ISSN No. 0976-5697

Volume 2, No. 1, Jan-Feb 2011

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Handwritten Arabic Text Recognition System using Window Based Moment Invariant Method

Mowaffak Othman Al_Barraq*
College of Computer and Information Technology,
Sana'a University,
Sana'a Yemen
Contact No.(00967-773922007)
mowaffak_albaraq@rediffmail.com

S.C.Mehrotra
Department of Computer Science and Information Technology.
Dr. Babasaheb Ambedkar Marathwada University
Aurangabad – 431001, Maharashtra State, India
mehrotrasc@rediffmail.com

Abstract: A sliding window approach has been used for segmentation of the handwritten Arabic texts into atomic characters or sub characters using features based on moment invariant technique. Each separated characters is represented by the n-dimensional space where n is number of windows related to the character. The Recognition rate based on Euclidian distance and cosine (θ) similarity approaches, is found to be 92.86 %.

Keywords: Windowing; Handwritten Arabic Texts; Moment Invariants; Features Extraction; Curve Fitting and Representation.

I. INTRODUCTION

Arabic is a language spoken by Arabs in over 22 countries, and roughly associated with the geographic region of the Middle East and North Africa as first language (Mother Tongue). It is also spoken as a second language by several Asian countries, (e.g. Iran, Indonesia, India, Malaysia, Pakistan, etc), in which Islam is the principle religion. Non-Semitic Languages such as Farsi, Urdu, Malay, and some West African languages such as Hausa have also adopted the Arabic alphabet for writing. Due to the cursive nature of the script, there are several characteristics that make recognition of Arabic distinct from the recognition of Latin scripts or Chinese.

An effective design of any Arabic recognition system needs to consider following characteristics [1,2,3,4]:

1. Arabic text content words (printed or handwritten) is cursive in general and written from right to left. Out of 22 Arabic letters are normally connected to each other on the baseline, other 6 letters are not connected with succeeding letters and recognition must consider this aspects.

2. Arabic Script is constituted of 28 characters, in addition to 10 numerals used in Hindi, punctuation marks, as well as spaces and special symbols. Each character can appear in four different manners depending on the position of the word [1,3,4], [Beginning Form (BF), Middle Form(MF), Isolated Form (IF),and End Form(EF)],as evident in table 1. The total number of forms near is about 107. Considering some specific characters in diacritical symbols, this goes up to 140.

3. The Arabic characters of a word are connected along a virtual baseline. A baseline may be identifying having the highest density of black pixels. The nonexistence of the baseline needs different segmentation approach from those used in other unconnected scripts.

4. Many Arabic characters have dots, which are positioned above or below the letter body.

Dots can be single, double or triple. Different Arabic letters can have the same body and differ in the number of dots identifying them.

Table I. Different forms of Arabic Alphabets

End	*Middle*	*Beg.*	*Isolate*	*Meaning*
ـا	-	ا	ا	**Alif**
ـب	ـبـ	بـ	ب	**Baa**
ـت	ـتـ	تـ	ت	**Taa**
ـث	ـثـ	ثـ	ث	**Thaa**
ـج	ـجـ	جـ	ج	**Jiim**
ـح	ـحـ	حـ	ح	**haa**
ـخ	ـخـ	خـ	خ	**khaa**
ـد	-	د	د	**Daal**
ـذ	-	ذ	ذ	**Dhaal**
ـر	-	ر	ر	**Raa**
ـز	-	ز	ز	**Zayn**
ـس	ـسـ	سـ	س	**Siin**
ـش	ـشـ	شـ	ش	**Shiin**
ـص	ـصـ	صـ	ص	**Saad**
ـض	ـضـ	ضـ	ض	**Daad**
ـط	ـطـ	طـ	ط	**Taa**
ـظ	ـظـ	ظـ	ظ	**Dhaa**
ـع	ـعـ	عـ	ع	**Ayn**
ـغ	ـغـ	غـ	غ	**Ghayn**
ـف	ـفـ	فـ	ف	**Faa**
ـق	ـقـ	قـ	ق	**Qaaf**
ـك	ـكـ	كـ	ك	**Kaaf**
ـل	ـلـ	لـ	ل	**Laam**
ـم	ـمـ	مـ	م	**Miim**
ـن	ـنـ	نـ	ن	**Nuun**
ـه	ـهـ	هـ	ه	**Haa**
ـو	-	و	و	**Waaw**
ـي	ـيـ	يـ	ي	**Yaa**

5. Arabic uses a short vowel referred to as Diacritic. A Diacritic is placed above or below the body of the Arabic character (ex: ًٌ، ، ِ, ٍ, ّ, َ, ُّ، ِّ، ْ).

© 2010, IJARCS All Rights Reserved

6. Some of the Arabic Character are located under baseline overlap characters (ex. ص , ش, س، ل, ز , ر ف،ق،ى،و،ض..). The overlapping makes the recognition more complicated, as it is difficult to determine spacing between characters and words

7. Some characters have a very small punctuation size with respect to the rest of the text. This makes them very difficult to recognize.

8. Arabic characters have different heights; therefore noise detection becomes very hard.

9. Many of the Arabic Words may have one or more sub-words . This is due to the fact that some characters are not connectable from the left side with the succeeding character difficult like " (الطالبات) it is one words contents four spaces two isolate letters (ا,ت), and two sub words (لبا ,لطا).

10. Some of the Arabic character consist of two parts or two characters (ex. ظ، ك ،ط ,لأ,ؤ,إْ , إ).

11. The character can be stretched over the baseline to more than one character space like given in this example (مليـــــــــــــار).

These and other special characteristics make it impossible to adapt English or other text recognition system to Arabic text recognition. The present paper described a window based approach to tackle above mentioned problems. The section 2 of the paper gives a literature Review. Section 3 describes details of window based methodology used in the work. Results and Discussions are explained in section 4. The paper is concluded in Section 5.

II. Literature Review

The main problem in AOCR Online or Offline is due to cursive property of Arabic script [8, 9,10].A structural classification method for recognizing on-line handwritten isolated Arabic characters first time was proposed by A. Amin et al [8]. In this work features such as the shape of the main stroke, the number of strokes, the number and position of the secondary parts have been extracted from the character. If these features do not permit the recognition of a character by a simple dichotomy consultation of the dictionary, a backtracking is performed in order to correct the shape of the main stroke. Secondary features of the main stroke such as the frame size, the start point and the curvature are also extracted using a distance function in order to remove the ambiguity and provide an exact match.

