A Statistical Global Feature Extraction Method for Optical Font Recognition

Bilal Bataineh[1], Siti Norul Huda Sheikh Abdullah[2], and Khairudin Omar[3]

Center for Artificial Intelligence Technology,
Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia, 43600, Bangi, Selangor, Malaysia
[1] bilal_bataineh82@yahoo.com, [2] mimi@ftsm.ukm.my, [3] ko@ftsm.ukm.my

Abstract. The study of optical font recognition has becoming more popular nowadays. In line to that, global analysis approach is extensively used to identify various font type to classify writer identity. Objective of this paper is to propose an enhanced global analysis method. Based on statistical analysis of edge pixels relationships, a novel method in feature extraction for binary images has proposed. We test the proposed method on Arabic calligraphy script image for optical font recognition application. We classify those images using Multilayer Network, Bayes network and Decision Tree classifiers to identify the Arabic calligraphy type. The experiments results shows that our proposed method has boost up the overall performance of the optical font recognition.

Keywords: Arabic calligraphy script, Font recognition, Global feature extraction, Gray level co-occurrence matrix, Statistical feature extraction.

1 Introduction

The feature extraction process is one of the most critical issue in document analysis techniques such as Optical font recognition (OFR). Generally, the feature extraction technique consists of two categories: global analysis [1][2] and local analysis [3][4]. The global analysis approach stresses on a region of text block (two lines or more) in the text context of geometric form. On the other side, the local analysis is an approach that concerns on disconnected parts in document images such as words or characters. The information about font helps in several applications such as reprinting documents, document characterization and classification, document layout analysis and improvement the multi-font optical character recognition (OCR).

Arabic language is one of international language. The Arabic alphabet has been widely adopted into other languages such as Jawi, Persian, Kurdish, Pashto, Urdu, and Hausa. In Arabic alphabet, there are three types of written forms: the printed, handwritten and calligraphy. The Arabic calligraphy is the oldest form, it is has many types, fixed accuracy standards and written by the specialist calligraphers. There are eight main Arabic calligraphy types: Old Kufi, Kufi, Thuluth, Naskh, Roqaa, Diwani, Persian and Andalusi. Fig. 1(a) until (h) show examples of one sentence written by calligraphy types.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 257–267, 2011.

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

In global analysis approach, usually the texture analysis techniques have used for extracting feature in document image [5][6]. Basically, the texture analysis techniques are designed for high level image such as grayscale and color images. However, the binary document images consist only two level color images and as far as the research concern, OFR is only interested in foreground text body. That may leads to limitation when applying texture analysis techniques in document images. Quite a number of previous OFR researches were focusing on the Latin [1] or Asian language [2][3]. Until now, there is negligence of the benefits of Arabic calligraphy definition such as classifying the documents layout, the purposes of the documents layout, documents library, the documents history classifying, improving the documents preparing and reprinting the documents as the original input image format.

The objective of this work is to propose an enhanced a statistical feature extraction technique for document images. The proposed method are tested on Arabic Calligraphy Script images as the sample case. The proposed method is compared with Gray Level-Co occurrence Matrix proposed by Haralick *et al.* [7]. Then, we evaluate both methods using Multilayer Network, Bayes network and Decision Tree classifiers to obtain the best performance for optical font recognition application. This paper is organized as follows. Section 2 reviews the state of art in previous OFR. Section 3 explains the proposed method and Section 4 presents and analyses the experimental results. Finally, conclusions are presented in Section 5.

2 State of the Art

Feature extraction is a process is between preprocessing and classification phases. These phase have used in several document analysis techniques. Optical font recognition OFR is one of the document analysis that require this stage.

2.1 Previous Work in OFR

Usually, the feature extraction has two categories: local analysis and global analysis, our interest is more on global analysis. The global approaches identify the text font by analyzing generated blocks of text from a document image.

At earlier, Zramdini and Ingold [1] presented a method to identify the weight, size, typeface, and slope of the printed English text image block. It employed a statistical

approach based on global typographical features using vertical and horizontal projection profile, the character main strokes and the connected components by a rectangular shape enveloping the connected components. A multivariate Bayesian classifier was used to recognize in a set of 280 font, size and style. The average accuracy rate of the font recognition was 95.6% after tested on high quality image text blocks. Some recognition process failed with some font types such as Lucida-Sans and Times fonts. Also, it is hardly to recognize the font for a short text.

