Numerals Recognition System using Feature Extraction and Minimum Distance Technique

Rasha Awad[1], Anwar H. Al-Saleh[2], Ali A. Al-Zuky[3], and Farah Adil[4]

[1]Computer Department, College of Basic Education, Mustansiriyah University, Baghdad, Iraq.
[2]Computer Department, College of Science, Mustansiriyah University, Baghdad, Iraq.
[3]Physics Department, College of Science, Mustansiriyah University, Baghdad, Iraq.
[4]Computer Department, Collage of Education, Mustansiriyah University, Baghdad, Iraq.
rashaheart_2005@yahoo.com, anwar.h.m@uomustansiriyah.edu.iq, prof.alialzuky@yahoo.com, farahadil355@gmail.com

Abstract:

Numerals Recognition is that strategy planned to determine the digit class from the text image after segmentation process, by convert the identifying and recognizing characters gating from entering images to ASCII cod or any other proportional machine editable structure. OCR systems are generally appropriate for many tasks such as data entry in business documents (e.g. passport, check and so on.), recognized automatic number plate, multi choice examination etc. Numerals recognition is getting increasingly more consideration since 10 years ago because of its wide extent and scope of uses.

Arabic number recognition system is proposed in this paper. The proposed system is designed using MATLAB 2017 programming language. This paper intends to develop a method to recognize numbers. The image is read by the program; the next process is feature extraction. Six features were extracted from each number. Using those extracted features, with database created for recognition purpose, the number image is classified by minimum distance technique, where a tested image is compared with the features of every number in database. The result has shown that all the numbers were correctly recognized.

Keywords: Optical Character Recognition (OCR), Classification, Filtering and Smoothing

1-Introduction

Optical character recognition (OCR) is process of classification of optical patterns contained in a digital image. The character recognition is achieved through segmentation, feature extraction and classification, where Most of the information available today is either on paper or in the form still photographs, to build digital libraries, this large volume of information needs to be digitized into images and the text converted to ASCII for storage, retrieval, and easy manipulation. However, Current OCR technology is largely restricted to finding text printed Against clean backgrounds, and cannot handle text printed against shaded or textured backgrounds, and/or embedded in images[1].

OCR technology has been used to convert the text in scanned Paper document into ASCII symbols. However, current commercial OCR systems do not work well if text is printed against shaded or hatched background, often found in documents such as photography, map, monetary documents, engineering drawings and commercial advertisements. Furthermore, these documents are usually scanned in gray scale color to preserve details of the graphics and pictures which often exit along with the text. For current OCR system, these scanned image need to be binarized before actual character segmentation and recognition can be done[2].

There are many studies that have been completed by researchers in the field of digital image processing and how to analyze and segment it and extract information from digital images to read texts, numbers and car registration boards. The segmentation of digital images plays a fundamental role in computer vision operations and is the process of splitting the digital image into parts and targets of distinguishing it For some of them, and depending on the characteristics or characteristics. [3] It is a difficult and complex problem, and there are many studies related to the current study. We include some of the most important of these studies [4].

Sika , et al., 2011 [5] suggested a system for detecting symbols that are letters and numbers and the units used are image enhancement , and the use of color conversion (from RGB to YUV) to obtain a mono image by segmenting the image into regions and evaluating those Regions. Then the image is rotated, OCR is recognized by OCR and the comparison of license

plates with the database, the researchers explained that the system is very suitable for use in parking lots.

D. Jiang, et al., 2012 [6]. They proposed a system for distinguishing the car registration plate, as they adopted the car image colors as inputs and registration numbers as outputs. This system includes three main stages to obtain the desired information, which is Locate the board determined , and then segment and distinguish .

L. E. George, et al., 2013 [7]. They proposed a system for distinguishing the car registration plate in Iraq. This system consists of three main stages, the first pre-treatment, which includes a dual image and its fragmentation. The second stage is to locate the registration plate and the last to distinguish the car registration plate.

Mohit, et al. , 2014 [8]. They suggested an automated detection system for the car registration plate based on the matching template, where they used in the study several pictures of cars and treatments were made using the Sobel effect and the Otsu method to reduce the threshold to remove noise and extract and separate codes.

Shilpa, et al. , 2015 [9] .They proposed a system for automatically recognizing the car registration plate and is done in three stages, which is extracting the place of the car registration plate from the image of the original scene. Secondly, finding the board area for work and finally handing the plate to the Box Approach system for the purpose of discrimination in order to identify the car by reading the plate symbols Recorded.

