Printed Arabic Character Recognition Using Variation Method and Discrete Cosine Transform

Hanaa Fathi Mahmood
Computer Science Department / Education College
University Mosul

Received 17 / 10 / 2011 **Accepted** 01 / 02 / 2012

الملخص

ان التمييز الالي للحروف كان موضوع البحث المركز للعقود الماضية، ولان الحروف العربية سواء المكتوبة باليد او المطبوعة ذات طبيعة متصلة لذلك كانت الأعمال على الكتابة العربية قليلة نسبياً مقارنة بالحروف الأخرى. يقدم البحث تقنية جديدة لتمييز الأحرف العربية المطبوعة.

بعد عملية إدخال صور الاحرف العربية تجري عليها عدة عمليات كإزالة الضوضاء باستخدام الفلتر الوسيط وتقسيم النص الى حروف باستخدام المساقط العمودية والأفقية واستخدمت خوارزمية Guo القياسية للتنحيف. واستخدمت طريقة التغيير و طريقة تحويل الجيب تمام المتقطع لاستخلاص صفات الحرف. وللتصنيف استخدمت شبكة radial basis function (RBF) network. بعد إجراء عدة اختبارات أعطت التقنية المستخدمة نتائج جيدة جدا.

Abstract

Automatic character recognition has been the subject of intensive research for almost last decades. Because of the complexity of printed and handwritten Arabic text a little research has been conducted on the automatic recognition of Arabic characters. This research proposed a new technique for recognizing printed Arabic character.

After acquisition Arabic character image a number of preprocessing steps are performed for the digitized image. These steps generally include smoothing by using median filter, the horizontal and vertical histogram profile are used for segmentation and a standard Guo thinning algorithm for thinning, Etc.. Variation Method and Discrete Cosine Transform Method are used for feature extraction. For

classification radial basis function (RBF) network are used. This method performs extremely well. This new technique is able to handle printed Arabic character task efficiently.

1. Introduction

Optical character recognition, usually abbreviated to OCR, is the mechanical or electronic translation of scanned images of handwritten, typewritten or printed text into machine-encoded text. It is widely used to convert books and documents into electronic files, to computerize a record-keeping system in an office, or to publish the text on a website. OCR makes it possible to edit the text, search for a word or phrase, store it more compactly, display or print a copy free of scanning artifacts, and apply techniques such as machine translation, text-to-speech and text mining to it. OCR is a field of research in pattern recognition, artificial intelligence and computer vision. [Wikipedia 2011]

The ultimate objective of any Optical Character Recognition (OCR) system is to simulate the human reading capabilities. That is why OCR systems are considered a branch of artificial intelligence and a branch of computer vision [Sunji Mori 1999] as well Character recognition has received a lot of attention and success for Latin and Chinese based languages, but this is not the case for Arabic and Arabic-like languages such as Urdu, Persian, Jawi, Pishtu and others [Abdelmalek 2004]. Researchers classify OCR problem into two domains. One deals with the image of the character after it is input to the system by, for instant, scanning in which is called Off-line recognition. The other has different input way, where the writer writes directly to the system using, for example, light pen as a tool of input. This is called On-line recognition. [Aburas 2008]

OCR systems require calibration to read a specific font; early versions needed to be programmed with images of each character, and worked on one font at a time. "Intelligent" systems with a high degree of recognition accuracy for most fonts are now common. Some systems are capable of reproducing formatted output that closely approximates the original scanned page including images, columns and other non-textual components. [Wikipedia 2011]

The potential of OCR systems is enormous because they enable users to harness the power of computers to access printed documents. OCR is already being used widely in the legal profession, where searches that once required hours or days can now be accomplished in a few seconds. [Wikipedia 2011]

2. Previous Work:

There are many offline OCR systems available for printed Arabic documents. Khelifi B. and Zaghden N proposed work to find similar text

regions basing on their fonts. They are extracted text regions, and then font matching is performed using fractal descriptors(box counting). Experiments are done for both maps and ancient documents.[Khelifi 2008]

Tang, Yuan Y., Tao, Yu, Tao, Jin, and Xi, Dihuav [Tang 1999] present method of feature extraction based on the principles of fractal geometry(box-counting approach) and wavelet to classify isolated Chinese character.

