles in EDM_2 are equal to the scoped image value EDM_2 (2,2).

We proposed several equations to include features from the EDM_1 and EDM_2 values. The proposed equations extract 22 features by calculating their correlation, homogeneity, pixel regularity, weights, edge direction and edge regularity as follows:

- *Correlation*: This feature explains the percentages of correlations between each relationship with all of other relationships. We define the correlation expression as follows:

$$\text{Correlation}(\theta) = EDM_1(x,y) \Big/ \left(\sum EDM_1(x,y)\right) + EDM_1(2,2), \quad (9)$$

where θ represents 0°, 45°, 90° and 135°, and (x,y) presents the relative position in EDM_1.

- *Homogeneity*: This feature indicates the percentages of distribution of relationships in the image and computes the percentage of each relationship number as follows:

$$\text{Homogeneity}(\theta) = EDM_1(x,y) \Big/ \left(\sum EDM_1(x,y)\right), \quad (10)$$

where θ represents 0°, 45°, 90° and 135°, (x,y) presents the relative position in EDM_1 and (x,y) presents the relative position in EDM_1.

- *Pixel regularity*: This feature explains the percentages of directions of the relationships and computes the percentage of each relationship compared to all angle relationships. We define the expression of pixel regularity as follows:

$$\text{Regularity}(\theta) = EDM_1(x,y)/EDM_1(2,2), \quad (11)$$

where θ represents 0°, 45°, 90° and 135° and $EDM_1(2,2)$ the number of pixels in the edges image.

- *Weight*: This feature explains the density of black pixels in the image pattern. It calculates the percentage of the number of occurrences of each edge pixel compared to the number of black pixels in the source of the image of Arabic calligraphy. We defined the weight information as follows:

$$\text{Weight} = EDM_1(2,2) \Big/ \sum (Iedge(x,y) = \text{black}), \quad (12)$$

where only 0°, 45°, 90° and 135° edges are considered, $I(x,y)$ presents the original image and EDM_1 (2,2) the occurrences of pixels in the edges image.

- *Edges direction*: This feature explains the visual main direction. It shows the dominant direction of the source image. It calculates the direction by finding a position that has the maximum number of relationships:

$$\text{Edges direction} = \text{Max}(EDM_1\ x,y), \quad (13)$$

where only 0°, 45°, 90° and 135° are considered and x, y presents the relative position in EDM_1.

- *Edges regularity*: This measure shows the directions of each pixel's relationships. These relationships represent the percentages for all directions in EDM_2 compared to the total number of edge pixels as follows:

$$\text{Edges regularity}(\theta^*) = EDM_2(x,y)/EDM_2(2,2), \quad (14)$$

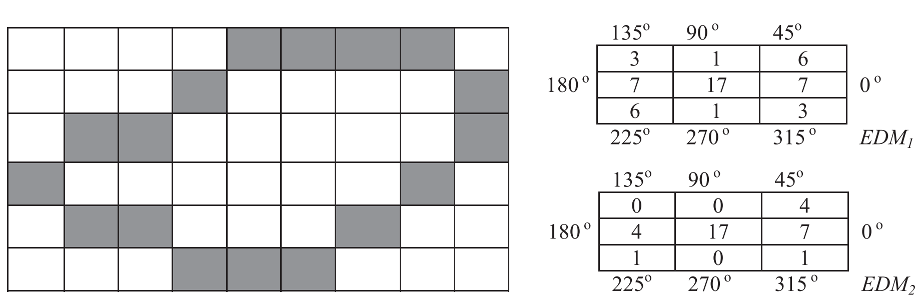
where θ represents 0°, 45°, 90°, 135°,180°, 225°, 270° and 135° and EDM_2 2, 2 the total number of edge pixels in the image.

4. Experiments and results

The samples for the dataset were collected from many sources, such as books, documents, artistic works, calligraphy software products and the internet. We collected 700 image samples, which consist of 100 samples from each type: Kufi, Diwani, Persian, Roqaa, Thuluth and Naskh. We merged Old Kufi with Kufi because Old Kuf has become extinct and the available samples were insufficient for this experiment. The various properties were determined in the preprocessing stage. The final image results are in bitmap format in a 512 × 512 pixel image size. The objective of this paper is to propose an enhanced global statistical feature extraction method and to find the best classifier performance for this application. Therefore, we applied artificial immune systems (AIS), a Bayesian network, decision table rules, a decision tree, and multilayer network classifiers.