Haidar J. Mohamad et al., 2019 [10] introduced new geometric features (19 features) to recognize characters. The suggested algorithm gives high-acceptable accuracy to recognized characters. To evaluate the error of the classification procedure between tested character and database Minimum Distance Technique (MDT) has been used. The recognition rate of experimental results is reach 99.8% for the proposed method.

ARO Taye Oladele et al., 2019 [11] proposed a recognition system of Arabic Numerals and Alphabet characters using back propagation NN technique adopted on extraction a diagonal feature method for training images characters, and are extracted by moving along their diagonals from the pixels of each zone. The extracted features then used to prepare feed

forward back propagation NN architecture to perform the tasks of recognition and classification. Recognition system performed effectively and gives better recognition accuracy of 92% for Alphabet characters and 91.66% for Arabic numerals.

This paper includes basic ideas for OCR systems and their recognition. By Clarification of the general character recognition form, which is image capture, character recognition software and the last stage of recognition and classification, in order to improve the results of the recognition system

2-Optical Character Recognition (OCR)

Optical character recognition, usually abbreviated to OCR, is computer software that is designed to translate text images into machine-editable text document that can be opened in any word processor or text editor. It helps to quickly digitize paper documents for further manipulations without any manual effort. OCR began as a field of research in pattern recognition, artificial intelligence and machine vision. [12]. OCR allows us to manipulate the printed data, using computer with minimum effort and time. It converts the scanned image to text-based document, which can be easily processed by a word processor or text editor. [13].

2-1Generic Model of an OCR System

One of the main characteristics of Optical Character Recognition is to study automatic reading. Therefore, the aim of OCR is to emulate the human ability to read at a much faster rate by associating symbolic identities with images of character. The language independent text recognition process can be broadly classified into three main categories, which are[13]:

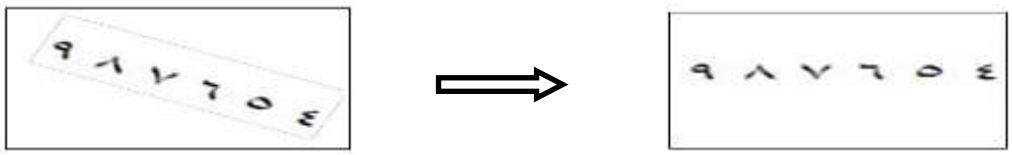
- Image Acquisition.
- OCR Software.
- Output Interface.

3-Numerical Analysis

It is also called pre-processing process in which number is extracted from the document image. This operation is very important for the better results. Different steps involved in the Pre-Processing phase are as follows

3-1 Skew Detection and Removal

Skew is the distortion or tilt, by an angle, in the input image i.e. the angle of the baseline that is not written horizontally. Sometimes it so happens that while scanning a page using a scanner, the page is tilted to a certain angle. As a result, whole text is tilted and the angle of the line makes it inappropriate for OCR processing. So we may say that skew detection and removal plays a significant role in output of OCR project as show in figure (1) [13].



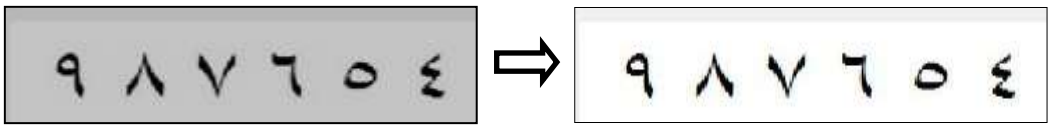
(a) Skewed Image (b) Image after Skew

Removal

Figure (1): Skew Detection and Removal

3-2 Binarization

The binarization is a process, which converts a gray scale image to a black and white image. The simplest way to do this is through thresholding, in which a histogram of the grey values of an image is computed and the cut-off point (valley between the two peaks) is calculated. All the pixels whose value is above the cut-off point are converted to one while all others to zero as show in figure(2) [13].