AlKhateeb et al. use DCT features and neural network classifier. They discard 80% of the DCT coefficients without sacrificing the recognition accuracy [AlKhateeb 2008].

Varieties of different approach have been applied for the recognition process of OCR.

In the proposed research, two feature extraction techniques were investigated for Arabic cursive character recognition. which is never use for Arabic OCR. The first is the variation method and the second is the discreet cosine transform technique. The output of each feature extraction technique was tested using Radial Basis Function (RBF) classifiers. The flow chart for new approach produced in this research is shown in Figure(1).

3. Characteristics of Arabic text:

Arabic is a popular script. It is estimated that there are more than one billion Arabic script users in the world. If OCR systems are available for Arabic characters, they will have a great commercial value. However, due to the cursive nature of Arabic script, the development of Arabic OCR systems involves many technical problems, especially in the segmentation stage.[Al-A'ali 2007].

The Arabic alphabet consists of 28 characters, where the shape of each character depends on its position within a word. Thus the characters are divided into four disjoint sets. These types are listed in Table 1 in details. The first set includes those characters which appear in an isolated form wherever their position are in different words. The second set includes characters at the head of words, naming beginning characters. The third type includes the characters within words naming middle characters. Finally, the last type includes those characters at the tail of words naming end characters. Thus after segmenting a given word, it will be known a priori which character set needs to be considered [Jannoud 2007] [Al-Zoubady 2006].