In 1981, Parhami and Taraghi [9] have presented a technique for the automatic recognition of printed Farsi text. The technique is applicable, with little or no modification, to printed Arabic text (Farsi has an alphabet similar to Arabic).The most important parts of the system are[12] isolation of symbols within each sub word and recognition. The main step in segmenting symbols is to determine the pen (script) thickness which is used to find connection columns. The application of the technique to Farsi newspaper headlines has been found to be 100% successful, as reported by the authors. However, fonts of smaller point-size will result in less than perfect recognition. The system is heavily font dependent and the segmentation process is expected to give degraded results in many cases.

In 1987 Almuallim and Yamaguchi [10] have proposed stroke based technique since it is difficult to separate a cursive word directly into characters. These strokes are then classified using their geometrical and topological properties. The relative positions of the classified strokes are examined, and the strokes are combined in several steps into a string of characters, which represent the recognized word. A maximum recognition rate of 91% was achieved. The system failure, in most of the cases was due to wrong segmentation of words.

In 1988 Ramsis et al. [11] adopted a method of segmenting Arabic typewritten characters after recognition.. They have assumed that the rightmost columns of a word, the number of which equals the width of the smallest character, constitute a character. Moments are calculated and checked against the feature space of the font. If a character is not found, another column is appended to the underlying portion of the word and moments are calculated and checked again. This process is repeated until a character is recognized or the end of the word is reached. The method allowed the system to handle overlapping and to isolate the connecting baseline between connected characters. This method seems to be sensitive to font type and input pattern variations. The system has also used intensive computations to compute the required accumulative moments. No figures are reported regarding the system recognition rate and efficiency.

In 1989, Amin and Mari [12] have presented a structural probabilistic approach to recognize Arabic printed text. The system is based on character recognition and word recognition. Character recognition includes segmentation of words into characters using vertical projections and identification of characters. Word recognition is based on Viterbi algorithm and can handle some identification errors. The system was tested on just a few words and no figures were reported about its performance. The method has inherent ambiguity and deficiencies due to interconnectivity of Arabic text.

In 1990, Al-Emami and Usher [13] have presented an on-line system to recognize handwritten Arabic words. Words are segmented into primitives that are usually smaller than characters. The system is taught by being fed by specifications of the primitives of each character. In the recognition process, parameters of each primitive are found and special rules are applied to select the combination of primitives that best matches the features of learned characters. The method required manual adjustment of some parameters. The system was tested against only 170 words, written by 11 different subjects for 540 characters.

In 1992, Zahour et al. [14] presented a method for automatic recognition of off-line Arabic cursive handwritten words based on a syntactic description of words. The features of a word were extracted and ordered to form a tree description of the script with two primitive classes: branches and loops. In this description, the loops were characterized by their classes and the branches by their marked curvature, their

© 2010, IJARCS All Rights Reserved

relationship, and whether they were in clockwise or counterclockwise direction. Some geometrical attributes were applied to the primitives that are combined to form larger basic forms. A character was then described by a sequence of the basic forms. The reported recognition rate of the system was found to be 86%.

In 1996, Abuhaiba [15] presented a text recognition system, capable of recognizing off-line handwritten Arabic cursive text. A straight-line approximation of an off-line stroke was converted to a one-dimensional representation. Tokens were extracted from this one-dimensional representation. The tokens of a stroke were re-combined to meaningful strings of tokens. Algorithms to recognize and learn token strings were presented. The process of extracting the best set of basic shapes that represented the best set of token strings that constituted an unknown stroke was described. The method was developed to extract lines from pages of handwritten text, arrange main strokes of extracted lines in the same order as they were written, and present secondary strokes to main strokes. Presented secondary strokes were combined with basic shapes to obtain the final characters by formulating and solving assignment problems for this purpose.

In 2000, A. Amin [16] produced recognition of printed Arabic text based on global features and decision tree learning techniques. He used a new technique for the recognition of Arabic text using the C4.5. Along with machine learning system which avoid the differences between two Arabic fonts. He aims to eliminate segmentation phase. therefore global features of the input Arabic word is used to extract features such as number of subwords, number of peaks within the subword, number and position of the complementary character , Finally, machine learning C4.5 is used to generate a decision tree for classifying each word . The system tested with 1000 Arabic words with different fonts (each word has 15 samples) and the correct average recognition rate obtained using cross-validation was 92%.

In 2000, M. Fakir, M .M. Hassani and C. Sodeyama [4] presented a method for recognition Arabic characters based on Hough Transform Technique . He analyst the texts by horizontal projection profile (HPP) so the less pixels is in between the lines , by scanned vertically the lines to determine the less or black pixels to segment the lines into words . and the vertical projection and a fixed threshold is used for segmenting a word into characters. Then features are extracted using Hough transform. Next, characters are classified using Dynamic programming matching technique and the extracted features. Characters classification by using Dynamic Programming matching technique, and features extracted in the Hough transform space. In the second one, simple topological features extracted from the geometry of the secondary parts are used by the topological classifier to completely recognize the characters. The topological features used to classify each type of the secondary part are the width, the height, and the number of the secondary part He obtained the recognition rate of about 95%.

In 2002, Taufiq Fahmi [17] presented a new character segmentation algorithm (ACSA) of Arabic scripts. The developed segmentation algorithm yielded on the segmentation of isolated handwritten words in perfectly separated characters. It was based on morphological rules, which were constructed at the feature extraction phase.

In 2002, A. Nosary, T. Paquer, L. H., and A.Bensefaia [18] presented a technique based on adaptation for handwritten text recognition writer This adaptation is realized by iterating word recognition steps that allow to label the writer representations (allographes) on the whole text, and re-estimation of character models. His approach relies in the ability of the system to learn in an unsupervised manner the writer specificities. The most particular point in this unsupervised learning principle lies in the control of lexical decisions, which directly influence the knowledge inferred at the level of writer allograph. This last point is beyond the scope of our future research. No result has been calculated in this system only the author expected adaptation approach it will improve the system performance.

In 2003, M.S. Korsheed [19], He presented a new method on off-line recognition of handwritten Arabic script. The method trains a single hidden Markov model (HMM) with the structural features extracted from the manuscript words. The HMM is composed of multiple character models where each model represents one letter from the alphabet. The performance of the proposed method is assessed using samples extracted from a historical handwritten manuscript. A new method for recognizing cursive handwritten Arabic words has been presented.. The HMM is composed of multiple character models where each model represents a letter from the alphabet. Using a simple Arabic spell-check, the system performance was enhanced by a rate of 9–10%. These recognition rates are expected to decrease when including more handwritten fonts.

In 2003, Ibrahim Abuhaiba [20] proposed the method for Discrete Arabic Script for better Automatic Document Understanding. His work laid in the groundwork for the design of new fonts to produce discrete Arabic script, for the first time, instead of cursive Arabic script. These fonts helped in automatic document understanding and could be used to print books, newspapers, periodicals, and all other printed materials. A strategy to break the cursive law of Arabic script was presented by reviewing what has happened to Arabic calligraphy since its start. It is important to note that all other properties of Arabic writing system were preserved when producing such fonts.