We evaluated the performance of the proposed method by comparing it with the gray-level co-occurrence matrix (GLCM) method proposed by Haralick et al. (1973). The GLCM measures angular second moment (ASM); contrast, correlation, entropy, homogeneity and variance. The method was implemented at angles of 0°, 45°, 90° and 135°, which resulted in about 24 features, as shown in Table 3.

The dataset was split into training and testing datasets. In our experiment, we tested a training dataset taken from a dataset of percentages between 60% and 70%. Based on our results, the proposed method obtains higher correct rates than the GLCM in



(a) (b)

Fig. 7. (a) is the edge image and (b) is its EDM_1 and EDM_2 results.

Table 3
The features list of the proposed method and the GLCM (Haralick et al.,1973).

Image type	Proposed method Laplacian edge image	GLCM Straight pixels
Angular second moment	0	4
Contrast	0	4
Correlation	4	4
Edges direction	1	0
Edges regularity	8	0
Entropy	0	4
Homogeneity	4	4
Pixel regularity	4	0
Variance	0	4
Weight	1	0
Total	22	24

Table 4
The classification results with 64% and 66% training.

Training dataset	Proposed method (%)		GLCM (Haralick et al.,1973) (%)	
	64%	66%	64%	66%
Bayes network	88.0478	89.0295	70.6349	70.0048
Multilayer Network	94.8207	93.6709	87.6984	86.9748
Decision tree	98.008	94.9367	86.1111	88.2353
Decision table rule	89.243	90.2954	71.4286	71.1016
AIS	84.0637	87.3418	48.8091	52.9328

all experiments. Different percentages of training and testing data sets were tested to determine the best performance. Based on the results, a decision tree with a 66% training dataset obtains the best performance for GLCM. As shown in Table 4, the correction rate of GLCM is 88.2353%. The proposed method achieves the highest performance with the decision tree and a 66% training dataset with an accuracy rate of about 94.9367%. The proposed technique outperformed a 64% training dataset with the decision tree. Based on the results shown in Table 4 and Fig. 8, we noted that the proposed method obtains the highest performance (about 98.008%) whereas the GLCM method obtains 86.1111%.

The decision tree consumed the second least processing time, about 1.81 s, whereas AIS, the decision table rule, the Bayesian network and the multilayer network needed 1.15, 0.25, 0.16 and 15.95 s, respectively. As a result, we note that the decision tree produces an acceptable performance time when looking at the accuracy rate of the results. We continued the experiment by analyzing the consistency of the results of the decision tree classifier using the proposed method and the GLCM with a 64% training dataset. We repeated this experiment five times and the results are shown in Table 5 and Fig. 9.

Based on the descriptive statistics provided in Table 6, the mean of values of the proposed method is 96.65%, which is higher than the mean for GLCM, which is 84.05%. The proposed method obtains a standard deviation of 0.9595, which is lower than the GLCM method. In relation to the above, the proposed method also produces a smaller standard error value of about 0.42911 compared to GLCM, which is 0.76742.

The correction rate results for each type of calligraphy are shown in Table 7 with a decision tree and a 64% training dataset. The highest accuracy was achieved with the Diwani, Naskh and Thuluth types of calligraphy, which achieve a 100% accuracy rate. Similarly, the lowest accuracy was obtained with the Persian calligraphy sample at 94.9% in GLCM, and the highest accuracy

Table 5
The results of five experiments of the proposed method using a decision tree classifier with 64% training.

	Experiment #1 (%)	Experiment #2 (%)	Experiment #3 (%)	Experiment #4 (%)	Experiment #5 (%)
Proposed	96.4143	95.6175	97.2112	96.0159	98.008
GLCM (Haralick et al., 1973)	82.5397	82.1429	85.3175	86.1111	84.127

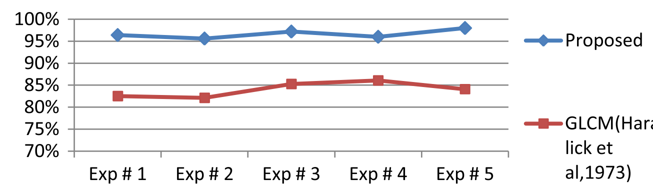


Fig. 9. The results of five experiments using a decision tree classifier with 64% training.

Table 6
The descriptive statistics for the results of the five experiments for the decision tree classifier with 64% training.

	Mean (%)	Standard deviation	Standard error
Proposed	96.65	0.9595	0.42911
GLCM (Haralick et al., 1973)	84.05	1.716	0.76742

Table 7
The correction rate for each class by proposed method and GLCM using a decision tree classifier with a 64% training dataset.