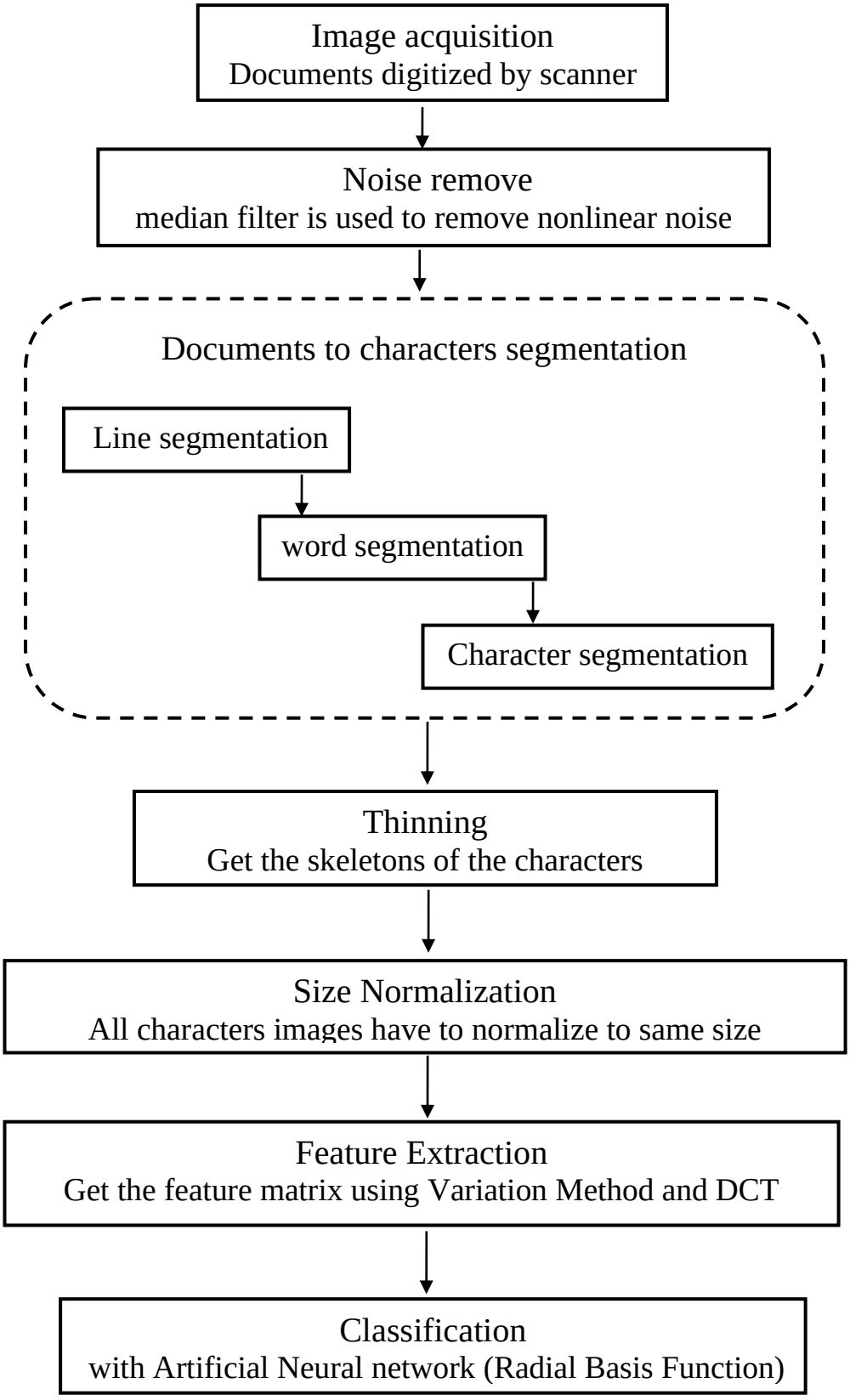


Figure (1): The steps of the proposed algorithm

Table (1): The basic alphabets of Arabic character and their shapes at different positions in the word.

Name	Isolated	Started	Middle	End
Alif	ا	ا	ـا	ـا
Ba	ب	بـ	ـبـ	ـب
Ta	ت	تـ	ـتـ	ـت
Tha	ث	ثـ	ـثـ	ـث
Jeem	ج	جـ	ـجـ	ـج
Hha	ح	حـ	ـحـ	ـح
Kha	خ	خـ	ـخـ	ـخ
Dal	د			ـد
Thal	ذ			ـذ
Ra	ر			ـر
Zay	ز			ـز
Seen	س	سـ	ـسـ	ـس
Sheen	ش	شـ	ـشـ	ـش
Sad	ص	صـ	ـصـ	ـص
Dhad	ض	ضـ	ـضـ	ـض
Tta	ط	طـ	ـطـ	ـط
Za	ظ	ظـ	ـظـ	ـظ
Ain	ع	عـ	ـعـ	ـع
Gain	غ	غـ	ـغـ	ـغ
Fa	ف	فـ	ـفـ	ـف
Qaf	ق	قـ	ـقـ	ـق
Kaf	ك	كـ	ـكـ	ـك
Lam	ل	لـ	ـلـ	ـل
Meem	م	مـ	ـمـ	ـم
Noon	ن	نـ	ـنـ	ـن
Ha	ه	هـ	ـهـ	ـه
Waow	و			ـو
ya	ي	يـ	ـيـ	ـي

4. Data Acquisition and Preprocessing:

The text is scanned off-line from the input document by a scanning device and is stored as a portable grey map (PGM) format file with a resolution of 300 dpi. Then, a number of preprocessing steps are performed for the digitized image. These steps generally include smoothing, segmentation thinning,. etc. Noise errors caused by the data

acquisition system, needs to be eliminated from the scanned document, median filter is used to remove nonlinear noise such, which. A 3*3 window is used to examine each pixel.

4.1 Segmentation

Segmentation is a necessary step in order to isolate the text image objects which will be passed to the recognition stage for recognize characters correctly, the image must be segmented to set of images which only contain one character. These character images will be passed to the OCR module for recognizing. This is accomplished by examining the horizontal histogram profile. Line separation is usually followed by a procedure that separates the text line into words, and in to characters see figure (2). It focuses on identifying physical gaps using only the components. Then the outer rectangle of the character image must be found. Outer rectangle is a rectangle with the least size that all pixels of character are in it. The outer rectangle can be found using horizontal and vertical projection of image.