Zhu *et al.* [2] described a new texture analysis approach to handle the font recognition for Chinese and English fonts. It employs the text block as a texture block where each font has a special texture. The 2-D Gabor filters was used to extract the texture features. The experiment implemented on four Chinese and eight English fonts. They achieved about 99.1% of average accuracy rate with machine documents with high quality. It ignored the scanned documents with low quality. This method obtained less recognition rate when applying a single character for each font.

In general, the local analysis approach suitable for a single dataset. It requires a modifications if other languages are included. It usually does not generalize on different languages. It strongly depends on the segmentation process and the affected noise. The global analysis approach can easily generalize on different languages. It doesn't require any basic modification if any changes or any additional samples are made in the dataset. Furthermore, only few researches focused on global feature extraction approach. Besides that, some research focused on Latin and Chinese languages, whereby other languages have their own significance in our live.

2.2 Global Feature Extraction Approach

The global feature extraction approach is branched into several categories[6,8]. The statistical method is an example of basic category which defines the texture of the images based on spatial distributions of gray level value in an image. It classifies based on statistical orders whereby the first-order static finds the value of each level and extracts the properties based on those values. Whilst, the second-order static finds the value by relating two gray-levels values with some geometrical relationship and indicates them as the important features. Apart from that, the third or higher order statics depend on finding a values of compound properties of the image [9].

The second-order statistics represent by The Gray Level Co-occurrence Matrix (GLCM) [7] and Gray-Level Difference Method (GLDM)[11]. In general, both methods are similar, but the GLCM is the most popular and use statistical method in the texture feature extraction [8,9]. GLCM has been proposed by Haralick *et al.* [7]. It is a matrix gives values explain the occurrences distribution in the image. The occurrences present the number of pair of two pixels of grayscale values are related with specific relationship. If C is a GLCM, I is an n × m image, $(\Delta x,\Delta y)$ denotes the value of the pairs of pixels that have the gray level value of i and j, the formula shown as following:-

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

Haralick *et al.* [7] has proposed a several equations use to compute a texture features from the GLCM. The GLCM has used for documents analysis research that

depending on global analysis approach research such as scripts and languages identification [11,12]. It has been used in other domains such as Fingerprint Classification [13] and Chinese Sign Language Recognition [14]. However, the statistical methods provide a statistical description of the image gray-levels. With binary images, we are dealing with two levels only that lead to reduce the effectiveness of the features and to some confusion.

3 The Proposed Method

We divide the framework of Arabic Calligraphy font recognition system into two parts: preprocessing and post-processing modules. In the pre-processing, besides generating texture blocks of the predetermined text, we also include edge deduction process [15]. Whilst the post processing involves two sub processes such as feature extraction using our proposed algorithm based on statistical method and recognition sub processes. Fig. 2 shows the flowchart of the proposed method where it requires the following steps: preprocessing and edge deduction, feature extraction and recognition.

The binary images representing documents or models contain two levels such as black and white levels. Our proposed method concerns all values within closed edge image. The edge image gives a clear contour representation of model type. It represents the simplest representation of any form such as the relationship between angle and adjacent pixels. This approach keeps the information of the shape and removes all the unwanted information that adversely affects the values of features. Unlike the previous techniques such as GLCM, they keep only relationship of the actual pixels value. The proposed statistical method is easy to implement. It can easily generalize on different of image types and less prone to noise.