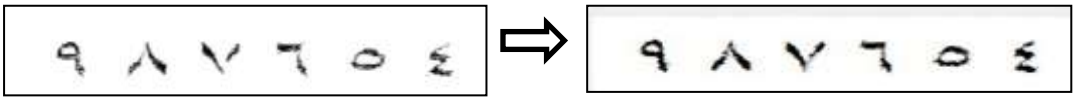


(a) Colored Image (b) Image after Binarization

Figure (2): Binarization

3-3 Filtering and Smoothing

'Filtering and Smoothing' is basically done to remove noise and distortion from the image, which may be produced in image acquisition. The filtering process is used to remove distortion and noise produced at the time of scanning due to shot noise, dark current noise, thermal noise or cross-coupling noise. Fine-textured noise is removed through smoothing as show in figure(3) [13].

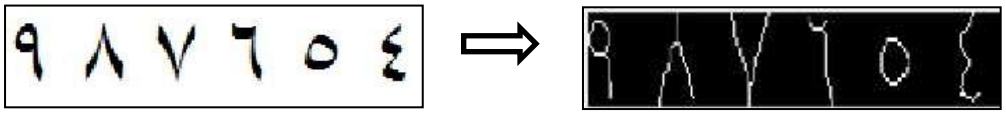


(a) Distorted Image (b) Image after Filtering and Smoothing

Figure (3): Filtering and Smoothing

3-4 Thinning

The thinning is a morphological process, which is an efficient way to express the structural relationships in the character recognition as it removes the selected foreground pixels from the image. It reduces the computational time and effort to traverse an image. But the technique is very much sensitive to noise; a little disturbance can change the shape of an image as show in figure (10).

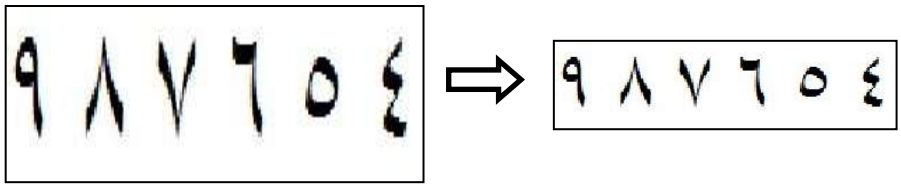


(a) Normal Image (b) Skeleton zed Image

Figure (4): Thinning

3-5 Normalization

Normalization is generally done to overcome size and orientation variation problem. Our purpose is to correctly recognize the similar shapes with different sizes and orientations. The translation, rotation and scaling is performed prior to the start of the actual OCR processing, to normalize the text for size as show in figure(5)[13].



(a) Image before Normalization (b) Image after Normalization
Figure (5): Normalization

4- Segmentation

Image segmentation is the process of dividing an image into multiple parts. This is typically used to identify objects or other relevant information in digital images. Generally, document is processed in hierarchical way. At first level lines are segmented using row histogram. From each row, words are extracted using column histogram and finally characters are extracted from words [1].

5-Feature Extraction

The heart of any character recognition system is the formation of feature vector to be used in the recognition stage. Feature extraction can be considered as finding a set of parameters (features) that define the shape of the underlying character as precisely and uniquely as possible. The term feature selection refers to algorithms that select the best subset of the input feature set. Methods that create new features based on transformations, or combination of original features are called feature extraction algorithms [1].

6- Classification

The classification stage is the decision making part of a recognition system and it uses the features extracted in the previous stage. Classifiers compare the input feature with stored pattern and find out the best matching class for input. In simple terms, it is this part of the OCR which finally recognizes individual characters and outputs them in machine editable form[14].

6-1 Minimum Distance Classification (MDC) Method

Image classification is a segmentation method that aggregates image pixels into a finite number of classes by certain rules so that each class represents a distinct entity with specific Properties. In general, it can be viewed as a label assignment by which image pixels sharing similar properties will be assigned to the same class. The minimum distance technique is a supervised method, which calculates the mean vectors for each class and calculates the Euclidian distance from each unknown pixel to the mean vector for each class. Then all pixels are classified to the nearest class unless a threshold is specified. This technique is highly recommended in all image classification applications. The advantage of this technique is that it not only is a mathematically simple and computationally efficient technique, but also provides better accuracy than others classification methods, in the case when the number of training samples is limited[14].

$$MDC = \min_c \left| \sum_{c=1}^{n_c} I_b(x,y) - \bar{\mu}_b(c) \right| \quad \Rightarrow class \quad = c \qquad (1)$$

where min_c is the minimum distance between pixel and mean of class, b represents the index of color bands (RGB), c represents index of class which have value from (1 To n_c) ,where n_c represents the number of image classes, I_b (x,y) is image pixel values, $\bar{\mu}$ is mean value of color band (b) in class (c)[14].