Figure (2): character segmentation

4.2 Thinning

Thinning is a morphological operation that successively erodes away the foreground pixels until they are one pixel wide or skeletons. A standard thinning algorithm [Guo 1989] is employed, The obtained shape is of one-pixel width with continuous lines carrying the important feature points of the script image. see figure(3).

The last preprocessing step is size normalization. It is the most important preprocessing phase that affect recognition rate directly [George 2002]. all character images have to normalize to 64 x 64 pixels.

Figure (3): Thinning character

5. Feature Extraction

Feature extraction addresses the problem of finding the most compact and informative set of features, to improve the efficiency or data storage and processing. Defining feature vectors remains the most common and convenient means of data representation for classification and regression problems. Data can then be stored in simple tables (lines representing "entries", "data points, "samples", or "patterns", and columns representing "features"). Each feature results from a quantitative or qualitative measurement, it is an "attribute" or a "variable". Modern feature extraction methodology is driven by the size of the data tables, which is ever increasing as data storage becomes more and more efficient.[Switzerland 2006].

6. Characters Features Extraction

This approach propose two method for feature extraction which is not used before for Arabic character recognition this method is illustrated below:

6.1 Variation Method

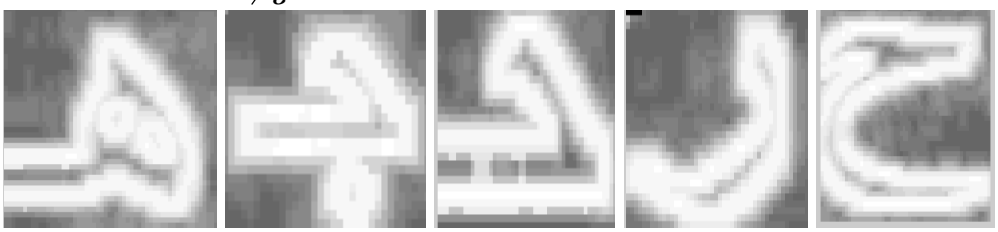
In this research the variation method is consider. This method, provides a means to estimating fractal dimension of an image, where the image intensity or amplitude can be written as a function of a spatial coordinate, Z=f(x,y).

This method requires the choice of various size **ε**. A covering element of radius **ε** is placed at each data point within the image. Next, the variation at each point ***vf(x,y, ε)***, within the element is then measured by taking the difference between the maximum, *zmax*, and minimum,

zmin, functional values contained within the covering elements region. These variations are then averaged over the image to form the **ε** - variation, ***Vf (ε)***, for the image. Note, in the finite case, this is a simple summing operation and, in the analytic or general case, this becomes an integral. The fractal dimension definition then takes the form:

$$D = \lim_{\varepsilon \to 0}\left(3 - \frac{\ln Vf(\varepsilon)}{\ln \varepsilon}\right) \qquad \text{.....(1)}$$

In order to estimate a finite sampled set, one takes the slope of the plot of $\ln V_f(\varepsilon)/\varepsilon^3$ versus $\ln 1/\varepsilon$.[Dubuc 1996][Dubuc 1989][Summers 1999]



Figure(4): variation method features for samples of Arabic characters (ح، ر،حـ، جـ، هـ)

The Variation Method is applied to calculate over each pixel of segmented and normalized characters images see figure (4). The extracted feature for each character image is written into a file in a specific format. This file is the input for a classifier.

6.2 Discrete Cosine Transform

The discrete cosine transform measures the contribution of the cosine function at different discrete frequency. The DCT transformation is applied to image blocks of NxN pixels in size where N is usually 2^N and provides an excellent energy compaction and fast algorithms exist. The fact that the DCT is discrete makes it especially easy for effective computations. The DCT coefficients could be computed using equation

$$f(u,v) = \frac{2}{N} C(u)C(v) \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} f(x,y) \cos\left(\frac{(2x+1)u\pi}{2N}\right) \cos\left(\frac{(2y+1)v\pi}{2N}\right) \qquad \text{.....(2)}$$

Where x and y are spatial coordinates in image blocks, and u and v are coordinates in the DCT coefficient block. The C terms are defined as: [Al-Hamadani 2006] [Sarhan 2009]

$$C(u)C(v) = \begin{cases} 1/\sqrt{2} & \text{for } u,v = 0 \\ 1 & \text{otherwise} \end{cases} \qquad \text{.....(3)}$$