In 2003, M. Sarfraz and S. Nazim, [3] have used technique for the Automatic Recognition of Arabic Printed text using Artificial Neural Networks. The main features of the system are preprocessing of the text, segmentation of the text to individual characters by using histogram and vertical and horizontal projection , Feature extraction using moment invariant technique and recognition using RBF Network for one size and the Naskh font.

In 2006, M. Syiam, and T. M. Nazmy, [21] produced Histogram Clustering Method for the Segmentation of the Arabic Word. This method gives the ability to process different user styles, and manages the variability of pen strokes. Also, a new algorithm for separating overlapped characters was proposed to support the proposed technique for segmentation. The feature extraction process was based on a combination between the PCA network and characters geometric features. A classifier for hundred of Arabic

© 2010, IJARCS All Rights Reserved

character images was designed using a decision tree induction algorithm, and MLP network. A segmentation correctness of 96% was achieved. and success rate for whole system was achieved 91.5 %.

In 2006, A. Amin and N. Al-darwish,[22] Introduced Machine Learning System .The system structural approach for feature extraction was used. The use of machine learning has removed the tedious task of manually forming rule-based dictionaries for classification of unseen characters and replaced it with an automated process which can cope with the high degree of variability that exists in printed and handwritten characters.

In 2007, Bhagile, and M. Albaraq [23] presented Arabic Printed Numeral Recognition System using a block based approach. The technique is applied on Arabic numeral with 0 to 9. The features have been extracted by sorting first ten most likely density points. The technique is found to work well for recognition of the English numerals, whereas it does not give satisfactory results for the Arabic numerals. The technique presented here is based on windowing approach and moment invariants features.

III. Methodology

A. *Creation of handwritten database*

As no database was available for Arabic Handwritten texts, the first step was to prepare an appropriate data base. The data has been collected by forty subjects who were aware of the Arabic script. The written pages from these forty subjects have been digitized by using a standard scanner with resolution of 300X300 Pixels. In the second stage the images are converted into black and white format. This forms the database used in the study.

B. *Preprocessing*

The preprocessing is performed for following tasks:

-Noise analysis and removal.

- Skew detection and correction and

- Evaluating the gap between the words and characters.

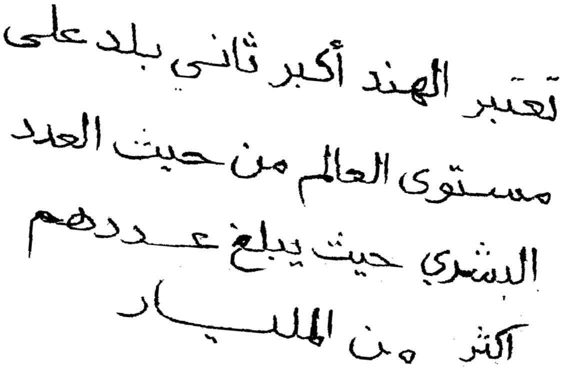

تعتبر الهند أكبر ثاني بلد على
مستوى العالم من حيث العدد
البشري حيث يبلغ عـــددهم
أكثر مـن الملـيـــار

Figure 1. Original texts with skewness

The filtering is done to reduce noise from the image. The isolated pixels are recognized and are removed from the images. Skew nature in line texts is detected through slope. The information is used to correct the skewed nature as shown in Figure 1 with horizontal line as shown in Figure 2.

An appropriate window size is selected and number of pixels is computed in each window. The number of pixels gets minimum at the end of alphabet and beginning of another alphabet. This feature was used to detect the beginning of each alphabet.

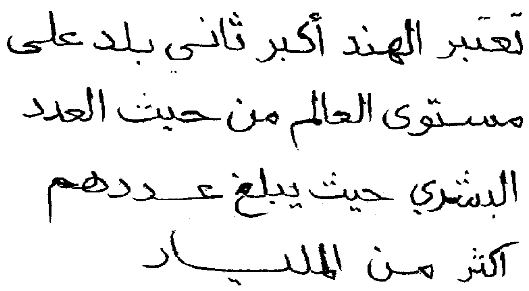

تعتبر الهند أكبر ثاني بلد على
مستوى العالم من حيث العدد
البشري حيث يبلغ عـــددهم
أكثر مـن الملـيـــار

Figure 2. Texts after removing skewness

C. *Text segmentation*

Segmentation of Arabic texts is a complicated task. An appropriate size of window is selected in the right part of the line. The window is slided from the right to the left. The windows sliding is done such that the next position overlaps half of the earlier position as can be seen in figure 3 and figure 4. The number of windows varies with the length of the word under consideration. For example, For example, the word 1 requires ten windows, whereas the word 5 needs only seven.

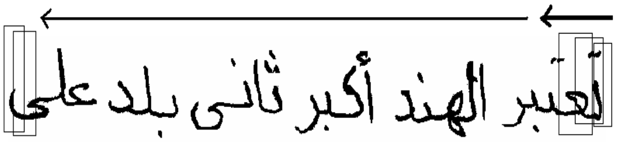


Figure 3. Windows moved from right to left

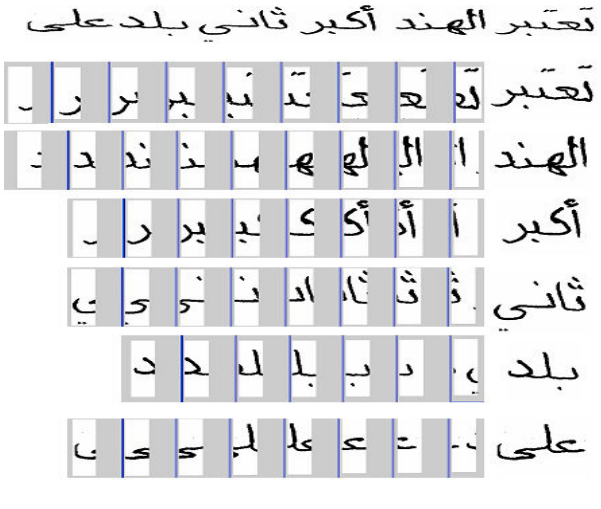


Figure 4. Characters with windows

© 2010, IJARCS All Rights Reserved

D. *Feature Extraction*

The features in each window are extracted through seven invariant moments as given by Hu[5]. Important prosperity of these moments is that they are invariant under reflection, rotation and scaling [5, 6]. In this work features have been extracted at each window. The Matlab Language has been used for the computation of the seven moments. An example of these values is given in table II for the characters shown in figure 4

