Journal of Theoretical and Applied Information Technology
10th March 2013. Vol. 49 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

GEOMETRICAL-MATRIX FEATURE EXTRACTION FOR ON-LINE HANDWRITTEN CHARACTERS RECOGNITION

SAAD M. ISMAIL[1], SITI NORUL HUDA SHEIKH ABDULLAH[2]

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

Email: [1]ababnah2002@yahoo.com, [2]mimi@ftsm.ukm.my

ABSTRACT

Most of Arabic handwriting recognition in the literature has focused only on recognizing offline script, and few of research take online case. So it's still remains as an active area of research. however there is a lack of studies in terms of recognizing Arab characters, especially on the online cases. The process of handwriting recognition faces a lot of challenges; feature extraction is the most important problem in character recognition. The main theme of this paper is new feature extraction method employed in online Arabic character recognition. An Arabic character recognition handwritten system cannot be successful, without using suitable feature extraction methods. In this work we have proposed the hybrid Edge Direction Matrixes and geometrical feature extraction method for on-line handwritten Arabic character recognition system. In addition, horizontal and vertical projection profile, and Laplacian filter have been used in the preprocessing phase**.** The training and testing of the online handwriting recognition system was conducted using our dataset; we have used 840characters from different writers, 504 characters for training, and 336 characters for testing. The evaluation was conducted on state of the art methods in the classification phase. The results have revealed that the proposed method gives best recognition rate for character category.

Keywords: *Online Recognition, Arabic Character, Geometrical Feature, Edge Direction Matrixes, Classification.*

1. INTRODUCTION

Character recognition is a process engulfed with a lot of challenges; feature extraction is the most important problem in character recognition. The performance of character recognition largely depends on, two main decisions such as: the feature extraction approach and the classification scheme. Feature extraction is a crucial part of character recognition. It greatly affects the recognition accuracy, if the features are not suitable for this task [12]. As mentioned in the literature [2]-[3], the feature extraction plays an important role in the overall process of handwriting recognition. Many feature extraction techniques [2][3][5][6][7][8][9] have been proposed to improve overall recognition rates; however, most of them depend on the size and slope of handwriting characters. They require very accurate resizing, slant correction procedure or technique; otherwise they will yield very poor recognition rates. Furthermore, most of existing techniques use only one feature of a handwritten character.

The main problem is encountered, while dealing with Arabic characters written by different persons, where the writers represent the same character differently in terms of size and shape. This variation is due to the individuality of the persons, who write the script, apart from the mood and situation of the writer [4]. The existence of 'dots' in Arabic is the other problem that provokes the difficulty of recognition process, the single, double or triple dots can be placed above or below the letter body. It is common that, Arabic letters have same body, but will differ in dots, which help the readers to identify those characters. Therefore, it is vital to recognize all the component edges and dots in a character. This work has three main steps such

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

as: preprocessing, feature extraction and classification (see figure2).

This research focuses on a new feature extraction technique, which uses several features of a character and combines them to create a hybrid features extraction. The remainder of this paper consists is organized as follows: the section two elucidates the proposed technique; the discussion and analysis of the experimental results are presented in section three; ultimately the section four presents the conclusion and suggestions for the future research.

2. RELATED WORKS

In the past, several researches have been done to solve the problem of Arabic character recognition. Various methods have been proposed based on the global feature extraction approach, of them, the first and foremost was proposed by [7]. They have developed the approach, by hybridizing statistical analyses of edge pixels relationships with geometrical relationship, and tested on Arabic calligraphy script image, for optical font recognition application. Later Naeimizaghiani [14] have presented an enhanced feature extraction method, which is a combination of two selected feature extraction techniques of Gray Level Co occurrence Matrix (GLCM) and Edge Direction Matrixes (EDMS) for character recognition system. The dataset of images that has been applied to the different feature extraction techniques includes the binary character with different sizes. This method was compared with GLCM and EDMS method, after performing the feature selection with neural network, bayes network and decision tree classifiers. Amin & Singh [2], and Amin [3] have presented a technique for the recognizing Arabic words and Chinese characters, with the C4.5 machine learning system. This technique has the following steps: digitization, pre-processing feature extraction, and classification. Mapping for the recognition of on-line handwritten characters was proposed by Khorsheed (Khorsheed 2003). This mapping creates the same output pattern, apart from of the orientation, position, and size of the input pattern. Abdullah et al. [12] have presented, an automatic license plate detection system, based on image processing and clustering. Enhanced geometrical feature topological analysis has been used as the feature extraction technique, while support vector machine has been applied as the classification technique.