The DCT has been used in many practical applications, especially in signal compression. For example, the compression achieved in the famous JPEG image format is based on the DCT. The strong capability of

the DCT to compress energy makes the DCT a good candidate for pattern recognition applications. Coupled with classification techniques such as Vector Quantization (VQ) and ANN, the DCT can constitute an integral part of a successful pattern recognition system. For example, the DCT was successfully used in face recognition applications. [Sarhan 2009]

To obtain the DC value that is represented by *f(0,0)* of the DCT block. Equation (2) is simplified to equation (4):

$$f(0,0) = \frac{1}{N^2} \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} f(x,y) \qquad \text{.....(4)}$$

The segmented and normalized characters images are divided into several blocks of fixed length (8*8 pixels). Then DCT (Discrete Cosine Transform) applied to calculation over each pixel of the frames.

The result of this DCT step is a set of 64 coefficients for a block of image of size 8x8 pixels. The resultant coefficients were zero and near-zero values. The DC coefficient holds most of the image energy and the average of the 63 coefficients of the block which is known as AC coefficients. The remaining 63 coefficients denote the intensity changes among the block image. In order to achieve a high compaction for the DCT coefficients, these coefficients were eliminated by the quantization operation. The quantization operation sets the near-zero coefficients to zero while sets the other coefficients to a reduced precision. The results of this operation are that the non-zero coefficients were located at the upper-left hand corner of the block and the zeros in the lower corner. These blocks should be changed to a linear form had known as a stream. To maximize the number of subsequent zeros in the stream, the block coefficients is not read line-by-line, but in a zigzagging patterns as shown in Figure (5). [Al-Hamadani 2006]

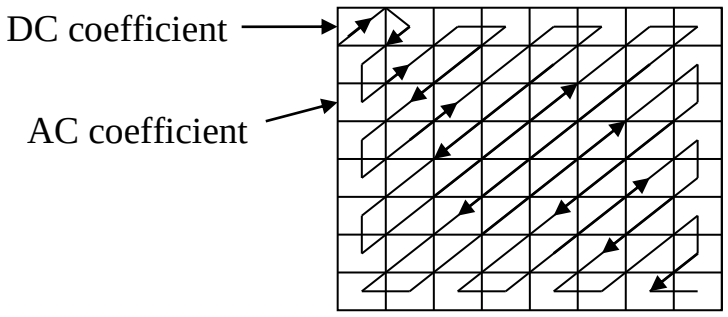


Figure (5): Properties of DCT Coefficients of 8x8 Blocks with Zigzag Pattern

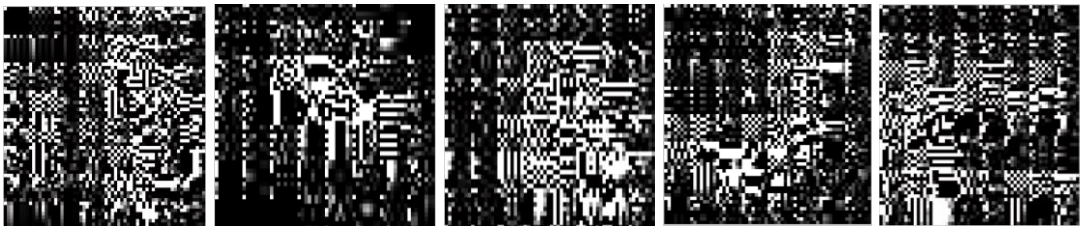


Figure (6): DCT features for samples of Arabic characters (ح، ر،حـ، جـ، هـ)

The DCT Method applied to calculation over each pixel of segmented and normalized characters images see figure (6). The extracted feature for each character image is written into a file in a specific format. This file is the input for a classifier.

