# Arabic (Indian) Handwritten Digits Recognition Using Multi feature and KNN Classifier

**Alia Karim Abdul Hassan**
*Computer science, University Of Technology*
110018@uotechnology.edu.iq

## Abstract

This paper presents an Arabic (Indian) handwritten digit recognition system based on combining multi feature extraction methods, such a upper_lower profile, Vertical _ Horizontal projection and Discrete Cosine Transform (DCT) with Standard Deviation σi called (DCT_SD) methods. These features are extracted from the image after dividing it by several blocks. KNN classifier used for classification purpose. This work is tested with the ADBase standard database (Arabic numerals), which consist of 70,000 digits were 700 different writers write it. In proposing system used 60000 digits, images for training phase and 10000 digits, images in testing phase. This work achieved 97.32% recognition Accuracy.

**Keywords:** Feature Extraction, projection profile, vertical_horizontal projection, Discrete cosine Transform Standard Deviation; KNN classifier.

## الخلاصة

تقدم هذه الورقة نظام التعرف على أرقام مكتوبة بخط اليد العربية على أساس الجمع بين أساليب الاستخراج متعددة المزايا، مثل الملف الجانبي العلوي، ورأسية _ الإسقاط الأفقي وتحويل جيب التمام منفصلة مع الانحراف المعياري. يتم استخراج هذه الميزات من الصورة بعد تقسيمها الى عدة كتل. المصنف KNN يستخدم لغرض التصنيف. يتم اختبار هذا العمل مع قاعدة بيانات ADBase القياسية (الأرقام العربية)، والتي تتكون من 70,000 أرقام تم كتابتها من قبل 700 شخص مختلف. في النظام المقترح يستخدم 60000 صورة رقم لمرحلة التدريب و 10000 صورة رقم في مرحلة الاختبار. حقق هذا العمل دقة تعرف على الارقام مقدارها ٪97.32.

**الكلمات المفتاحية:** استخلاص الخواص، الاسقاط العمودي والافقي، المصنف KNN.

## 1. Introduction

The process of transforming the Arabic text which is presented in its spatial form of graphical marks, into its symbolic representation is called as Offline Arabic handwriting recognition [Benouareth *et al.*, 2008]. Off-line handwritten digit recognition in different languages of the world plays a significant role in several applications, such as automated processing of bank checks and automatic sorting of postal mail. Any handwritten digit recognition has several challenges due to the variety of the handwriting style, sizes and orientations of digit samples between different writers [Lawgali, 2016]. The proposed recognition system has three stages, preprocessing, feature extraction and classification recognition stags. The preprocessing stage, tries to remove the noise data, And the feature extraction stage is the process of extracting useful information from the binary handwriting digit image to be used in recognition stage. Finally is classification and recognition stage, which classified all data into 10 classes, then recognize the unknown handwriting digit image to which class it denoted.

## 2. Related Work

Several researchers have developed and design methods for Arabic handwritten digits. In [AlKhateeb and Marwan, 2014] proposed a multiclass classification system used discrete cosine transform (DCT) coefficients approach for feature extraction, then these features are used to train a Dynamic Bayesian Network (DBN) for classification. The proposed system used a ADBase database to evaluate his system, and the results average is 85.26 %. In [Lawgali, 2015] in this work presents a system based on two experiments, one of them used DCT technique on the handwritten digital image to extracted DCT coefficients, which denoted the feature vector. This work tested on ADBase database which containing 70,000 images. The recognition rates for this experiment is 97.25 %. In [Parvez and Mahmoud, 2010] they proposed an approach that finds the polygonal approximation where the direction and length features are extracted from the polygonal approximation. In the recognition stage, they used Fuzzy Attributed Turning Functions and used to define a dissimilarity measure for comparing polygonal shapes. The system is tested on an Arabic numerals database called ADBase database , and the average recognition accuracy was 97.18%.

## 3. The Data set [El-Sherif and Abdleazeem, 2007]

The ADBase standard database (Arabic numerals) composed of 70,000 digits written by 700 persons (writer), each one wrote each digit (from 0 to 9) twenty times. This database collected from different institutions and schools. The written digits were scanned with 300 dpi resolution and adjusted it to produce binary images directly. Figure 1 shows samples of this database. The database is divided into training set which has 60,000 digits and 6000 test set has 10,000 digits when 1000 images per class. The ADBase is available at ('http://datacenter.aucegypt.edu/shazeem') for researchers.