7- The Proposed System and Results

The first stage of proposed system is image acquisition. This stage was done using two types of fonts, Arial and Calibri, for numbers from 0 to 9. Each font is with size: 16, 22, 28, 36, 42, 50, and 72.

8- Convert to Binary

In order to segment the numbers, of the input image, it should convert to the binary level. For binarization the image a threshold value is required. In this phase, the MATLAB function im2bw is used, to select the optimal thresholding depending on the intensity level of each input image. So, if the value of the pixels in the license plate image is less than the threshold value,

it is expressed as "1"; and if it is greater than the threshold value it is expressed as "0". In this way, the image is converted into the binary level. Figure (6) shows the input image and the image after converted into the binary level.

Figure (6)**:** Input image before and after converted to binary

9-Segmentation

In this stage each digit in the binary image was extracted using rectangle function, as shown in figure (7).

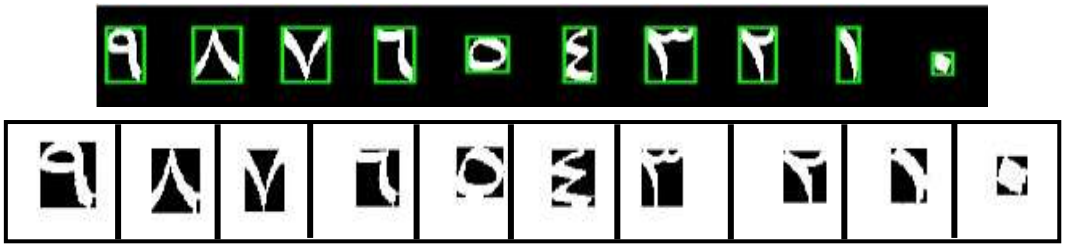


Figure (7): segmentation stage

10-Thinning Process and Normalization

This process is considered important process in a preprocessing part. Thinning of a binary pattern is generally considered as a process of iterative deletions of pixels along the edge of the pattern until the pattern is thinned to line drawing, i.e. skeleton. Before done template matching technique each segmented character, number and word must be resizing to a fixed size. The size of every number may be different because the images are taken from different sizes. So, in this system the numbers images are normalized into 100x60 images as a standard size of number. Figure (8) shows the resulting final skeleton after applying the thinning and Normalization techniques.

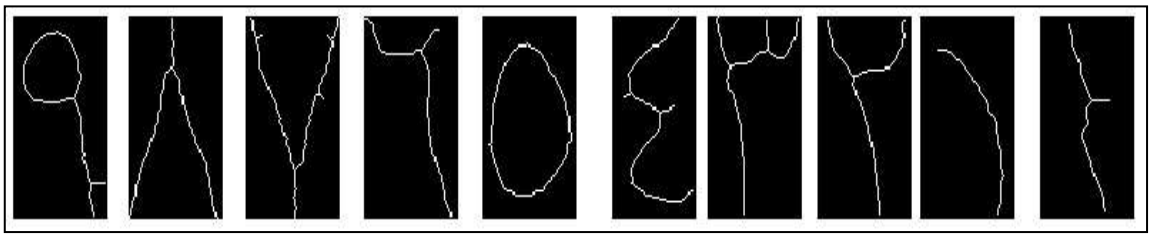


Figure (8): applying thinning technique for each segment number.

11-Feature Extraction

The purpose of feature detection is to obtain those features, which preserve the useful information about the image to the largest extent. The aim of feature selection is to determine those principal feature components depending on a certain classification task in order to achieve an effective classification. The above idea shows that the output of feature detector reflects the information of the image. The feature after selection must contain the information that is useful to distinguish different classes for image classification. From the above viewpoint, we propose vector of six features for each segment number, Vf = [Area, Areath, P1, P2, P3, P4, P5, P6], where Area and Areath, are the area of number image before and after thinning, respectively. P1, P2, P3, P4, P5, P6 are the sum of pixels of an image, where the image is divided into six parts, as shown in figure (9), and computes the sum of pixels for each part.