7. Classification

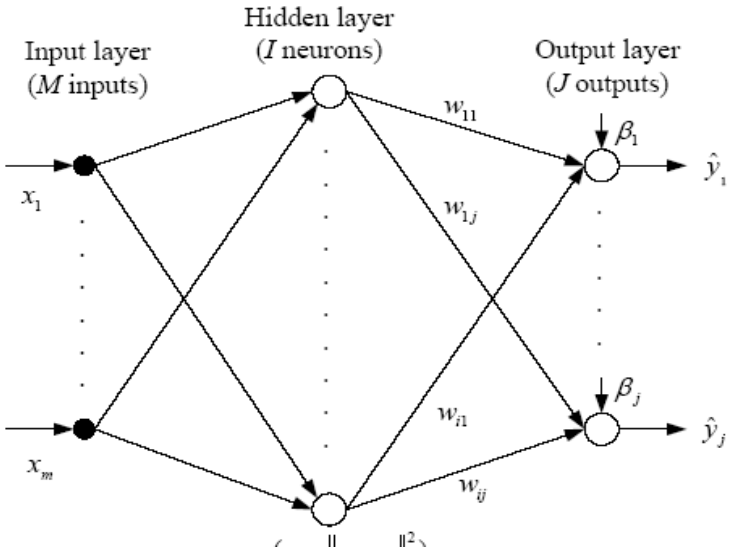
In character recognition, the main task is extraction of features from data. Classification methods are well developed and they generally present low errors if the features are suitable for the task. The classification stage consists of two parts, training and testing. In the training phase, the features of character data are computed and fed to the classifier for training purposes. In the testing phase, features of the unknown input character are extracted. The constructed feature vector is sent to classifier to match the nearest class.

classifiers chosen for this task was Radial Basis function (RBF) network.

7.1 Radial Basis Function Neural Network

A radial basis function (RBF) network is a special type of neural network that uses a radial basis function as its activation function. RBF networks are very popular for function approximation, curve fitting, time series prediction, control and classification problems. The radial basis function network is different from other neural networks, possessing several distinctive features. Because of their universal approximation, more compact topology and faster learning speed, RBF networks have attracted considerable attention and they have been widely applied in many science and engineering fields [Kurban 2009].

Radial basis function (RBF) is a multi-layer neural network consisting of an input layer, hidden layers, and an output layer see figure (7). Nodes in each layer are fully connected to those in the layers above and below, and nodes in hidden layers (basis function nodes) have kernel functions usually given as Gaussian profiles. Each connection is associated with a synaptic weight but the unit weight is assigned to all connections between the input layer and hidden layers. [Ng 1991]



Figure(7): Network architecture of the RBF [Kurban 2009].

The RBF network is trained first by unsupervised learning to determine the characteristics of the hidden layer and then by supervised learning. In the unsupervised learning process, the means and variances of the basis functions for the hidden layers are determined using K-means clustering algorithm [Hush 1993]. The supervised learning process is followed by presenting each input-output pattern to the network and calculating the basis function node outputs. The basis function node outputs and the desired outputs are used to determine the network output weights.

A classifier is used to identify the characters by using their features obtained by applying variation method and DCT. These are then compared and saved as models for the training stage.

8. Result and Discussion

In order to investigate the effectiveness of the proposed algorithm, a series of tests were performed using Radial basis function neural network where Numbers of neurons in input layer equal the number of output for feature extraction of each method used in this research. The output layer contained one node for each class, so the number of neurons in output layer is 28.

Fifty input documents are used which have a hundreds of characters of all types (isolated, beginning, middle and end characters) as a training and testing data with different font type (Arabic Transparent, Simplified Arabic, Arial, Courier) font size varying from 12 to 18 point size. Thirty documents are used for training and twenty documents are used as testing.

Recognition is composed according to the type of font position of the primary part within the word; i.e. Isolated, Beginning, Middle, and End. Both methods (variation method and DCT) were used and compared in terms of the information content of features extracted from the characters with the same ANN structure in the classification stage.

The results in this research are displayed in tabular form for each set of experiments, Table(2,3,4) shows the performance of the new algorithm which proposed in this research when using feature extracted from two feature extracted methods.