Recently [4] have presented approach for on-line Arabic handwritten characters recognition. This approach has utilized structural features and decision tree learning techniques and has three phases: First, the user writes the character on an exclusive window on the screen, and then the coordinates of the pixels forming the character is acquired and saved in a special array. Second, a bounding box of 5x5 is drawn around the character, and the five features are drawn out from the character, which is used in step three for recognizing the character using a decision tree learning techniques. This approach has been tested on a set of 1400 different characters, written by ten users. Each user wrote the 28 Arabic characters five times, in order and the approach has achieved about 75% recognition rate. [15] has proposed an approach for online Arabic handwriting characters recognition. This approach has been based on decision tree and matching algorithm to learn the stroke direction of the Arabic character. In this approach a dataset collected for handwriting samples has been used as training set, and tested on a set of 140 different characters written by five users, each user wrote 28 characters randomly. This approach has achieved 97% recognition rate.

3. THE ARABIC SCRIPT

Arabic alphabets consist of 28 characters. The form of the character is based on its location within a word. In Arabic text, the script alphabet where the successive letters in a word are connected to each other by a baseline is utilized. As reported before to house the baseline, the Arabic alphabet takes the following form: isolated, initial, medial and final. But many letters in the alphabet hardly ever stick to this rule and posses various forms for the medial and final shapes. When one of these non-linked characters is implicated in a word, the preceding letter considers it as final (or isolated) form, and the non-joiner assumes it as initial (or isolated) form (see table 1). In addition, most of Arabic letters have same body but will differ in dots, which are helpful to identify them [1]. In this work we have focused on the isolated character shape.

Table1. The Arabic Isolated Characters

No Dots	One Dot	Tow Dots	Three Dots
ا	ب	ت	ث
ح	ج	ق	ش
د	خ	ي	

Journal of Theoretical and Applied Information Technology
10th March 2013. Vol. 49 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ر	ذ		
س	ز		
ص	ض		
ط	ظ		
ع	غ		
ل	ف		
م	ن		
ه	ك		
و			

4. THE OCR STRUCTURE

The proposed system consist the following steps:

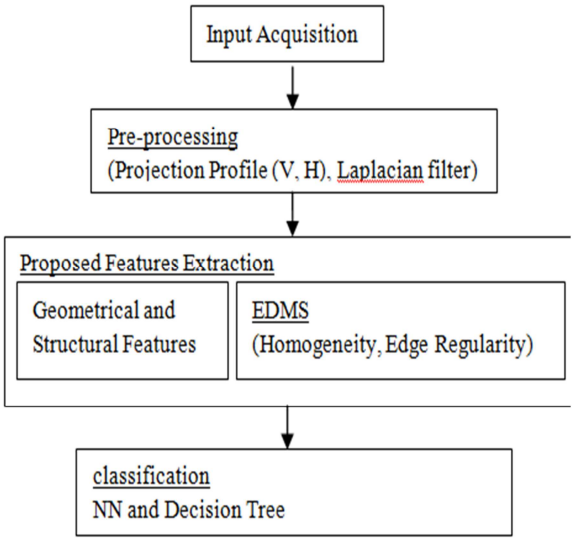


Figure.1 OCR System Structure

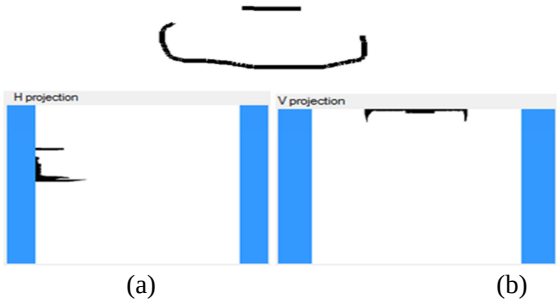
4.1 Preprocessing

To recognize Arabic characters, the preprocessing stages should be applied before the recognition stage. In this work, the preprocessing stage involves in finding the vertical and horizontal projection profile, using the Laplacian filter.

4.1.1 Horizontal and Vertical Projection Profile

Projection profile has been widely used in detecting lines and words[11]. So, the horizontal projection profile is computed to determine the edges that in turn will determine the shape of character and the dots above or below the character. The proposed algorithm uses horizontal and vertical projections profile from different regions, to extract the features from the projection profiles of character (see figure 2).