Table II. The values of ϕ_1 to ϕ_7 as a function of windows

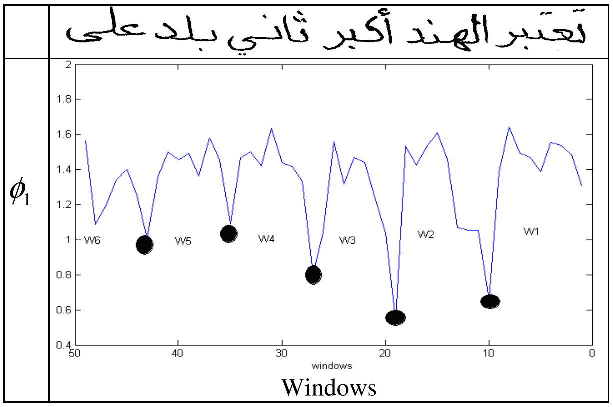
n / ϕ_i	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	ϕ_6	ϕ_7
1	1.31	3.50	7.22	7.33	14.63	9.08	16.28
2	1.49	4.39	8.75	7.89	16.43	10.23	17.58
3	1.54	4.70	9.36	8.83	18.09	11.38	19.18
4	1.56	4.76	9.72	9.69	19.46	12.12	21.27
5	1.39	3.66	9.04	9.15	18.25	10.98	20.81
6	1.47	4.90	10.83	10.12	21.00	12.91	22.75
7	1.49	5.59	11.42	12.43	28.08	16.82	26.56
8	1.64	8.08	11.93	9.00	20.37	15.13	22.17
9	1.08	4.05	6.74	7.06	14.58	11.90	14.66
10	0.66	1.52	7.11	6.32	13.88	7.91	17.79
11	1.05	2.80	6.96	6.55	13.58	8.21	16.31
12	1.05	2.88	7.76	6.87	14.29	8.38	17.94
13	1.07	4.10	7.64	8.31	17.94	10.98	16.37
14	1.46	6.06	11.09	9.08	20.15	14.43	21.18
15	1.61	5.35	11.44	9.80	21.08	12.48	20.82
16	1.53	5.32	8.81	8.59	18.24	12.17	18.69
17	1.42	7.77	9.02	8.19	18.56	12.95	18.45
18	1.53	4.40	12.15	10.02	21.45	12.22	21.46
19	0.20	0.33	1.42	1.35	2.77	1.21	10.03
20	1.04	2.73	8.29	9.24	19.03	11.44	19.31
21	1.24	3.65	9.18	10.00	23.64	15.03	24.63
22	1.44	4.67	10.87	9.63	20.94	13.39	22.20
23	1.47	4.66	11.12	9.52	20.59	12.53	20.80
24	1.32	4.21	9.31	9.54	21.80	14.00	22.10
25	1.56	7.56	10.07	8.05	20.61	14.63	18.74
26	1.04	3.54	6.95	7.13	14.52	9.44	16.36
27	0.81	1.93	6.08	5.90	11.92	6.88	15.29
28	1.33	3.66	9.06	11.40	22.41	16.13	22.06
29	1.42	4.23	9.43	9.09	21.07	15.02	21.02
30	1.44	4.97	9.24	8.30	18.31	11.64	17.72
31	1.64	5.45	9.66	9.97	21.00	13.75	20.52
32	1.42	3.78	8.81	8.55	17.26	10.45	19.14
33	1.50	5.12	9.99	8.80	18.62	11.42	18.75
34	1.47	4.51	8.32	9.51	18.62	11.86	19.10
35	1.09	2.72	7.28	7.10	14.32	8.49	16.29
36	1.46	7.06	10.26	7.54	23.50	13.99	17.58
37	1.58	5.76	9.90	11.49	22.87	14.68	22.44
38	1.36	3.74	12.71	10.96	24.77	14.42	26.41
39	1.49	5.21	8.90	9.11	20.74	12.29	18.40
40	1.46	8.05	10.93	8.65	19.06	12.85	19.21
41	1.50	9.08	9.42	8.90	18.85	15.71	21.96
42	1.08	4.23	7.34	6.86	14.86	9.42	15.27
43	1.51	7.30	9.31	10.20	19.94	13.84	19.75
44	1.25	5.19	8.71	6.95	18.04	10.24	15.80
45	1.40	4.38	8.96	7.29	15.87	9.88	19.56
46	1.34	3.84	9.04	9.45	21.62	14.67	20.17
47	1.20	4.49	9.09	6.59	15.45	9.41	16.17
48	1.09	4.73	8.71	6.25	14.59	9.17	15.94
49	1.55	4.90	7.57	7.05	15.95	11.02	16.88

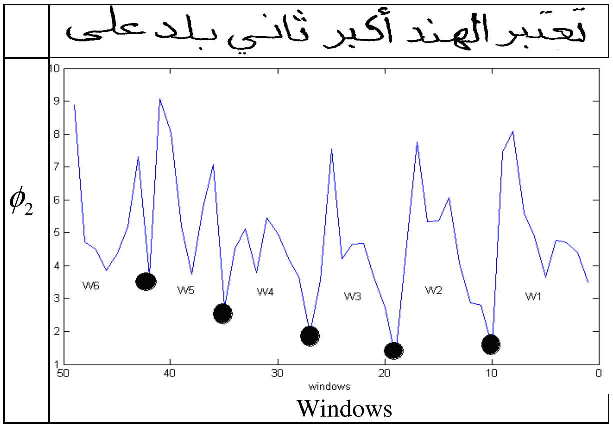
E. *Feature Representation*

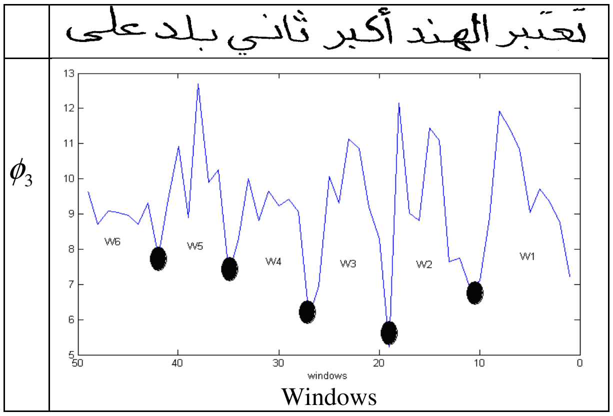
A window with an appropriate size is allowed to slide along be extracted through the window.

In our case the window size is selected in such a way that the cursive nature of the script can be extracted through the window. The values of the seven invariant moments, ϕ_i , are computed as a function of windows movement .

These are shown in figure (5) for ϕ_1 to ϕ_7 as a function of window movement. It can be seen from the figure that the beginning and end the word in a sentence can easily be identified from the values of minima. The features of each word can be extracted from these plots related to a sentence. The technique is very effective for extraction features of a word in a given sentence.