Table(2): shows the performance of the proposed algorithm for different font type When using Variation method

Font type	Isolated	Started	Middle	End
Arabic Transparent	96.7 %	90.9 %	87.5 %	87.9 %
Simplified Arabic	99.5 %	93.2 %	91.4 %	91.3 %
Arial	98.7 %	91.6 %	90.4 %	92.3 %
Courier	97.9 %	89.9 %	89.9 %	92.1 %

Table(3): shows the performance of the proposed algorithm for different font type When using DCT

Font type	Isolated	Started	Middle	End
Arabic Transparent	94.6 %	88.7 %	82.9 %	98.8 %
Simplified Arabic	96.2 %	90.4 %	83.5 %	99.5 %
Arial	93.6 %	89.6 %	85.4 %	98.8 %
Courier	97.2 %	85.7 %	79.8 %	96.9 %

Table(4): shows the performance of the proposed algorithm for 2 feature extraction methods and for all type of fonts

Feature extraction method	Isolated	Started	Middle	End
Variation method	98.2 %	91.4 %	89.8 %	91.4 %
Discrete cosine transform	95.4 %	88.6 %	82.9 %	98.5 %

The proposed method was efficient and best results are achieved with the features extracted by Variation method for isolated character the worst result for middle character when using DCT method. The two methods was efficient for isolated and end characters.

9. Conclusion

This research proposed a new method of Printed Arabic Character Recognition where the variation method and discrete cosine transform are used for extract features of normalized characters.

Classification and Normalization of off-line Printed Arabic characters has been proven to be efficient on proposed approach. Significant increase in accuracy levels has been found on comparison of this method with the others for character recognition see table(5). With the addition of sufficient pre processing the approach offers a simple and fast structure for fostering a full OCR system. The experimental results show that the tow methods (variation method and discrete cosine transform) achieved good performance when they are using for printed Arabic character. As observed from the results of research show that both methods gave very good results for the isolated and end characters. Due to the use of size normalization and thinning in preprocessing stage there is no effect of font size on all experiments.

Table (5) Performance analysis for the proposed algorithm with existence ones.

Author	Publish Date	Technique		Reco-gnition
		Feature extraction	Classification	
Majida A.	2011	Genetic Algorithm	Template features	95%
Slimane F. and Kanoun S	2010	Gaussian Mixture Models (GMMs)	Likelihoods with GMMs	99.1%
Sabri A. and Ashraf S.	2009	Fast Hartley transform (FHT)	Nearest neighbor	97%
Zamanifar K. and Izakian H.	2008	Chain code	Nearest neighbor	97.4 %
AlKhateeb	2008	DCT	Neural network	80%
Zheng L.	2006	symmetric map (S-GCM)	minimum distance	97%

The use of Radial Basis Function (RBF) neural network effect good results obtained in this research and This may be attributed to the fact that the Gaussian function in the hidden layer of the RBF network.

In future research, experiments will be using non-resized character images. Also, further experiments will be using non-thinned character image, experiments using another type of font, using handwriting character, using any other type of character like English or Turkish etc., comparison with other kinds of algorithms and other type of neural networks.

10. References

1) Aburas A. A., Rehiel S. A., 2008, JPEG for Arabic Handwritten Character Recognition: Add a Dimension of Application, Advances in Robotics, Automation and Control,, pp. 472, I-Tech, Vienna, Austria.

2) Abdelmalek Z. 2004, ORAN: A Basis for Arabic OCR system, Proceeding of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, Hong Kong, pp. 703-706.

3) Al-A'ali, Mansoor, Ahmad, Jamil, 2007, Optical Character Recognition System for Arabic Text Using Cursive Multi-Directional Approach, Journal of Computer Science 3 (7) pp: 549-555.

4) Al-Hamadani,Israa, 2006, Fast Access Image Retrieval System, A Ph.D. Thesis, College of Computer and Mathematical Sciences University of Mosul, Iraq

5) AlKhateeb J, Ren J, Jiang J, Ipson S, El Abed H., 2008, Word-based handwritten Arabic scripts recognition using DCT features and neural network classifier. Proc 5th Int Multi-Conf on Syst, Signals and Devices.

6) Altuwaijri, M. and M. Bayoumi, 1994, Recognition of Arabic characters using neural networks. ICECS, pp: 720-725. Dec 19-22, Cairo, Egypt.