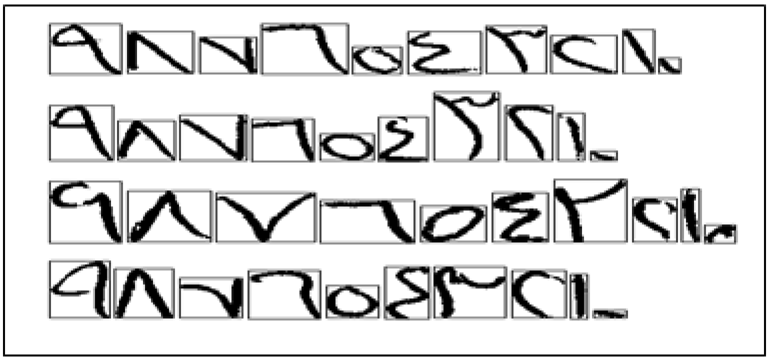


Figure1 Different Sample Images of Arabic Handwritten Digits

## 4. Basic Concepts and Definitions

A set of techniques and methods were used to propose the new recognition method for Arabic word without segmentation are:

**Normalization:** is an important task in the recognition specifically the size normalization which is used to reduce size variation and adjust the image size in order to enhance the recognition process accuracy [Lawgali, 2015].

**Upper and Lower profile:**

**T**he geometrical and topological characteristics of a pattern that represent a type of structural features [Vinciarelli, *et. al.*, 2008]. Upper (or lower) profile is computed by finding the distance (pixel count) of each column from the top (and bottom) of the bounding box of digit to the closest the black pixel in that column [ Sahlol and Suen, 2014].

**Vertical_ Horizontal Projection method:**

**T**he sum of the black pixels perpendicular to the y axis represents the Vertical profile which is computed by scanning the digit column wise along the y-axis and counting the number of black pixels in each column. The horizontal projection profile is the sum of the black pixels perpendicular to the x axis. The digit is traced horizontally along the x-axis. The row wise sum of a number of black pixels presents in each row [Sahlol and Suen, 2014].

## 4.4 Discrete Cosine Transform (DCT)

DCT is a technique which used to convert the image data in the spatial domain into its elementary frequency components in the frequency domain [A. Al-Haj 2007]. The important characteristic of DCT is its ability to convert the energy of the image into a few coefficients. DCT groups coefficient in 2 dimensional array where the coefficients of high value in the upper left corner and coefficients with low value in the bottom right corner [AlKhateeb, *et. al.*, 2008]. DCT frequencies used in the field of pattern recognition when using DCT coefficients as features which becomes efficient in many recognition problems [McLaren, *et. al.*, 2014].

## 4.5 The K Nearesat Neighbors

The K Nearest Neighbor is a classification technique use vectors in a multidimensional feature space, each with a class label as training samples which are stored at the training phase. In the **classification phase**, the distances (the Euclidean distance is more popular) between each training sample and tested sample is calculated. K is a user-defined constant. The *K* training samples that have the smallest distances (nearets) to the test sample are found and identified their labels. By using the majority vote on the neighbor samples will declare the class of the test sample [Hirwani, *et.al.*, 2014, S. Abdleazeem , Ezzat El-Sherif 2008]. The basic steps of KNN as described by [EL Kessab, *et .al.*, 2015]:

**KNN basic steps**

1) Define integer k.
2) Find the distances between the *x test* and xi uses eq.1

$$d(x_{test}, x_j)^2 = \sum_{i=1}^{N} =1 (x_{test,i} - x_{i,j}) \qquad \ldots 1$$

3) Retain k observations with smaller distances
4) Count these k obseravtions in each class, determining the correspondents classes.
5) Choosing the most represented class using eq.2

$$class(X_{test}) = argmax\ X_k \sum_{xj \in KNN} d(\ x_{test}, x_j\ ) \quad \ldots\ 2$$