© 2010, IJARCS All Rights Reserved

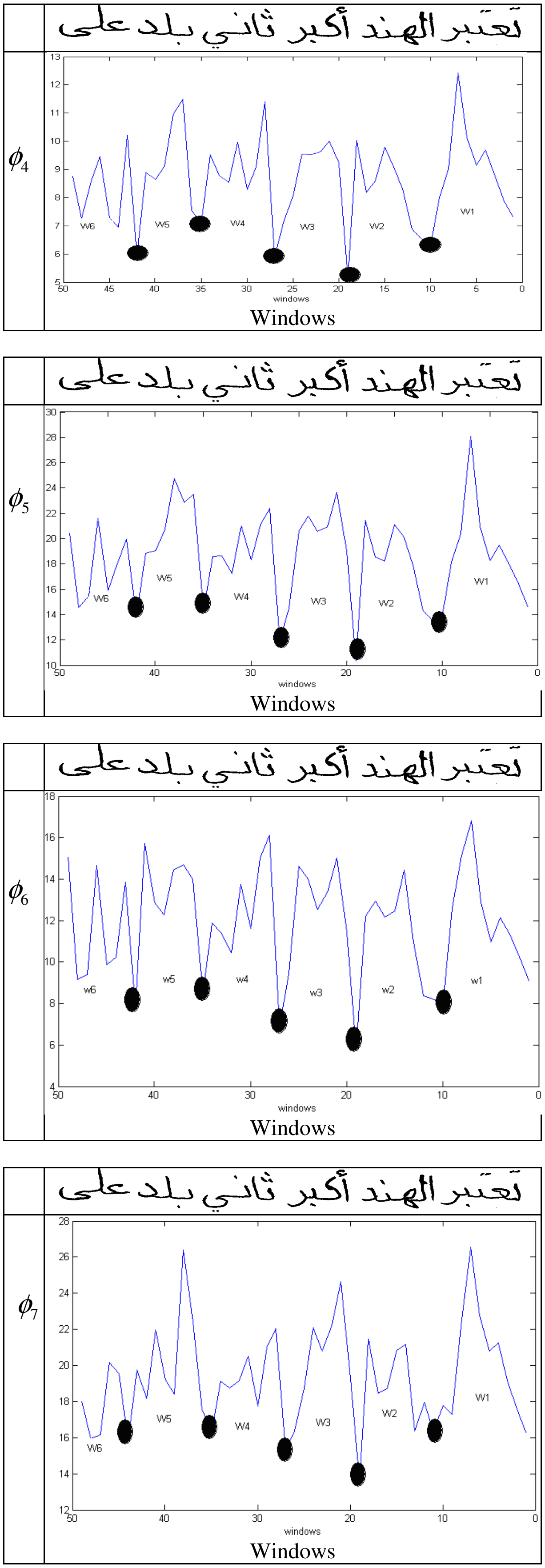


Figure 5. Plots of ϕ_1 to ϕ_7

The first line of the text given in figurer 3 is considered for the analysis and the words are separated using the maxima and minima approach. The result is illustrated in table II in which plots of feature vectors as a function of sliding window for each separated word (given on the right) are shown on the left side. Notice that (*α1*= word1),(*α2*= word2) …… (*α6*=word6)

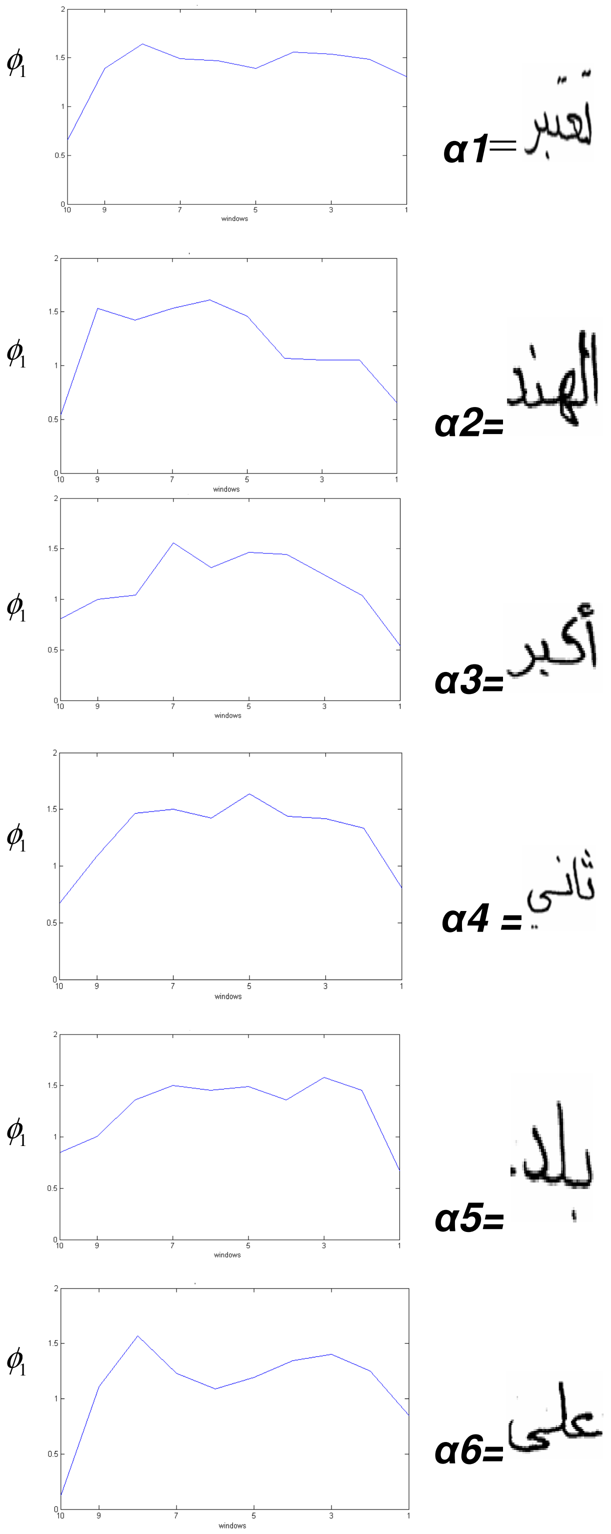


Figure 6. The words with its feature vectors for ϕ_1

F. *Recognition Algorithms*

The recognition system has been designed based on the vector approach using Minimum Distance Classifier (MDC) and Cosine (θ) Classifier (CC) as follows:

1. *Recognition Using Minimum Distance*

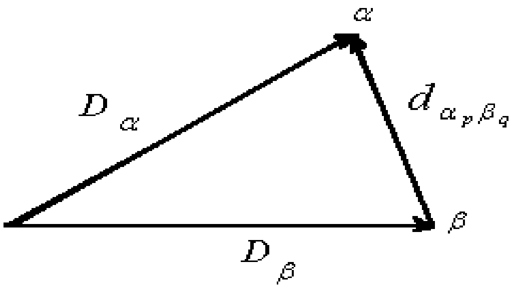


Figure 7. Distances between two features vectors (α_p , β_q)