- Horizontal profile: sum of black pixels perpendicular to the x axis
- Vertical profile: sum of black pixels perpendicular to the y axis



(a) (b)

Figure2. (A) Horizontal Projection And (B) Vertical Projection Profiles For (ت) Character

4.1.2 Laplacian Filter

A Laplacian filter forms another basis for edge detection methods. A Laplacian filter is used to compute the second derivatives of an image, which measures the rate at which the first derivatives change. This helps to determine the changes in adjacent pixel values in an edge.

Kernels of Laplacian filters, usually contain negative values in a cross pattern (similar to a plus sign), which is centered within the array. The corners are either zero or positive values. The center value can be either negative or positive. The following array figure 3 is an example of a 3 X 3 kernel for a Laplacian filter.

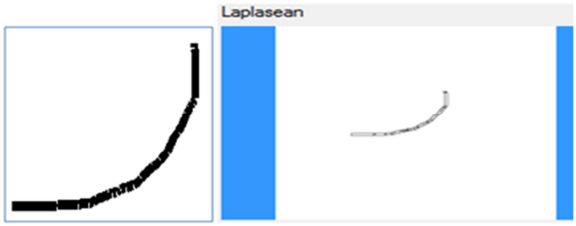
1	1	1
1	8	1
1	1	1

Figure3. A 3 By 3 Kernel For A Laplacian Filter

In this research we have applied Laplacian filter with 3×3 kernel matrix (Figure 3), because, it is a powerful technique to detect edges in all directions and it is also effective to solve salt and pepper noise. The figure 4(a) shows the original character and the figure 4(b) shows a character after applying Laplacian filter. It represents the final result of preprocessing.

Journal of Theoretical and Applied Information Technology
10th March 2013. Vol. 49 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195



(a) (b)

Figure 4. (A) Original Character (B) Filtered Image By Laplacian With 3x3 Kernel Matrixes.

4.2 The Proposed Feature Extraction

In this research two combinational types of features are used as follows:

4.2.1 EDMS Features

To extract the features of Arabic character, statistical analysis technique presented by using edge direction matrixes (EDMS) has been used [7]. Eight neighboring kernel matrices have been applied and associated with each pixel, according to their two neighboring pixels. A relationship has been established between the scoped pixel, S(x, y) and their neighboring pixels as depicted in figure 5(a). The eight pixels are used to transform the encompassing into the position values as shown figure 5(b). Based on the previous illustration, this method has been introduced, depending on two perspectives, such as: Finding the first order relationship, and finding the second order relationship.

	1	2	3
1	(x-1, y-1)	(x, y-1)	(x+1, y-1)
2	(x-1, y)	S(x, y)	(x+1, y)
3	(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)

Figure 5. (A) The Eight Neighboring Pixels, And (B) The Direction Angles Of The Neighboring Pixels

In the first order relationship, initially, the 3×3 edge direction matrix (EDM1) was created as shown in figure 5(b). Each cell in EDM1 contains a position within 0 until 315 degree value. Secondly, the relationship is determined of the scoped pixel in the edge image ledge (x,y) by calculating the number of occurrence for each value in EDM1.

The Algorithm

For each pixel in Iedge (x,y)

If Iedge (x,y) is black pixel at center **then**

Increase number of occurrence at EDM1 (2,2) by 1.

If Iedge (x+1,y) is black pixel at 0° **then**

Increase number of occurrence at EDM1 (2,3) by 1.

If Iedge (x+1,y−1) is black pixel at 45° **then**

Increase number of occurrence at EDM1 (1,3) by 1.

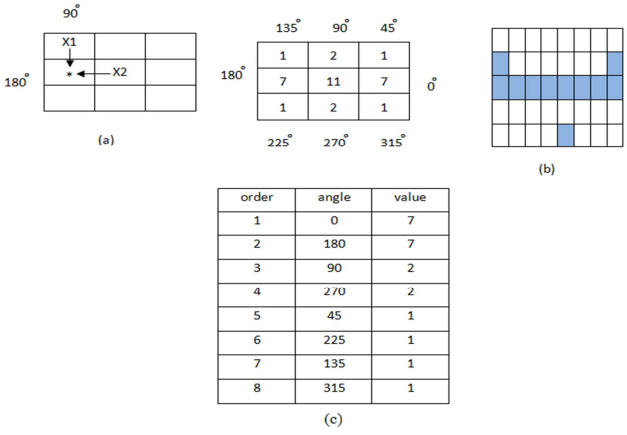
If Iedge (x,y−1) is black pixel at 90° **then**

Increase number of occurrence at EDM1 (1,2) by 1.