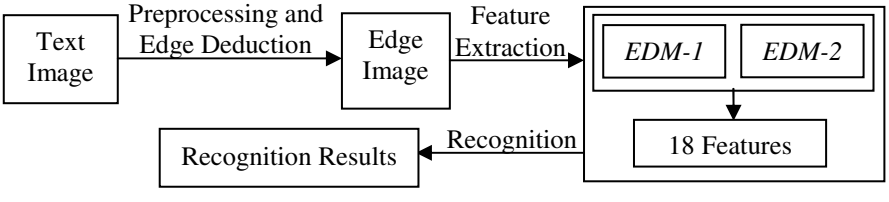
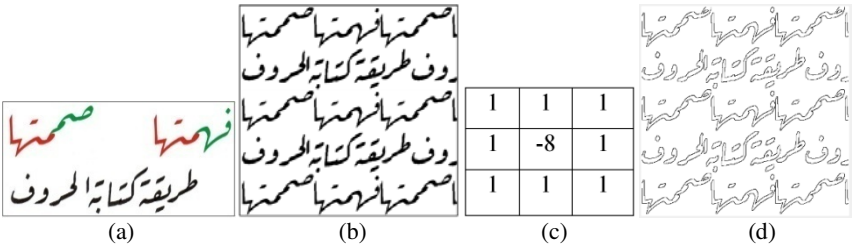


Fig. 2. The flow chart of the proposed Arabic Calligraphy font recognition system

3.1 Preprocessing and Edge Deduction

Usually, the input images have different properties. For that, we apply some methods to overcome this problem. In this stage, we prepare the image using binarization, skew correction, and text normalization subsequently. We obtain the binarization threshold value by using fixed global thresholding method. Then, we use Hough Transform [16] to determine the skew angle in the skew deduction sub-phase. In the text normalization sub-phase, 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(Fig. 3(b)). Lastly, we apply Laplacian filter with 3×3 kernel matrix (Fig. 3(c)), because it is a powerful technique to deduct edges in all directions and it is also effective to solve salt and pepper noise. Fig. 3(d), shows an image after applying Laplacian filter. It represent the final result of preprocessing, it is a 512×512 edge text block image with 96 DPI.



(a) (b) (c) (d)

Fig. 3. (a) The original Image, (b) the image results after generating texture blocks, (c) Laplacian Filter value and (d) filtered image by Laplacian

3.2 Features Extraction

We apply an eight neighbouring kernel matrix and associate each pixel according to their two neighbouring pixels. We present a relationship between the Scoped pixel, *S(x, y)* and their neighboring pixels as depicted in Fig. 4(a). We use and transform an encompassing eight pixels into position values as Fig. 4(b). Based on the previous illustration, we introduced our method based on two perspectives: (1) Find the first order relationship, and (2) Find the second order relationship.

(x-1, y-1)	(x, y-1)	(x+1, y-1)
(x-1, y)	*S(x, y)*	(x+1, y)
(x-1, y+1)	(x, y+1)	(x+1, y+1)

(a)

	1	**2**	**3**
1	135°	90°	45°
2	180°	Scoped pixel	0°
3	225°	270°	315°

(b)

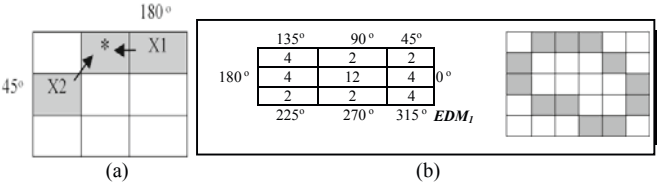
Fig. 4. (a)The eight neighboring pixels, (b) the direction angles of the neighboring pixels, also it represents the edge direction matrix (*EDMS*) and their cells properties

In the first order relationship, we firstly create an 3×3 edge direction matrix (EDM_1) as Fig. 5(b). Each cell in EDM_1 contains a position within 0 until 315 degree value. Secondly, we determine the relationship of the scoped pixel in the edge image $Iedge(x,y)$ by calculating the number of occurrence 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° **then**
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.

In the first order relationship, each pixel in the edge image relates with two pixels. For example, as Fig. 5(a) the scoped pixel presents 180° for X1 and 45° for X2. That means each pixel presents a two relationships in (EDM_1).



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

(c)

Fig. 5. (a) The two neighboring pixels, (b) the edge image and its EDM_1, (c) the order of the angle's importance

In the second order relationship, each pixel will present by one relationship only. We firstly create an 3×3 edge direction matrix (EDM_2). Secondly, we determine the relationship importance for *Iedge(x, y)* by sorting the values in EDM_1 descendingly as shown in Fig. 5(b) and (c) respectively. We take the most importance relationship of the scoped pixel in *Iedge(x, y)* by calculating the number of occurrence for each value in EDM_2. The relationship orders must follow as below:-

(i) If there are more than one angle which have the same number of occurrence, then the smaller angle, is selected firstly.
(ii) Next, the reversal angle is selected subsequently.