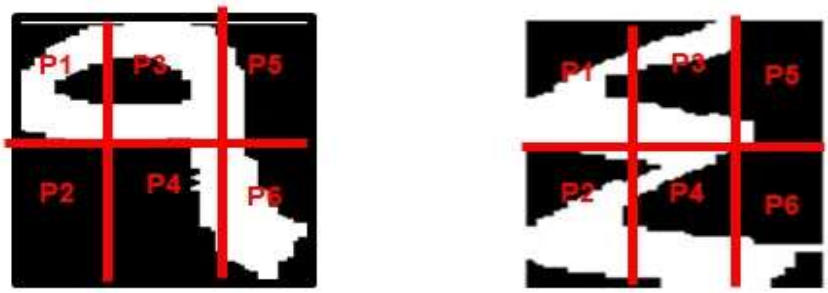


Figure (9): determined features P1, P2, , P6 for two numbers images 9 and 4.

12-Database Building

The database in this paper was built using two fonts, Arial and Calibri, for numbers from 0 to 9. Each font with size, 16, 22, 28, 36, 42, 50, 72, then has been obtained 140 number's image. For each number's image, propose vector of six features, Vf = [Area, Areath, P1, P2, P3, P4, P5, P6], where Area and Areath represent the area of input image (the number of pixels that equal 1) before and after thinning operator, respectively. Pi is the area of part i, where i = 1, 2, …, 6, illustrated in figure(9), so has been obtain 140 vectors (Vf), each vector with 8 Column for each input image number to database, then the size of obtained database is (140x8).

The following table(1) illustrate some results of the program's execution for different inputs images for two numbers images 9 and 4, the proposed system can recognize all numbers with high accuracy and in short time.

Table (1): determined features P1, P2… P6 for two numbers images 9 and 4.(a) before normalizing, and (b) after normalizing

(a)

Font & Size	A	Ath	P1	P2	P3	P4	P5	P6
4 arial-16	2595	172	374	339	623	828	254	339
4 arial-22	2511	151	481	423	564	709	198	306
4 arial-28	2493	164	476	443	560	671	190	316
4 arial-36	2415	165	429	413	527	674	213	312
4 arial-42	2427	161	455	431	532	663	197	303
4 arial-50	2432	163	413	405	550	656	225	333
4 arial-72	2435	166	443	418	518	649	227	329
9 arial-16	2385	161	644	0	762	60	375	660
9 arial-22	1648	164	365	0	515	0	344	486
9 arial-28	1837	158	446	0	555	0	299	600
9 arial-36	1829	168	450	0	607	0	290	557
9 arial-42	1824	160	472	0	628	9	256	541
9 arial-50	1736	172	460	0	603	7	255	492
9 arial-72	1965	165	545	17	729	17	244	516

(b)

Font & Size	A	Ath	P1	P2	P3	P4	P5	P6
4 arial-16	0.36	0.78	0.4	0.4	0.57	0.82	0.28	0.34
4 arial-22	0.33	0.62	0.6	0.5	0.51	0.7	0.22	0.3
4 arial-28	0.32	0.72	0.6	0.6	0.5	0.66	0.21	0.31
4 arial-36	0.29	0.73	0.5	0.5	0.47	0.67	0.24	0.31
4 arial-42	0.3	0.7	0.5	0.5	0.47	0.66	0.22	0.3
4 arial-50	0.3	0.71	0.5	0.5	0.49	0.65	0.25	0.33
4 arial-72	0.3	0.74	0.5	0.5	0.46	0.64	0.25	0.33
9 arial-16	0.28	0.7	0.8	0	0.71	0.06	0.41	0.66
9 arial-22	0.04	0.72	0.4	0	0.45	0	0.38	0.48
9 arial-28	0.1	0.68	0.5	0	0.5	0	0.33	0.6
9 arial-36	0.1	0.75	0.5	0	0.55	0	0.32	0.55
9 arial-42	0.1	0.69	0.5	0	0.57	0.01	0.28	0.54
9 arial-50	0.07	0.78	0.5	0	0.55	0.01	0.28	0.49
9 arial-72	0.14	0.73	0.6	0	0.68	0.02	0.27	0.51

13-Numerical Recognition:

The final stage of proposed system is recognizing of each segmented number image. The basic process takes place in numerical recognition is to convert the numbers image to a text file that can be edited and used as such by any other program or application that needs it. There are many techniques that used for this purpose. In this paper minimum distance technique (MDT) was used for recognition process which gives best recognition. After the number's image is generate a vector of coefficients that represent the important information (or main details for the character's image), Vf, this vector is assumed to uniquely represent input image since it carries the important details of that image, Then Euclidean distance between this vector and each vector in database will be measured as show in table (2),,that illustrate some results of the minimum distance technique (MDT) execution for two numbers images 9 and 4. Finally, the minimum distance points to the corresponding character, and then the character is recognized.