If Iedge (x−1,y−1) is black pixel at 135°**then**

Increase number of occurrence at EDM1 (1,1) by 1.

In the first order relationship, each pixel in the edge image relates with two pixels. For example, as shown figure 6(a) the scoped pixel presents 180° for X2 and 45° for X1. This means that, each pixel presents two relationships in (EDM1).



order	angle	value
1	0	7
2	180	7
3	90	2
4	270	2
5	45	1
6	225	1
7	135	1
8	315	1

(c)

Figure 6. (A) The Two Neighboring Pixels, (B) The Edge Image And Its EDM1, (C) The Order Of The Angle's Importance

In the second order relationship, each pixel has been presented by one relationship only. Firstly the 3×3 edge direction matrix (EDM2) has been created. Secondly, the relationship importance for

Journal of Theoretical and Applied Information Technology
10th March 2013. Vol. 49 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Edge (x, y) has been determined by sorting the values in EDM1 descending order as shown in Figure 6(b) and (c) respectively. We have taken the most important relationship of the scoped pixel in Edge(x, y) by calculating the number of occurrences for each value in EDM2. The relationship orders must follow as follows:-

a. If more than one angle has the same number of occurrences, then the smaller angle is selected first.

b. Next, the reversal angle is selected subsequently.

The algorithm of the second order of EDM2 relationship is as follows:

The Algorithm Steps

Step 1: Sort descendingly the relationships in EDM1(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 EDM2(x, y).

Lastly, several features from the EDM1 and EDM2 values are presented. Some features have been summarized by calculating their homogeneity and edges regularity as follows:

Edges Regularity (θ*) = EDM2 (x, y) / EDM2 (2, 2)

where θ represents H0°, H45°, H90° and H135°, θ* presents 0°, 45°, 90°, 135°,180°, 225°,

1- Homogeneity: This feature represents the percentages of each direction, to all available directions in the edge character as follows: Homogeneity (θ) =EDM1(x, y)/(ΣEDM1(x, y)).
2- Edges Regularity**:** This feature represents each real direction in EDM2, to the number of scoped pixels in the edge image as follows:

4.2.2 Geometrical And Structural Features

To extract the features of Arabic character, as well as EDMS feature, we have used geometrical features see figure7. Five features have been applied such as: width and the highest of horizontal and vertical base, width of dot, number of occurrence in (H, V) projection and comparison between the two parts of projection profile.

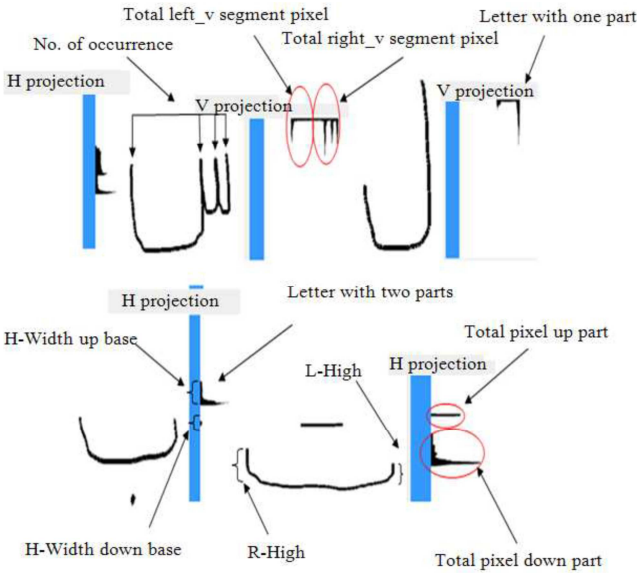


Figure7. Geometrical And Structural Features

4.3 Normalization

Normalization is a process that changes the range of pixel intensity values. Data transformation such as, normalization may improve the accuracy and efficiency of classifiers, like neural networks, and decision tree classifiers. A lot of methods are available for data normalization including, min-max normalization, z-score normalization and normalization by decimal scaling. Such methods provide better results, if the data to be analyzed have been normalized, that is, scaled to specific ranges such as, [0.0,1.0]. An attribute is normalized by scaling its values; so that, they fall within a small-specified range, such as, 0.0 to 1.0. In this research the normalization has been applied on dataset for NN and decision tree classifier. The normalization has been calculated based on the following equation:

$$V = \frac{\text{Value}-\text{Max}}{\text{Min}-\text{Max}} \quad (1)$$

Where, V is the output value, the value is the current input value, and Max and Min are the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

maximum value and the minimum value in the range of all input values subsequently.