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

Step 1: Sort descendingly the relationships in $EDM_1(x, y)$.
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 two neigbouring pixels,
Step 5: Compare the relationship values between two available relationships,
Step 6: Increase 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 as shown in Fig.6.

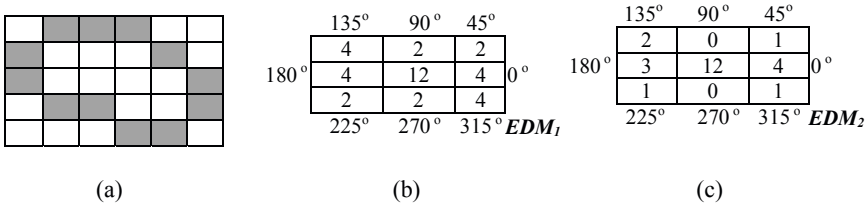


Fig. 6. (a) is a sample image, (b) is its EDM_1 and (c) is its EDM_2 values results

Lastly, we proposed several features from the EDM_1 and EDM_2 values. We summarize 18 features by calculating their homogeneity, pixel regularity, weights, edge direction and edge regularity as the followings:

- **Edges Direction:** This feature represents the main direction of the model, It is calculated by finding the largest value in EDM_1 as follows:

$$Edges\ Direction = \mathrm{Max}(\mathrm{EDM}_1(x, y)) \tag{2}$$

- **Homogeneity:** This feature represent the percentages of each direction to all available directions in the edge image as follows:

$$Homogeneity(\theta) = \mathrm{EDM}_1(x, y)/(\sum \mathrm{EDM}_1(x, y)) \tag{3}$$

- **Weight:** This feature represents the density of the model in the image. It is calculated based on the percentage of edges on the size of the model as follows:

$$Weight = \mathrm{EDM}_1(2,2) / \sum(\mathrm{Iedge}(x, y) = \mathrm{Black}) \tag{4}$$

- **Pixel Regularity:** This feature represent each direction in EDM_1 to the number of scoped pixels in the edge image as follows:

$$Pixel\ Regularity(\theta) = \mathrm{EDM}_1(x, y) / \mathrm{EDM}_1(2,2) \tag{5}$$

- **Edges Regularity:** This feature represent each real direction in EDM_2 to the number of scoped pixels in the edge image as follows:

$$Edges\ Regularity(\theta_*) = \mathrm{EDM}_2(x, y) / \mathrm{EDM}_2(2,2) \tag{6}$$

where θ represents 0°, 45°, 90° and 135°, θ_* presents 0°, 45°, 90°, 135°,180°, 225°, 270° and 135°, and *(x, y)* presents the relative position in EDM_1 and EDM_2.

4 Experiments and Results

The dataset have collected from many resources such as documents, artistic works and calligraphy software products. We have collected 700 samples that consist of 100 samples from each type of Kufi, Diwani, Persian, Roqaa, Thuluth and Naskh. We compare our method with GLCM[7]. In this work, the measures which have been implemented are Contrast, Homogeneity, Angular Second Moment (ASM), Energy, Entropy, Variance and Correlation with 0°, 45°, 90° and 135° angles, that derived to 28 features. Also, we applied Multilayer Network, Bayes network and Decision Tree

classifiers to compare the proposed method performance within different classifiers. The dataset was split into training and testing datasets in percentage between 60% to 70%. Based on our experimental results, the proposed method gives higher performance than the GLCM in all experiments. The experiments performance of each classifier with the different training dataset were convergent, so we chosen the 60% training dataset. Based on the results in Table 1 and Fig. 7, we can note that the proposed method gains the highest performance about 97.85% with the decision tree whereas GLCM method obtains 87.143% with the Multilayer Network.