	Proposed (%)	GLCM (Haralick et al.,1973) (%)
Andalusi	97.1	94.6
Kufi	97.1	71
Diwani	100	97.3
Naskh	100	82.9
Roqaa	96.9	92.3
Persian	94.9	71.9
Thuluth	100	87.8

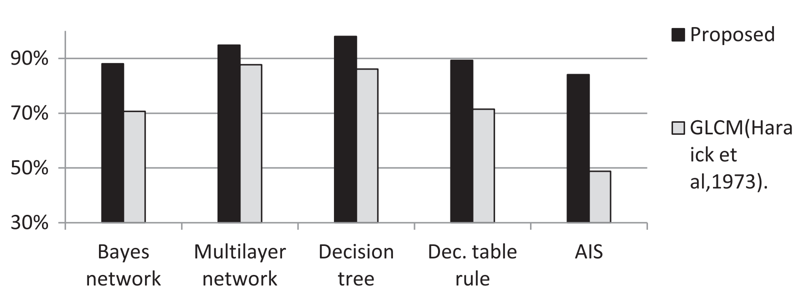


Fig. 8. The graph of the accuracy rate of classifier techniques with 64% training.

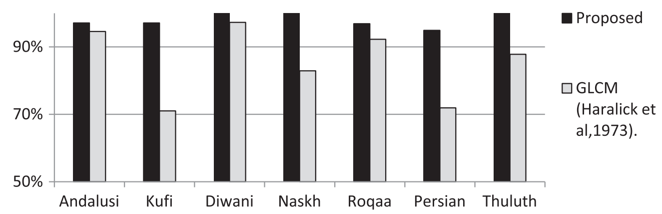


Fig. 10. The correction rate for each class by proposed method, GLCM features and decision tree, with 64% training dataset.

was obtained with the Diwani calligraphy sample at 97.3%. The lowest accuracy rate was obtained with the Kufi calligraphy sample at 71% and the Persian calligraphy sample at 71.9%. Taking these results into account, the proposed method obtained an accuracy rate for each type of calligraphy higher than that for the GLCM. Fig. 10 presents the confusion matrix results of the proposed method and the GLCM.

5. Conclusion

The objective of this paper is to propose an enhanced global feature extraction method based on statistical analysis of the behaviors of edge pixels in binary images. The proposed method as applied Arabic calligraphic script. In summary, after preparing and normalizing the document images, we found a set of values can be presented as the relationships between the edge pixels. These values are stored in related matrices called EDM_1 and EDM_2. Then, by applying the values of these matrices in the proposed algorithms, we obtained about 22 representative features for Arabic calligraphic type. We have compared our method with the GLCM, and we also applied multilayer network, Bayesian network and decision tree classifiers. The proposed method gave higher performance compared with the GLCM. The highest performance of the proposed method was 98.008% obtained with the decision tree on a 64% training dataset. Our global feature extraction method is based on statistical approach and is easy and practical to apply for scripts. The method requires simple preprocessing before execution of the feature extraction process. Furthermore, the proposed method can be used for other language scripts and objects without major modification.

Acknowledgments

We would like to thank previous CAIT researchers such as Jon Timmis at the University of York, UK, and Azuraliza Abu Bakar for offering ideas to improve this research. We also like to thank UKM University for Research Grants UKM-OUP-ICT-36-186/2010 and UKM-GGPM-ICT-119-2010 and FRGS UKM-TT-03-FRG0129-2010.

References

Abdullah, S. N. H. S., Khalid, M., & Yusof, R. (2006). License plate recognition using multi-cluster and multilayer neural networks. In *2nd International conference on information and communication technologies* (pp. 1818–1823).

Abuhaiba, I. (2005). Arabic font recognition using decision trees built from common words. *Journal of Computing and Information Technology – CIT, 13*, 211–223.

Azizi, A., & Pourreza, H. R. (2009). A novel method using contour let to extract features for iris recognition system. Emerging intelligent computing technology and applications. *Lecture Notes in Computer Science*, 544–554.

Busch, A., Boles, W., & Sridharan, S. (2005). Texture for Script Identification. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 1720–1732.

Castro, L. N. D., & Timmis, J. (2002). *Artificial immune systems: A new computational intelligence approach*. Springer, pp. 57–58.

Chen, T., Chen, Y., & Chien, S. (2009). Fast image segmentation and texture feature extraction for image retrieval. In *IEEE 12th international conference on computer vision workshops, ICCV Workshops, Kyoto* (pp. 854–861).

Daniel, J., Andrés, P.-U., Héctor, S., Miguel, B., Patrick, V., & Marco, T. (2008). A survey of artificial neural network-based modeling in agroecology. In B. Prasad (Ed.), *Soft computing applications in industry* (Vol. 226, pp. 247–269). Heidelberg: Springer Berlin.