5. EXPERIMENTAL RESULTS

In these experiments the performance of the proposed feature extraction has been evaluated by comparing it with different methods of feature extraction that have been used in OCR. The Edge Direction matrixes (EDMS) method and the Naeimizaghiani [14] method are used. These methods were used in these experiments, because they have been used in OCR. Based on the literature review, Naeimizaghiani is based on hybrid of GLCM and EDMs. The GLCM and EDMs are among the best statistical global feature extraction methods that have been applied in document images. They have been used in many document analysis techniques, such as, in OFR [7] and language identification [16], writer identification[17]. To find the best performance, the decision tree and multilayer neural network classifiers were applied.

The dataset has been split into training and testing datasets. In this experiment, the training dataset is determined from percentages between 60% and 70%. Based on the experimental results, the proposed method has obtained higher accuracy rates than EDMS and Naeimizaghiani feature extraction methods in all experiments. Different percentages of training and testing data sets have been tested to determine the best performance. Based on the results, a decision tree with a 60% training dataset has obtained the best performance for EDMS was 68.5%. While, the best performance of Naeimizaghiani method was 79.8% in 62% training dataset with decision tree classifier. As shown in Table 2 and Figure 8, the proposed feature extraction method achieves highest performance with the 61% training dataset with neural network about 98.85%, as shown in table 2. Based on the results shown in Figure 7, it is noted that the proposed has method obtained the highest performance than of EDMS and Naeimizaghiani feature extraction method in all experiments. Table 3 presents the standard deviation for the results of five experiments for the EDMS, Naeimizaghiani [14] and the proposed methods, using decision tree classifier and using Arabic characters dataset with 61% training. The proposed method has obtained a standard deviation of (0.4), which was lower than the EDMS method about 6.0, and Naeimizaghiani method about (2.14). Therefore, the features of the proposed method are more effective than the other methods.

Table 2. The classification results of the EDMS, (Naeimizaghiani et al. 2011) and proposed method from 60% to 70% splitting of the training dataset and the classification results of neural network and decision tree.

	decision tree/ proposed	NN/ proposed	decision tree/ EDMS	NN/ EDMS	Decision tree/ (Naeimizaghiani. et al)	NN/(Naeimizaghiani et al)
60%	93.85	97.76	68.5	59.9	79.3	71
61%	96.55	98.85	65.2	62.7	79.4	72.4
62%	95.2	97.6	64.3	59	79.8	72
63%	96.4	96.4	65.7	65	79	67
64%	97.5	98.7	66.4	61.3	77	70.2
65%	94.9	97.4	64.4	64.6	76.76	73.6
66%	94.74	97.36	63.4	61	77.5	72.82
67%	96	99	68	61	79	71
68%	97.2	97.2	65.8	63.1	76.9	72.3
69%	94.2	95.7	65.3	68.6	77.2	69.7
70%	97	97	66.3	63.8	76	70.37

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

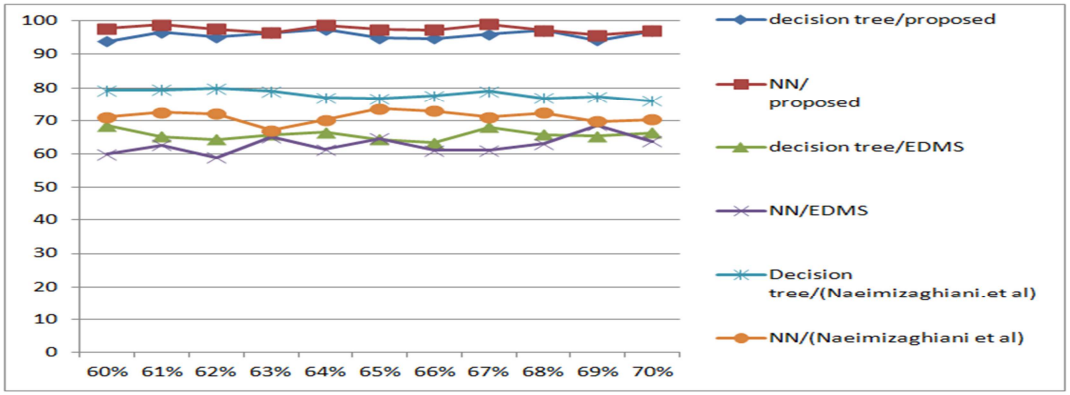


Figure 8. The Classification Results Of The EDMS Bataineh Et Al(Bataineh Et Al. 2011), Naeimizaghiani (Naeimizaghiani Et Al. 2011) And Proposed Methods From 60% To 70% Splitting Of The Training Dataset And The Classification Results Of Neural Network And Decision Tree.