Table 1. The classification results with 60% training

	Proposed method	**GLCM[7]**
Bayes network	92.473%	77.857%
Multilayer Network	95.341%	87.143%
Decision Tree	97.85%	85.714%

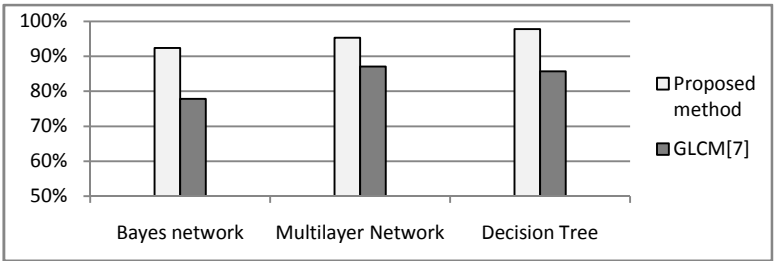


Fig. 7. The classification results of the *GLCM* and proposed method with 60% training

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 60% training dataset. We repeated this experiment five times and the results are shown in Table 2.

Table 2. The results of five experiments using a decision tree classifier with 60% training

	Exp # 1	**Exp # 2**	**Exp # 3**	**Exp # 4**	**Exp # 5**
Proposed	97.85%	97.133%	95.699%	94.624%	96.416%
GLCM [7]	85.714%	79.643%	85.714%	83.929%	84.286%

Based on the experiments values of the Table 2, a statistical descriptive provided in Table 3. The mean of values of the proposed method is 96.3444%, which is higher than the GLCM, which is 83.8572%. The proposed method obtains a standard deviation of 1.25201, which is lower than the GLCM, which is 2.49218. That meant the results of the experiments of the proposed methods is most fixed than the results of GLCM. In relation to the above, the proposed method also produces a smaller standard error value of about 0.55992 compared to GLCM, which is 1.11454.

Table 3. The descriptive statistics for the results of the five experiments for the decision tree classifier with 60% training

	Mean	Standard Deviation	StandardError
Proposed	96.3444%	1.25201	0.55992
GLCM [7]	83.8572%	2.49218	1.11454

Based on the results in Table 4, the correction rates for each calligraphy type shows with decision tree and 60% training dataset. The highest accuracy is achieved on Persian and Thuluth. These calligraphy types achieved by 100% accuracy rate. While in the same case, the lowest accuracy is exhibited with the Diwani about 95.5%. In GLCM, the highest accuracy achieved on the Kufi about 95.2%. The lowest accuracy rate are the Persian about 76.2%. In conclusion, the proposed method has given accuracy rate for each calligraphy type higher than the GLCM. Fig. 8 summarizes the results of proposed method and GLCM respectively.

Table 4. Confusion Matrix results of the proposed approach by using Decision Tree Classifier, 60% training dataset

	Proposed	GLCM [7]
Andalusi	97.4%	92.1%
Kufi	97.9%	95.2%
Diwani	95.5%	86.5%
Naskh	97.5%	78.9%
Roqaa	97%	91.9%
Persian	100%	76.2%
Thuluth	100%	80.4%

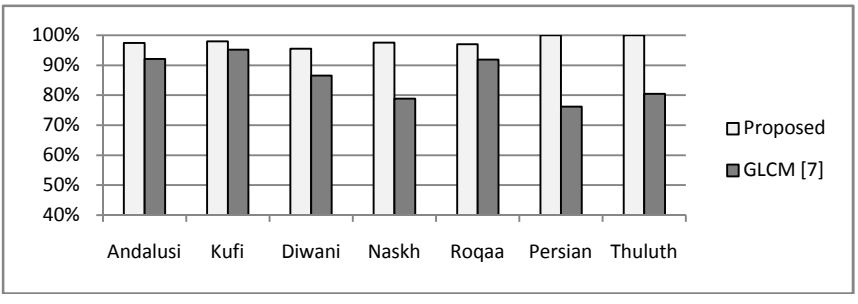


Fig. 8. The correction rate for each class by proposed technique and GLCM features and decision tree, with 60% training dataset

5 Conclusion

In this paper, we propose a global feature extraction method for the binary document images. This method is based on statistical analysis of the relationships between the pixels of an edge image that contains black and white levels as an example. We

applied the proposed method on optical font recognition application to identify the Arabic calligraphy font type. In general, firstly we transform an input image containing texture blocks into closed edge or contour image. Then, we find the values of EDM_1 and EDM_2. In line with that, we obtained 18 features of Arabic calligraphic font type. The proposed method is compared with GLCM on different classifiers. The proposed method outperformed the GLCM method on three chosen classifiers. The highest performance of the proposed method was 97.85% using decision tree on a 60% training dataset. Our method requires simple preprocessing before execution. As conclusion, our proposed global feature extraction method is simple and able to generalize on other image datasets.