The feature vector ϕ_i *(* α_p *)* corresponding to the word *(* α_p) is given as follows:

$$\phi_i(\alpha_p) = \left\{\phi_i^{(k)}(\alpha_p), \quad k = 1,2,3..........n_w\right\} \qquad (1)$$

where the word α is written by the p^{th} subject and n is number of windows for the word. In this work, as there are forty subjects, the value of *p is* 40 .In this vector space , the parameters for a given vector are computed as follows:-

- *The length of the vector corresponding to a word (* α_p *) in the feature vector space can be computed as follows :*

$$D_{\alpha_p} = \sqrt{\sum_{k=1}^{n_w} [\phi_i^{(k)}(\alpha_p)]^2} \qquad (2)$$

- *The distance between two features vectors corresponding to two words(* α_p *,* β_q *) , can be computed as follows*:

$$d_{\alpha_p\beta_q} = \sqrt{[\sum_{k=1}^{n_w} \phi_i^{(k)}(\alpha_p) - \phi_i^{(k')}(\beta_q)]^2} \qquad (3)$$

where $\phi_i^{(k)}$ (α_p or β_q) is the feature Corresponding the word (α_p or β_q) written by p^{th} , q^{th} subject respectively , $n_{w`}$, $n_{w'}$ represent number of windows for the word (α_p , β_q) respectively and *i* is the i^{th} invariant vector , *ḱ= ḱ (k, n_w, $n_{w'}$), if n_w = $n_{w'}$ and k =ḱ* , otherwise it is computed as follows :

$$k' = \frac{(n_{w'} - 1)k}{n_w + 1} + (\frac{n_w - n_{w'}}{n_w + 1}) \qquad (4)$$

The value of $\phi_i^{(k')}(\beta_q)$ is interpolated by using the polynomial method. The difference between two feature vectors corresponding to different words and different subjects are computed by using "Eq. (3)" the values are averaged over all the subjects for a given word.

These average along with standard division are summarized in the tables IV,V,VI and VII corresponding to ϕ_1, ϕ_2 and ϕ_7 respectively **.**

2. *Recognition using Cosine (θ)*

The angle value between two features vectors corresponding to two words (α_p , β_q) are computed as follows:

$$Cos(\theta_{\alpha_p,\beta_q}) = \frac{\sum_{k=1}^{n_w} \phi_i^{(k)}(\alpha_p) * \phi_i^{(k')}(\beta_p)}{D_{\alpha_p} D_{\beta_q}} \qquad (5)$$

To enhancement the recognition rate we also computed the cosine (θ) technique. For computing the angel values between two feature vectors corresponding to different words and different subjects by using i.e." Eq. (5)" .The values are averaged over all the subjects for a given word. These values of average along with standard division are summarized in tables VIII,IX,X,XI and XII corresponding to ϕ_1, ϕ_2 and ϕ_7 respectively.

IV. Results and Discussions

Experiments have been performed to test the above (AOCR) system. The developed Arabic texts recognition system has been tested using randomly selected texts. The system is designed in Matlab 7.0 for the recognition of Handwritten Arabic Naskh type and font nearly 14. The system developed is a single font, one size system. The system was gained 83.09% successful recognition rate by using minimum distances classifier technique and 92.86 % successful recognition rate for cosine(θ) technique respectively and the results obtained as follows:

A. *Experiment using Minimum Distance Classifier*

- Minimum Distances values obtained by Eq(3). For ϕ_1, ϕ_2 and ϕ_7 between all subjects are calculated and results cited in tables blue.

  Notice: *that we mean by (α1, β1=Word1, α2, β2=Word2.........and α6, β6=Word6), words are written by different Subjects and extracted form handwritten texts .*

© 2010, IJARCS All Rights Reserved

Table III. The Euclidean minimum distance matrix for subject1

Words	β1	β 2	β 3	β 4	β 5	β 6
α1	0	3.74	3.73	0.1	0.13	0.04
α 2	3.74	0	0.1	2.63	5.27	3.05
α 3	3.73	0.1	0	2.62	5.26	3.04
α 4	0.1	2.63	2.62	0	0.45	0.02
α 5	0.13	5.27	5.26	0.45	0	0.3
α 6	0.04	3.05	3.04	0.02	0.3	0

- Average and Standard Deviation for ϕ_1 between all words written by all subjects are summarized in table IV.

Table IV. Average values of d_{α_p,β_q} and correspond standard deviation for the word (α_p, β_q) with ϕ_1

Words		β1	β 2	β 3	B4	β 5	β 6
α1	AVE	0.22	2.65	5.19	5.09	12.52	5.61
	STDEV	0.26	1.65	6.13	4.30	14.70	6.80
α 2	AVE	2.02	0.36	1.64	1.56	6.32	3.67
	STDEV	1.34	0.34	3.41	1.56	8.27	1.50
α3	AVE	5.88	3.26	3.00	3.23	4.99	4.78
	STDEV	6.10	3.72	3.88	2.60	7.13	6.10
α 4	AVE	0.78	1.75	3.69	3.08	8.96	4.82
	STDEV	0.86	1.74	6.47	3.62	12.45	5.39
α 5	AVE	4.71	4.68	6.50	3.20	2.76	8.25
	STDEV	11.80	7.60	8.76	3.18	11.28	11.64
α 6	AVE	6.05	3.61	5.93	1.29	2.41	6.62
	STDEV	4.64	1.68	5.94	1.58	3.08	6.18

- Average and Standard Deviation for ϕ_2 between all words written by all subjects are summarized in table V.

Table V. Average values of d_{α_p,β_q} and correspond standard deviation for the word (α_p, β_q)with ϕ_2.

Words		β1	β 2	β 3	B4	β 5	β 6
α1	AVE	0.22	2.65	5.19	5.09	12.52	5.61
	STDEV	0.26	1.65	6.13	4.30	14.70	6.80
α 2	AVE	2.02	0.36	1.64	1.56	6.32	3.67
	STDEV	1.34	0.34	3.41	1.56	8.27	1.50
α3	AVE	5.88	3.26	3.00	3.23	4.99	4.78
	STDEV	6.10	3.72	3.88	2.60	7.13	6.10
α 4	AVE	0.78	1.75	3.69	3.08	8.96	4.82
	STDEV	0.86	1.74	6.47	3.62	12.45	5.39
α 5	AVE	4.71	4.68	6.50	3.20	2.76	8.25
	STDEV	11.80	7.60	8.76	3.18	11.28	11.64
α 6	AVE	6.05	3.61	5.93	1.29	2.41	6.62
	STDEV	4.64	1.68	5.94	1.58	3.08	6.18

- Average and Standard Deviation for ϕ_7 between all words written by all subjects are summarized in table VI.