Table (2): Results of the MDT for two numbers images 9 and 4

	min-dis	out	min-dis	out
1	0	4	0.231	9
2	0.122	4	0.252	9
3	0.126	4	0.286	9
4	0.1868	4	0.294	9
5	0.233	4	0.299	9
6	0.345	4	0.364	9
7	0.536	4	0.385	9
8	1.17	4	0.5	9
9	1.18	4	0.565	9
10	1.231	4	0.673	9

14-Conclusion

This paper has been proposed a new structure of off line OCR system which is not based on ANN, to avoid the time consuming problems. Moreover, it benefited from the image property which produces a unique vector which helps to identify each number. By using this unique vector, the proposed system has recognized the input number's image using minimum distance technique (MDT) between the input vector and the vectors in the database, then the shortest distance pointed to the corresponding number. The

result was considerably high in terms of accuracy and recognition rate. The result has shown that all the numbers were correctly recognized.

References

[1] Hiral Modi, M.C.P., A Review on Optical Character Recognition Techniques. International Journal of Computer Applications, 2017. 160(6): p. 20-24.

[2] S. Kranthi, K. Pranathi, and A. Srisaila, "Automatic Number Plate Recognition", International Journal of Advancements in Technology, Vol. 2, No. 3, PP. 408-422, July, 2011.

[3] Ismael Saad Eltoum, Zhaojun Xue" Automatic Gate Control System Based On Vehicle License Plate Recognition" International Journal of Engineering Research & Technology, Vol. 3 - Issue 8 (August - 2014)

[4] M. I. Khalil، "Car Plate Recognition Using the Template Matchin Method"، International Journal of Computer Theory and Engineering، Vol. 2، No. 5، PP. 683-687، October، 2010.

[5] .Cika, P., Zukal, M. and Sebela, M., "Vehicle License Plate Detection and Recognition Using Symbol Analysis", 34th International Conference on Telecommunications and Signal Processing (TSP), IEEE, PP. 589-592, 2011.

[6] D. Jiang, M. Tulu, E. Tiruneth, and G. Ashenafi, "Car Plate Recognition System", In Fifth international conference on Intelligent Networks and Intelligent Systems, IEEE, PP. 9 -12, 2012.

[7] L. E. George , Nada Najeel Kamal," Iraqi License Plate Recognition System" College of Science, Baghdad University, 2013.

[8] Mohit Kumar Pandey,et," Automatic Vehicle Regestration Plate Recognition System Using Soft Computing Techniques", Department of Electronics and Telecommunication Engineering, S.G.S.I.T.S., Indore, Madhya Pradesh, India, al International Journal of Computer and Electronics Research [Volume 3, Issue 5, October 2014].

[9] Shilpa Pawar, et," Automatic Number Plate Recognition", Savitribai Phule Pune University, Pune – Maharashtra – 411015, Department Of Electronics & Telecommunication, International Conference on Inter Disciplinary Research in Engineering and Technology [ICIDRET] 222.

[10] Haidar J. Mohamad, Seham A. Hashim, Anwar H. Al-Saleh " Recognize printed Arabic letter using new geometrical features " Indonesian Journal

of Electrical Engineering and Computer Science Vol. 14, No. 3, pp. 1518~1524, June 2019.

[11] ARO Taye Oladele, MUSA Abdullahi Yola, "Recognition of Alphabet Characters and Arabic Numerals Using Back Propagation Neural Network " Sensors Journal of Computer Science and Control Systems, Volume 11, No. 2, 2019.

[12] Sobia Tariq Javed, Ameera Maqbool, Sehrish Jameel and Samia Asloob Qureshi, "Urdu Nastaleeq OCR", BS final year report, 2005

[13] R. El-Hajj, L. Likforman-Sulem and C. Mokbel, "Arabic Handwriting Recognition Using Baseline Dependant Features and Hidden Markov Modeling", in the 8th International Conference on Document Analysis and Recognition, ICDAR 2005, Seoul, Korea, (2005).

[14] Chitra D.,"Image Classification Tool for Land Use / Land Cover Analysis: A comparative Study of Maximum Likelihood and Minimum Distance Method ", International Journal of Geology, Earth and Environmental Sciences, Vol. 2, No.3, 2012.