Acknowledgments. This research is based on two fundamental research grants from Ministry of Science, Technology and Innovation, Malaysia entitled "Logo and Text Detection for moving object using vision guided" UKM-GGPM-ICT-119-2010 and "Determining adaptive threshold for image segmentation" UKM-TT-03-FRGS0129-2010. We also would like to thank previous the CAIT researchers such as Prof. Dr. Jon Timmis, University of York, UK and Assoc. Prof. Dr. Azuraliza Abu Bakar.

References

1. Zramdini, A., Ingold, R.: Optical Font Recognition Using Typographical Features. IEEE Transactions On Pattern Analysis And Machine Intelligence 20(8), 877–882 (1998)
2. Zhu, Y., Tan, T., Wang, Y.: Font Recognition Based On Global Texture Analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(10), 1192–1200 (2001)
3. Sun, H.-M.: Multi-Linguistic Optical Font Recognition Using Stroke Templates. In: The 18th International Conference on Pattern Recognition, Hong Kong, pp. 889–892 (2006)
4. Ding, X., Chen, L., Wu, T.: Character Independent Font Recognition on a Single Chinese Character. IEEE Transactions on Pattern Analysis And Machine Intelligence 29(2), 195–204 (2007)
5. Joshi, G., Garg, S., Sivaswamy, J.: A generalized framework for script identification. International Journal on Document Analysis and Recognition 10(2), 55–68 (2007); ISSN:1433-2833
6. Tuceryan, M., Jain, A.K.: Texture Analysis. In: Chen, C.H., Pau, L.F., Wang, P.S.P. (eds.) The Handbook of Pattern Recognition and Computer Vision, 2nd edn., ch. 2.1, pp. 207–248. World Scientific Publishing Co., Singapore (1998)
7. Haralick, R.M., Shanmugam, K., Dinstein, I.: Textural Features for Image Classification. IEEE Trans. Systems, Man and Cybernetics 3(6), 610–621 (1974)
8. Petrou, M., García Sevilla, P.: Image Processing, Dealing with Texture. John Wiley & Sons, Ltd., Chichester (2006)
9. Bian, N.: Evaluation of Texture Features For Analysis of Ovarian Follicular Development. Master Thesis, Department of Computer Science. University of Saskatchewan, Saskatoon, Canada (2005)
10. Conners, R.W., Harlow, C.A.: A Theoretical Comparison of Texture Algorithms. IEEE Transactions on Pattern Analysis And Machine Intelligence PAMI-2(3), 204–222 (1980)
11. Busch, A., Boles, W., Sridharan, S.: Texture for Script Identification. IEEE Transactions on Pattern Analysis And Machine Intelligence 27(11), 1720–1732 (2005)

12. Peake, G., Tan, T.: Script and Language Identification from Document Images. In: Proc.Workshop Document Image Analysis, San Juan, Puerto Rico, vol. 1, pp. 10–17 (1997)
13. Yazdi, M., Yazdi, M., Gheysari, K.: A New Approach for the Fingerprint Classification Based on Gray-Level Co-Occurrence Matrix. Proceedings of World Academy of Science, Engineering and Technology 30 (July 2008)
14. Quan, Y., Jinye, P., Yulong, L.: Chinese Sign Language Recognition Based on Gray-Level Co-Occurrence Matrix and Other Multi-features Fusion. In: 4th IEEE Conference Industrial Electronics and Applications, ICIEA 2009, Xi'an, pp. 1569–1572 (2009)
15. Bataineh, B., Abdullah, S.N.H.S., Omer, K.: Generating an Arabic Calligraphy Text Blocks for Global Texture Analysis. In: International Conference on Advanced Science, Engineering and Information Technology (ICASEIT 2011), Kuala Lumpur, Malaysia (January 2011)
16. Singh, C., Bhatia, N., Kaur, A.: Hough transform based fast skew detection and accurate skew correction methods. Pattern Recognition 41(3), 3528–3546 (2008)