Table 3. The Standard Deviation

Std. dev.	EDMS	(Naeimizaghiani et al.)	Proposed
	6	2.14	0.4

6. CONCLUSION

The main aim proposed in this work was to apply hybrid Edge Direction Matrixes and geometrical feature extraction method for on-line handwritten Arabic character recognition system, to improve the accuracy rate. This had been achieved by a set of steps. This work consists of three main phases: In the pre-processing phase, the Arabic character had been pre-processed based on the common pre-processing methods, such as: vertical and horizontal projection profile and Laplacian filter. In the feature extraction phase, the Geometrical and EDMS features had been proposed. In the recognition phase, we had applied two classification techniques such as: the NN and decision tree. Based on the results obtained, it had proved that the proposed method had produced the best accuracy rate, compared with the EDMS and Naeimizaghiani methods after

applied on NN and decision tree classifier. the proposed approach had produced about 96.7% accuracy rate.

7. ACKNOWLEDGMENT

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-GGPMICT-119-2010 and "Determining adaptive threshold for image segmentation" UKM-TT-03-FRGS0129 2010. We also would like to thank Dr. Bilal Bataineh, who is a member of the research group.

REFERENCES:

[1] Aburas, A. A. & Gumah, M. E. 2008. Arabic handwriting recognition: Challenges and solutions. Information Technology, 2008. ITSim 2008. International Symposium on, hlm. 1-6.

[2] Amin, A. and Singh, S. "Recognition of hand-printed Chinese characters using decision trees/machine learning C4. 5 system." Pattern Analysis & Applications, 1(2): 130-141, 1998.

[3] Amin, A. "Recognition of printed Arabic text based on global features and decision tree learning techniques." Pattern recognition, 1309-1323, 2000.

[4] Al-Taani, A. T. and Al-Haj, S. "Recognition of On-line Arabic Handwritten Characters using Structural Features." Journal of Pattern Recognition Research 5(1), 23-37, 2010.

[5] Bakhtiari, C. "Arabic Online Handwriting Recognition". Citeseer, thesis, 2007.

[6] Biadsy, F., El-Sana, J. and Habash, N.. "Online arabic handwriting recognition using hidden markov models," The 10th International Workshop on Frontiers of Handwriting Recognition. 2006.

[7] Bataineh, B., Abdullah, S. and K. Omar. "A statistical global feature extraction method for optical font recognition." Intelligent Information and Database Systems, 257-267, 2011.

[8] Cheung, A., Bennamoun, M. and Bergmann, N. W. "An Arabic optical character recognition system using recognition-based

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

segmentation." Pattern recognition, 215-233, 2001.

[9] El-Sheikh, T. and El-Taweel, S. "Real-time Arabic handwritten character recognition." Pattern recognition, 1323-1332, 1990.

[10] El-Sheikh, T. S. and Guindi, R. M. "Computer recognition of Arabic cursive scripts." Pattern recognition , 293-302, 1988.

[11] Ha, J., Haralick, R. M. & Phillips, I. T. 1995. Document Page Decomposition by the Bounding-Box Project. Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on, hlm. 1119-1122.

[12] Guo, H. & Zhao J. 2011. Research on FeatureExtraction for Character Recognition of NaXi Pictograph. Journal of Computers 6(5): 947-954.

[13] Khorsheed, M. S. "Recognising handwritten Arabic manuscripts using a single hidden Markov model." Pattern Recognition Letters, 2235-2242, 2003.

[14] Naeimizaghiani, M., Abdullah, S. N. H. S., Bataineh, B. & Pirahansiah, F. 2011. Character Recognition Based on Global Feature Extraction. 1-4.

[15] Omer, M. A. H. and M. Shi Long. "Online Arabic handwriting character recognition using matching algorithm." Computer and Automation Engineering (ICCAE), 2010 The 2nd International Conference, Singapore, Volume 2, 259-262, 2010.

[16] Busch, A., Boles, W. W. & Sridharan, S. 2005. Texture for Script Identification. Pattern Analysis and Machine Intelligence, IEEE Transactions on 27(11): 1720-1732.

[17] Shahabi, F. & Rahmati, M. 2006. Comparison of Gabor-Based Features for Writer Identification of Farsi/Arabic Handwriting.