7) Al-Zoubady, L. Alnsour, A. 2006, Arabic Handwritten Characters Recognized by Neocognitron Artificial Neural Network, University of Sharjah Journal of Pure & Applied Sciences Volume 3, No. 2 p 1-17.

8) Dubuc B. and Dubuc, S, 1996, Error Bounds on the Estimation of Fractal Dimension, Siam J. Numerical Analysis, vol. 33, no. 2, pp. 602-626.

9) Dubuc B., 1989, Evaluating the Fractal Dimension of Profiles, Physics Rev. A, vol. 39, no. 3, pp. 1500-1512,.

10) George D. da C. Cavalcanti, Rodrigo C. Doria, Edson C. de B. C. Filho. 2002. Feature Selection for Off-line Recognition of Different Size Signatures. Proceedings of the 12th IEEE Workshop on Neural Networks for Signal Processing, pp. 355- 364.

11) Guo, Z., and Hall, R. W. 1989, Parallel thinning with two-subiteration algorithms. Communications of the ACM 32, 3 (March), 359–373.

12) Hush L. R. and B. G. Horne 1993, Progress in Supervised Neural Networks - What's New Since Lippmann? IEEE Signal Processing Magazine: 8-39.

13) Jannoud,I. A., 2007. Automatic Arabic Hand Written Text Recognition System, American Journal of Applied Sciences 4 (11) p: 859-866.

14) Khelifi B., Zaghden N., 2008, Unsupervised Categorization of Heterogeneous Text Images Based on Fractals, IEEE.

15) Kurban, Tuba. Beşdok, Erkan, 2009 A Comparison of RBF Neural Network Training Algorithms for Inertial Sensor Based Terrain Classification, Sensors 2009, 9, 6312-6329.

16) Majida A.,2011, Recognition of Different Size Arabic Isolated Characters Using Genetic Algorithm Journal of Applied Sciences Research, 7(6): 907-915.

17) Ng K. and Lippmann R. P. 1991, A Comparative Study of the Practical Characteristics of Neural Network and Conventional Pattern Classifiers. Technical Report 894, Massachusetts Institute Of Technology, Lincoln Laboratory.

18) Sunji Mori, H. Nishida, H. Yamada, 1999, Optical Character Recognition. John Wiley & Sons.

19) Sarhan, Ahmad M., 2009, Iris Recognition Using Discrete Cosine Transform and Artificial Neural Networks, Journal of Computer Science (5): 369-373.

20) Summers R.M., L.M. Pusanik, J.D. Malley, J.M. Hoeg, 1999, Fractal Analysis of Virtual Endoscopy Reconstructions. SPIE MedicalImaging: Physiology and Function from Multidemensional Images, C.-T. Chen and A.V. Clough, eds., vol. 3660, pp. 258-269.

21) Tang, Yuan Y.; Tao, Yu; Tao, Jin; Xi, Dihuav,1999, New method of feature extraction using fractals and wavelets, SPIE Vol. 3715, p. 248-258.

22) Wikipedia Contributors, 2006, Biometrics Information on Wikipedia.com", Wikipedia, The Free Encyclopedia, Cited at: http://en.wikipedia.org/wiki/Biometrics.

23) Sabri A., Ashraf S., 2009, The use of Hartley transform in OCR with application to printed Arabic character recognition, Springer, Pattern Anal Applic 12:353–365.

24) Slimane F., Kanoun S., 2010, Gaussian Mixture Models for Arabic Font Recognition, 2010 International Conference on Pattern Recognition, p:2174-2177.

25) Switzerland, Z. Guyon, I., 2006, Feature Extraction Foundations and Applications. Springer. Netherlands p:1.

26) Zamanifar K., Izakian H., 2008, Multi-Font Farsi/Arabic Isolated Character Recognition Using Chain Codes, World Academy of Science, Engineering and Technology 43.

27) Zheng L., 2006, Machine Printed Arabic Character Recognition Using S-GCM, The 18th International Conference on Pattern Recognition (ICPR'06) IEEE.