## 5. The Proposed Arabic(Indian) Hand Written Digits Recognition System

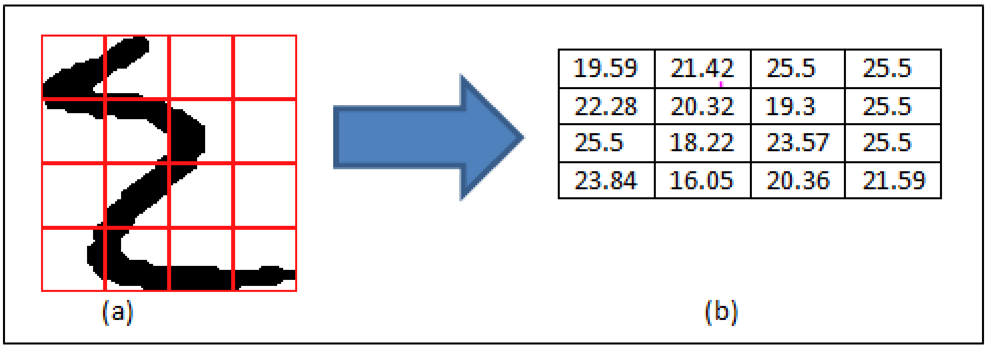
In this work proposed system for recognition Arabic handwritten digits, it includes three stages: are preprocessing, feature extraction, and classification stages. In the preprocessing stage normalized each image to removing the variation in the images. The next stage is feature extraction from normalized image using a hybrid techniques are DCT coefficients, upper_lower profiles and vertical_ horizontal projection methods. In the final stage deciding the Unknown query digit to which class it belongs by applying KNN classifier. Algorithem-1 describe the proposed recognition system main steps.

```
Algorithm-1: Arabic (Indian) Digit Recognition system
Input: digit binary image
Output: digit class C
Step1:Preprocessing by normalizing input images at 32*32 size.
Step2:Feature extraction
   2.1: DCT –SD vector feature
        2.1.1: convert the normalized binary image into a two-dimensional array p.
        2.1.2: divided the image array p into 4*4 blocks each block has 8*8 size .
        2.1.2: calculate DCT coefficients for each block
        2.1.4: find the Standard Deviation for each block
        2.1.5: put theses values in 1 dimensional array which represented the feature vector FV1
   2.2: Upper_lower profile vector feature
        2.2.1: For each column and row in the normalized digital image Calculate the distance from
               the digit boundary box to digit edge.
        2.2.3: each column (or row) denoted a single attribute, put them in single vector which
               called VF2.
   2.3: Vertical –horizontal projection vector feature.
     Count the Black pixel in each row and column to use as the attributes for the third feature
vector VF3.
   2.4: feature vector F={VF1,VF2,VF3}.
Step3: KNN classification
        3.1: set k=4
        3.2: split the data set into two sets, a training set and a test set.
        3.3: training phase store the feature vectors of train set with their class labels.
        3.4: classification phase, calculate the distances between each training vector and tested
             vector.
        3.5: find out The K training vectors which closed (nearest) to the test vector (F).
        3.6: By using the majority vote will declare the class C of the test vector.
 Step4: return (C class of testing digit).
```

In Algorthem-1 step1 represent the **Preprocessing stage in** only the normalization step applied in this stage by normalized all images at 32*32 size. In the ADBase database most of preprocessing steps applied during the development stage such as scanning papers, noise reduction, image binaryzation, segmentation. Step2 is the f**eature extraction stage in which** different methods were to be are DCT, profile projection, and vertical_horizontal projection methods. Where in the first set of features constructed by dividing the normalized digit image into 16 (4*4) blocks.

DCT applied to the each block (8*8 size), then calculate SD (**Standard Deviation** ) value of it. These values are put into a feature vector in order to build the first vector called VF1( Figure 2 shows DCT based feature vector construction process). In the second type of feature extraction method is upper_lower profiled. For each column and row in digital image Calculate the distance from the digit boundary box to digit edge, each column (or row) Denoted a single attribute , then put on a list to build the next vector which called VF2(figure 3 shows upper_lower profile for digit ٩ ) .

The third type of feature extraction method is vertical –horizontal projection. By this method, for each normalized digit image count the Black pixel in each row and column (Figure 4 shows the V_H projection for Arabic digit ٤). Then the all constructed vectors are combined in a single feature vector F. Finally Step3 the classification process by using KNNclassifier with K=4**.** Typically, two-thirds of the data are represent the training set, and the remaining one-third represents the test set. In the **training phase** storing the feature vectors of the training set with their class labels. In the **classification phase**, calculate the distances (Euclidean distance) between each training vector and tested vector. Then find out The *K* training vectors which closed (nearest) to the test vector. By using the majority vote will declare the class of the test vector, when the value k=4 the proposed system has a high accuracy.



| | | | |
|---|---|---|---|
| 19.59 | 21.42 | 25.5 | 25.5 |
| 22.28 | 20.32 | 19.3 | 25.5 |
| 25.5 | 18.22 | 23.57 | 25.5 |
| 23.84 | 16.05 | 20.36 | 21.59 |