Table VI. Average values of d_{α_p,β_q} and correspond standard deviation for the word (α_p, β_q) with ϕ_7

Words		β1	β 2	β 3	B4	β 5	β 6
α1	AVE	0.19	3.25	2.80	3.62	3.24	4.54
	STDEV	0.24	7.40	5.16	5.45	4.51	10.20
α 2	AVE	3.49	0.27	1.07	1.46	1.51	1.97
	STDEV	7.38	0.25	2.07	2.57	2.89	2.86
α3	AVE	4.91	2.89	2.14	1.06	1.51	5.52
	STDEV	7.18	3.45	3.61	0.59	0.89	7.13
α 4	AVE	2.37	1.35	0.55	0.72	0.75	3.08
	STDEV	4.69	2.58	0.47	0.93	0.42	5.22
α 5	AVE	2.12	1.47	0.65	1.13	0.83	2.97
	STDEV	4.38	2.93	0.84	1.52	0.79	5.60
α 6	AVE	5.71	2.60	2.94	2.44	2.59	2.15
	STDEV	10.34	2.98	4.98	5.46	5.77	4.13

The total percentage rate using Minimum Distance Classifier for all subjects along with all ϕ_i's are summarized in table VII.

Table VII. Recognition rate for all subjects and all ϕ_i's using Minimum Distances Classifier.

Subjects	Subject1	Subj.2	Subj.3	Subj.4	Average
Subject1	100	53.33	66.66	66.66	71.66
Subject2	53.33	100	86.94	89.9	82.54
Subject3	83.33	86.94	100	92.9	90.79
Subject4	66.66	89.9	92.9	100	87.37
Total percentage					83.09%

B. *Experiment using Cosine (θ) Classifier*

- Cosine (θ) (The angle values between tow vectors for ϕ_1, ϕ_2 and ϕ_7 between all subjects) using Eq.(5) and the results obtained summarized in tables VIII, IX, X and XI respectively.

Table VIII. Average values of $Cos(\theta_{\alpha_p,\beta_q})$ and correspond standard deviation for the word (α_p, β_q) with ϕ_1.

Words		β1	β 2	β 3	B4	β 5	β 6
α1	AVE	1.00	0.98	0.98	0.90	0.92	0.90
	STDEV	0.01	0.01	0.02	0.07	0.04	0.06
α 2	AVE	0.96	0.99	0.97	0.89	0.90	0.88
	STDEV	0.03	0.01	0.02	0.10	0.05	0.06
α3	AVE	0.92	0.97	0.98	0.88	0.90	0.91
	STDEV	0.10	0.01	0.02	0.08	0.03	0.07
α 4	AVE	0.97	0.98	0.97	0.88	0.90	0.90
	STDEV	0.02	0.01	0.03	0.11	0.06	0.06
α 5	AVE	0.97	0.96	0.96	0.91	0.97	0.91
	STDEV	0.04	0.04	0.05	0.07	0.04	0.05
α 6	AVE	0.88	0.82	0.87	0.81	0.86	0.88
	STDEV	0.03	0.10	0.04	0.13	0.11	0.11

- Cosine (θ) (The angle values between tow vectors for ϕ_2 along with all subjects are summarized in table IX)

© 2010, IJARCS All Rights Reserved

Table IX. Average values of $Cos(\theta_{\alpha_p,\beta_q})$ and correspond standard deviation for the word(α_p, β_q) with ϕ_2.

Words		*β1*	*β 2*	*β 3*	*B4*	*β 5*	*β 6*
α1	AVE	0.79	6.08	7.07	6.76	7.38	9.39
	STDEV	0.78	1.73	1.91	4.51	1.28	2.50
α 2	AVE	5.99	1.74	2.18	3.72	13.74	15.74
	STDEV	2.09	1.72	1.82	2.45	2.23	2.83
α3	AVE	7.64	2.05	2.14	4.09	15.40	17.40
	STDEV	0.99	1.33	2.18	2.08	1.17	2.64
α 4	AVE	2.96	2.89	3.10	4.52	10.71	12.72
	STDEV	2.37	1.95	1.61	4.28	2.55	2.46
α 5	AVE	7.19	12.90	12.81	13.58	1.93	2.16
	STDEV	2.35	2.60	3.04	5.09	1.45	0.82
α 6	AVE	8.48	14.19	14.10	14.87	4.95	3.74
	STDEV	5.60	5.39	5.34	7.45	2.12	3.76

- Cosine (θ) (The angle values between tow vectors for ϕ_7 along with all subjects are summarized in table X)

Table X. Average values of $Cos(\theta_{\alpha_p,\beta_q})$ and correspond standard deviation for the word(α_p, β_q) with ϕ_7.

Words		*β1*	*β 2*	*β 3*	*B4*	*β 5*	*β 6*
α1	AVE	0.99	0.98	0.99	0.98	0.99	0.99
	STDEV	0.01	0.01	0.01	0.98	0.99	0.01
α 2	AVE	0.96	0.99	0.97	0.98	0.97	0.98
	STDEV	0.03	0.03	0.01	0.01	0.01	0.01
α3	AVE	0.97	0.97	0.99	0.99	0.98	0.99
	STDEV	0.03	0.02	0.01	0.00	0.01	0.00
α 4	AVE	0.83	0.97	0.98	0.99	0.98	0.99
	STDEV	0.33	0.03	0.01	0.01	0.01	0.01
α 5	AVE	0.96	0.96	0.98	0.98	0.98	0.98
	STDEV	0.03	0.03	0.01	0.01	0.01	0.01
α 6	AVE	0.96	0.96	0.98	0.98	0.98	0.98
	STDEV	0.04	0.03	0.01	0.01	0.01	0.01

- The recognition rate using cosine (θ) for all subjects with all ϕ_i 's are summarized in table XI.

Table XI. Recognition rate for all Subjects and all ϕ_i 's using cosine (θ)

Subjects	**Subject1**	**Subj.2**	**Subj.3**	**Subj.4**	**Average**
Subject1	100.00	98.00	96.60	90.00	96.15
Subject2	98.00	100.00	86.94	86.20	92.79
Subject3	96.60	86.94	100.00	86.00	92.39
Subject4	90.00	84.33	86.20	100.00	90.13
Total percentage					92.86%

It should be noticed that the recognition rate for the ϕ_1, ϕ_2 *and* ϕ_7 are 66.66 %, 83.33% and 83.33% respectively .If all ϕ_i *'s* are calculated together the recognition rate is 83.09 % .