Ding, X., Chen, L., & Wu, T. (2007). Character independent font recognition on a single chinese character. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 195–204.

Haralick, R. M., Shanmugam, K., & Dinstein, I. (1973). Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, 610–621.

Holmes, D. E., & Lakhmi, C. J. (Eds.). (2005). *Innovations in Bayesian networks. Mathematical & computational methods. Studies in computational intelligence*. Springer.

Jiang, X. (2009). Feature extraction for image recognition and computer vision. In *2nd IEEE international conference on computer science and information technology* (pp. 1–13).

Joshi, G., Garg, S., & Sivaswamy, J. (2007). A generalized framework for script identification. *International Journal on Document Analysis and Recognition*, 55–68.

Journet, N., Ramel, J. Y., Mullot, R., & Eglin, V. (2008). Document image characterization using a multiresolution analysis of the texture: application to old documents. *International Journal on Document Analysis and Recognition*, 9–18.

Ma, H., & Doermann, D. (2005). Font identification using the grating cell texture operator. In *Proceedings of document recognition and retrieval XII* (pp. 148–156).

Mohamed, M. H., & AbdelSamea, M. M. (2008. An efficient clustering based texture feature extraction for medical image. In *11th International conference on computer and information technology, ICCIT 2008, Khulna* (pp. 88–93).

Najab, A., Khan, I., Arshad, M., & Ahmad, F. (2010). Classification of settlements in satellite images using holistic feature extraction. In: *12th International conference on computer modeling and simulation, Cambridge* (pp. 267–271).

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 62–66.

Pal, U., Wakabayashi, T., & Kimura, F. (2009). Comparative study of Devnagari handwritten character recognition using different feature and classifiers. In *10th International conference on document analysis and recognition, Barcelona* (pp. 1111–1115).

Peake, G. S., & Tan, T. N. (1997). Script and language identification from document images. In *Proceedings of workshop document image analysis* (pp. 10–17).

Petrou, M., & García Sevilla, P. (2006). *Image processing. Dealing with texture*. Chichester, UK: John Wiley & Sons, Ltd.

Quan, Y., Jinye, P., & Yulong, L. (2009). Chinese sign language recognition based on gray-level co-occurrence matrix and other multi-features fusion. In *4th IEEE conference on industrial electronics and applications, ICIEA 2009, Xi'an* (pp. 1569–1572).

Ramanathan, R., Soman, K. P., Thaneshwaran, L., Viknesh, V., Arunkumar, T., & Yuvaraj, P. (2009a). A novel technique for english font recognition using support vector machines. In *International conference on advances in recent technologies in communication and computing* (pp. 766–769), Kottayam, Kerala.

Ramanathan, R., Soman, K. P., Thaneshwaran, L., Viknesh, V., Arunkumar, T., & Yuvaraj, P. (2009b). Tamil font recognition using gabor filters and support vector machines. In *International conference on advances in computing, control, and telecommunication technologies* (pp. 613–615), Kottayam, Kerala.

Rokach, L., & Maimon, O. (2005). Top-down induction of decision trees classifiers – A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 476–487.

Sellahewa, H., & Jassim, S. A. (2010). Image-quality-based adaptive face recognition. *IEEE Transactions on Instrumentation and Measurement*, 805–813.

Shahabi, F., & Rahmati, M. (2006). Comparison of Gabor-based features for writer identification of Farsi/Arabic handwriting. In *Tenth international workshop on frontiers in handwriting recognition. 10th IWFHR* (pp. 545–550).

Singh, C., Bhatia, N., & Kaur, A. (2008). Hough transform based fast skew detection and accurate skew correction methods. *Pattern Recognition*, 3528–3546.

Sun, H. M. (2006). Multi-linguistic optical font recognition using stroke templates. In *The 18th international conference on pattern recognition (ICPR'06), Hong Kong* (pp. 889–892).

Tuceryan, M., & Jain, A. (1999) (pp. 207–248). *Texture analysis. The handbook of pattern recognition and computer vision* (2nd ed.,). World Scientific Publishing Co.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques* (2nd ed.). San Francisco: Morgan Kaufmann Publishers.

Yazdi, M., Yazdi, M., & Gheysari, K. (2008). A new approach for the fingerprint classification based on gray-level co-occurrence matrix. In *Proceedings of world academy of science, engineering and technology* (pp. 313–316).

Zhu, Y., Tan, T., & Wang, Y. (2001). Font recognition based on global texture analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1192–1200.

Zramdini, A., & Ingold, R. (1998). Optical font recognition using typographical features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 877–882.