**Figure 2: a) 4*4 blocks for digit image b) SD values for all blocks**

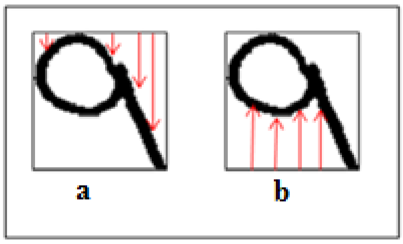


**Figure 3: Profile method a)upper profile b) lower profile**

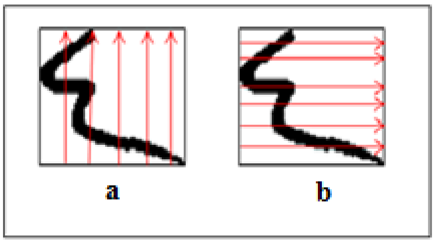


**Figure 4: Projections for arabic digit a)vertical b) horizontal**

## 6. Experimental Results and Discussion

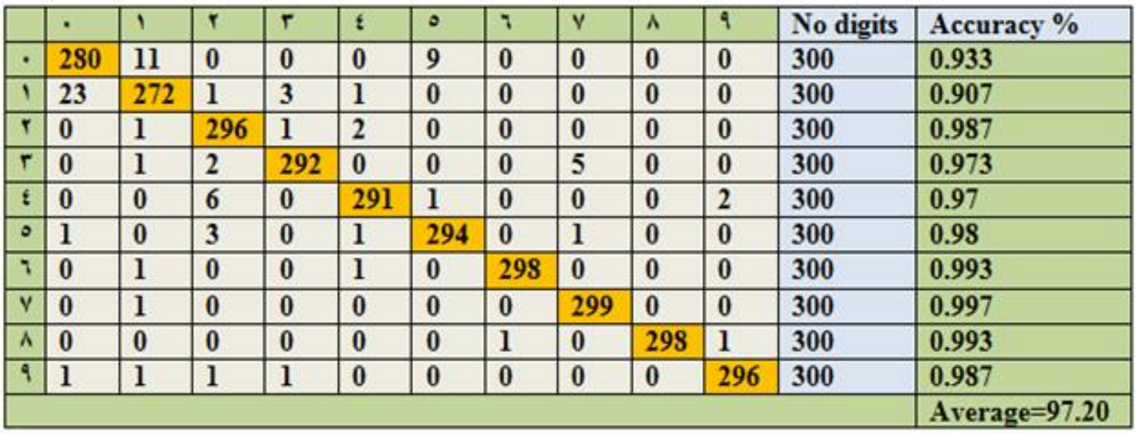
The proposed system for Arabic digit recognition were implemented using Visual basic 2008 Programming Language on computer with Intel ® core™ i5 CPU and ram 4 GB. ADBase database is used which consists of 70,000 Arabic digits written by 700 writers with different ages. It is split into two sets the training set and testing set. The training set has 6000 images (6000 digits per class) and the test set including 10000 digits (1000 digits per class). To compute the proposed system accuracy the following experiment is made:

**Experiment1:** Select 10000 digit image from this ADBase database (7000 images in the training set and 3000 images for testing set) called it **GROUP 1**.

**Experiment 2:** use the all Data set (60000 for Training and 10000 for testing) called it GROUP 2. In other words, used (70% of data for training and 30% for testing).