And by using cosine (θ) Classifier the recognition rate can enhancement to 86.94% , 96.60% and 98% using ϕ_1, ϕ_2 *and* ϕ_7 respectively.

In case we use all ϕ_i *'s* the recognition rate can increased up to 92.86 %.

V. Conclusion

In this paper, an attempt has been made to apply a technique based on windowing for segmentation the texts into atomic character or sub character, moment invariants technique has been used for feature extraction. The feature vector space represents the values of the invariant moments in sliding windows across a word / alphabet. The recognition system is based on the distance between two points in the features vectors space. The angle between two vectors feature vectors space has also been used for the recognition system. The system has achieved a success recognition rate of 83.09% using minimum *distances* classifier technique and for the cosine (θ) technique it is 92.86 %. We conclude that the cosine (θ) is more sensitive feature for the recognition.

Acknowledgement

This work has been supported by Ministry of Technical Education and Vocational Training Republic of Yemen.

VI. References

[1] V. Bhagile, M. Al-Baraq, R.J. Ramteke, S.C.Mehrotra, "Recognition System for Arabic Printed Numeral :A size Independent Approach", IEEE IDICON . Annual Symposium of IEEE Bangalore India Section 6-8 sept.,2007.

[2] A. Elgammal. Bilingual " (Arabic / English) Document Image Analysis System With Font Independent Arabic Text Recognition". Master's thesis, Computer Science Department – Alexandria University, 1996.

[3] M. Sarfraz , S.Nazim and A. Al-Khuraidly " Offline Arabic Text Recognition System" Proc. of the 2003 International Conference on (GMAG'03) 0-7695-1985-7/03 $17.00 © 2003 IEEE.

[4] M. Fakir and M.M Hassani, "Automatic Arabic Characters recognition by moment invariants", Colloqu International de telecommunications, Fes, Morocco, 1997, pp 100 –103.

[5] M. K. Hu, "Visual Pattern Recognition by Moment Invariant", IRE Trans. On Information Theory, vol. IT – 8, 1962, pp 179-187.

[6] R. J. Ramteke, P. D. Borkar, S. C. Mehrotra, "Recognition of Isolated Handwritten Numerals: An Invariant Moment Approach", at International Conf. Cognition and Recognition (ICCR 2005),Mysore, (Karnataka), India on 22nd – 23rd Dec. 2005.

[7] R. J. Ramteke, S. C. Mehrotra, "Feature Extraction Based on Moment Invariants for Handwriting Recognition" at 2006 IEEE Int. Conf. on Cybernetics and Intelligent System (CIS), Bankok, Thailand on 7th – 9th June 2006. (Finalist for Best Student Paper Award).

[8] A. Amin, A. Kaced, J. P. Haton and R. Mohr, "Handwritten Arabic Characters Recognition by the IRAC System",

Proc. 5th Int. Conf. on Pattern Recognition, Miami, 1980, pp. 729-731.9. I. S. I. Abuhaiba, "Recognition of Off-Line Handwritten Cursive Text," Ph.D. thesis, Department of Electronic and Electrical Engineering, Loughborough University, Loughborough, U. K., 1996.

[9] B. Parhami and M. Taraghi, "Automatic Recognition of Printed Farsi Texts", Pattern Recognition, 14(1– 6) (1981), pp. 395-403.

[10] H. Al-Muallim and S. Yamaguchi, "A method of recognition of Arabic Cursive Handwriting", IEEE Trans. Pattern Anal. Mach. Intell. PAMI – 9, 1987,pp 715-722.

[11] R. Ramsis, S.S. El-Dabi, and A. Kamel, "Arabic Character Recognition System", IBM Kuwait Science .Centre, Report , No.KSC027, 1988.

[12] A. Amin and J.F. Mari, "Machine Recognition and Correction of Printed Text," IEEE Transactions on Systems, Man, and Cybernetics, 19(5), (1989), pp. 1300-1306.

[13] S. Al-Emami and M. Usher, "On-line recognition of handwritten Arabic characters," IEEE Trans. on Pattern Analysis and Machine Intelligence, 12(7), pp. 704-710, July 1990.

[14] B. Taconet, A. Zahour, and A. Faure, "A New Global Off-Line Recognition Method for Handwritten Words", in From Pixels to Features III: Frontiers in Handwriting Recognition, ed. S. Impedovo and J.C. Simon., Amsterdam: Elsevier Science Publishers B.V., 1992, pp. 327-338.

[15] I. S. I. Abuhaiba, "Recognition of Off-Line Handwritten Cursive Text," Ph.D. thesis, Department of Electronic and Electrical Engineering, Loughborough University, Loughborough, U. K., 1996.

[16] A. Amin," Recognition of printed Arabic text based on global features and decision tree learning techniques", Pattern Recognition, 33(8), August 2000, 1309–1323.

[17] Toufik Sari "Off-line Handwritten Arabic Character Segmentation Algorithm". ACSA Proceedings of the Eighth Int. Workshop on Frontiers in Handwriting Reco. (IWFHR'02) 0-7695-1692 , IEEE 2002.

[18] A. Nosary, T. Paquer, L. H ,A.Bensefaia "Handwritten Text Recognition Through Writer Adaptation" .Proceedings of the Eighth International Workshop on Frontiers in Handwriting Recognition (IWFHR'02) 0-7695-1692-0/02 $17.00 © IEEE 2002.

[19] M.S. Korsheed "Recognition Handwritten Arabic Manuscripts Using A Single Hidden Markov Model" Elsevier B.V. All rights reserved.doi:10.1016/S0167-8655(03) see front matter 2003.

[20] Ibrahim S. I. Abuhaiba "A Discrete Arabic Script For Better Automatic Document Understanding," April 2003 The Arabian Journal for Science and Engineering, Volume 28, Number 1B.

[21] M. Syiam, T. M. Nazmy "Histogram Clustering And Hybrid Classifier For handwritten Arabic characters Recognition", Proc. of the 24th lasted Inter. multi-conference in signal processing ,pattern reco. and App. feb-15-17,2006 ,Australia.

[22] A. Amin , N al-darwish, "Structural Description To Recognizing Hand-Printed Arabic Characters Using Decision Tree Learning Techniques", proceeding of the international Journal of Computer and Applications volu. 28, No.2,2006 , paper no.202-1551.

[23] V. Bhagile , M. Al-Baraq, R.J. Ramteke, S.C.Mehrotra, "Arabic Printed Numeral Recognition System :A Block Based Approach ", International Conference on ,ACVIT-07,28th – 30th Nov. IEEE 2007.

[24] Rudra Pratab, "Getting started with Matlab 7", copyright 2006 Oxford University press ,In first Indian Edition in 2006.

[25] A. Amin, "Off-Line Arabic Character Recognition system State of the Art", Pattern Recognition, Vol. 31, No. 5, 1998, pp 517-530.