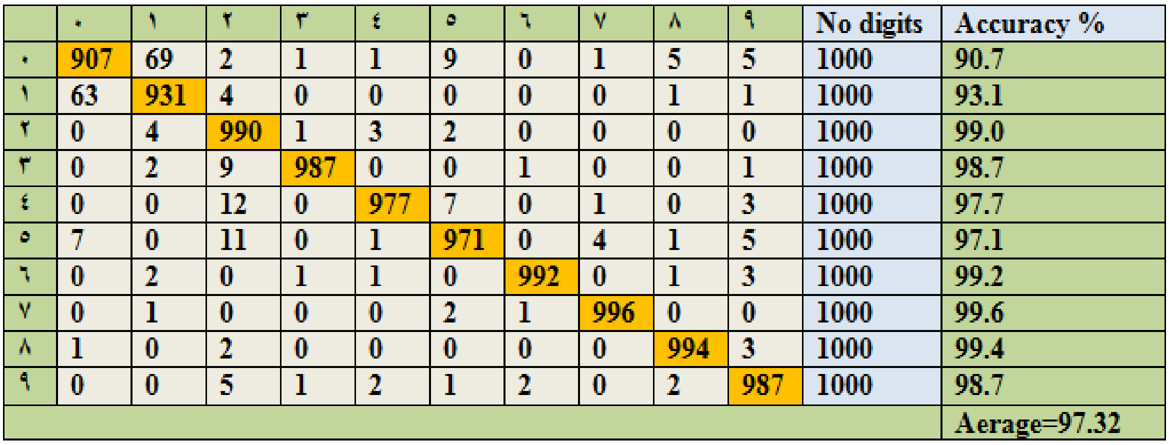
In table 1 the confusion matrix for GROUP 1 data with recognition accuracy for each digit class. And the average accuracy is 97.20%, the maximum error occur with 0 and1. In next experiment when using **GROUP 2** all data set (60000 digits for training phase and 10000 for testing phase). The recognition rate for this data is 97.32%. According to confusion matrix in table 2 the proposed work had the highest recognition accuracy for a number "٧". And also had the lower recognition rate of numbers "٠", " ١ ".

**Table1 : Confusion matrix and recognition rates for Arabic handwritten digits for GROUP 1**



| | ٠ | ١ | ٢ | ٣ | ٤ | ٥ | ٦ | ٧ | ٨ | ٩ | No digits | Accuracy % |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ٠ | 280 | 11 | 0 | 0 | 0 | 9 | 0 | 0 | 0 | 0 | 300 | 0.933 |
| ١ | 23 | 272 | 1 | 3 | 1 | 0 | 0 | 0 | 0 | 0 | 300 | 0.907 |
| ٢ | 0 | 1 | 296 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 300 | 0.987 |
| ٣ | 0 | 1 | 2 | 292 | 0 | 0 | 0 | 5 | 0 | 0 | 300 | 0.973 |
| ٤ | 0 | 0 | 6 | 0 | 291 | 1 | 0 | 0 | 0 | 2 | 300 | 0.97 |
| ٥ | 1 | 0 | 3 | 0 | 1 | 294 | 0 | 1 | 0 | 0 | 300 | 0.98 |
| ٦ | 0 | 1 | 0 | 0 | 1 | 0 | 298 | 0 | 0 | 0 | 300 | 0.993 |
| ٧ | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 299 | 0 | 0 | 300 | 0.997 |
| ٨ | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 298 | 1 | 300 | 0.993 |
| ٩ | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 296 | 300 | 0.987 |
| | | | | | | | | | | | | Average=97.20 |

Table 2 : Confusion matrix and recognition rates for Arabic handwritten digits for GROUP 2



| | ٠ | ١ | ٢ | ٣ | ٤ | ٥ | ٦ | ٧ | ٨ | ٩ | No digits | Accuracy % |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ٠ | 907 | 69 | 2 | 1 | 1 | 9 | 0 | 1 | 5 | 5 | 1000 | 90.7 |
| ١ | 63 | 931 | 4 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1000 | 93.1 |
| ٢ | 0 | 4 | 990 | 1 | 3 | 2 | 0 | 0 | 0 | 0 | 1000 | 99.0 |
| ٣ | 0 | 2 | 9 | 987 | 0 | 0 | 1 | 0 | 0 | 1 | 1000 | 98.7 |
| ٤ | 0 | 0 | 12 | 0 | 977 | 7 | 0 | 1 | 0 | 3 | 1000 | 97.7 |
| ٥ | 7 | 0 | 11 | 0 | 1 | 971 | 0 | 4 | 1 | 5 | 1000 | 97.1 |
| ٦ | 0 | 2 | 0 | 1 | 1 | 0 | 992 | 0 | 1 | 3 | 1000 | 99.2 |
| ٧ | 0 | 1 | 0 | 0 | 0 | 2 | 1 | 996 | 0 | 0 | 1000 | 99.6 |
| ٨ | 1 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 994 | 3 | 1000 | 99.4 |
| ٩ | 0 | 0 | 5 | 1 | 2 | 1 | 2 | 0 | 2 | 987 | 1000 | 98.7 |
| | | | | | | | | | | | | Aerage=97.32 |

## 7. Comparative Study

In Table 3 a comparative study to the proposed method with previous works using the same data set(ADBase database), the comparative study showed the proposed system has a higher accuracy than the other related works.

Table 3 Comparison Results

| Author | Feature extraction method | Classifier | Accuracy |
|---|---|---|---|
| [AlKhateeb and Alseid, 2014] | (DCT) coefficients | (DBN) | 85.26 |
| [Parvez and Mahmoud, 2010] | Directions and length feature | (FATF) | 97.18 |
| [Lawgali 2015] | DCT | ANN | 97.25 |
| Proposed system | V&H_projection, DCT_SD,upper_lower profile | KNN | 97.32 |

## 8. Conclusion

In this paper, presented a system for recognizing the handwritten Arabic digit, which use a combination of three types DCT_SD, V_H projection and upper_lower profile methods for feature extraction and KNN classifier is used for classification. ADBase database is used. From experimental results, it is found that the using a set of more than one type of features is a better method to enhance the recognition rate. After examining the recognition rate for each digit we note that the recognition accuracy is between a high accuracy is 0.996 for digit 7 low accuracy is 0.907 for digit 0. The accuracy of this system is 97.32%. The type and size of databases have an influence on handwritten Arabic digit recognition systems, so may be used another database on this system.

## References

Abdleazeem, S. ; Ezzat El-Sherif , **Arabic Handwritten Digit Recognition**, IJDAR, 11:127–141, Springer 2008 .

Al-Haj, A. "**Combined DWT-DCT Digital Image Watermarking**", Journal of Computer Science, 3 (9), pp. 740-746, 2007.

AlKhateeb , J.H; Marwan Alseid **, "DBN - Based learning for Arabic Handwritten Digit recognition using DCT features"**, 6th International Conference on CSIT, pp. 222-226,IEEE, 2014.

AlKhateeb, J.H. ; Jinchang R., Jianmin J., S. S. Ipson, and H. El-Abed 2008, **Word-based Handwritten Arabic Scripts Recognition Using DCT Features and Neural Network Classifer,"** in 5th International Multi-Conference on Systems, Signals and Devices, 2008, pp. 1-5.

Benouareth, A.; A. Ennaji and M. Sellami 2008,"**Arabic Handwritten Word Recognition Using HMMs with Explicit State Duration"**, EURASIP Journal on Advances in Signal Processing, Article ID 247354, Vo l ,2008.

EL Kessab, B.E. ; C. Daoui, B. Bouikhalene and R. Salouan **, "Isolated Handwritten Arabic Numerals Recognition using The K Nearest Neighbor and The Hidden Markov Model Classifier",** Facta Universitatis (NIˇS) Ser. Math. Inform. Vol. 30, No 5 (2015), 731–740.

El-Sherif E. and S. Abdleazeem , "**A Two-Stage System for Arabic Handwritten Digit Recognition Tested on a New Large Database**", International Conference on Artificial Intelligence and Pattern Recognition, Orlando, Florida, USA, **(2007)**, pp. 237–242.

Hirwani, A. N. Verma, S. Gonnade**," Efficient Handwritten Alphabet Recognition Using LBP based Feature Extraction and Nearest Neighbor Classifier"**, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 11, November 2014.

Lawgali **, A. "Handwritten Digit Recognition based on DWT and DCT"**, International Journal of Database Theory and Application Vol.8, No.5 ,pp.215-222, (2015).

Lawgali , A. "**A Survey on Arabic Character Recognition"**, International Journal of Signal Processing, Image Processing and Pattern RecognitionVol. 8, No. 2 pp. 401-426, (2015).

Lawgali , A. "**Recognition of Handwritten Digits using Histogram of Oriented Gradients"**, International Journal of Advanced Research in Science, Engineering and Technology, Vol. 3, Issue 7 , July 2016

McLaren, M.; N. Scheffer, L. Ferrer, and Y. Lei , "**Effective Use of DCT for Contextualizing Features For Speaker Recognition** " , IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),2014.

Parvez M.; and S. Mahmoud , "**Arabic Handwritten Alphanumeric Character Recognition using Fuzzy Attributed Turning Functions**", First International Workshop on Frontiers in Arabic Handwriting Recognition, pp. 9-14, (2010).

Sahlol, A.; C. Suen 2014, "**Off-line System For The Recognition of Handwritten Arabic Character"**, Computer Science & Information Technology (CS & IT). pp. 227–244, 2014. © CS & IT-CSCP 2014.

Vinciarelli, A.; S. Bengio, and H. Bunke 2004 "**Off-line Recognition of Unconstrained handwritten texts using HMMs and statistical language models**". IEEE Transactions on Pattern Analysis and Machine Intelligence , 26 (6), 709–720, (2004).