PRINTED ARABIC TEXT RECOGNITION ALGORITHMS

طرق ادراك النصوص العربية المطبوعة

Mofreh M. Salem * Aida O. Abd El-Gwad *
Fatma E. Z. Abou Shadi ** Mohamed Y. Keshk *

* *Computers & Control Dept.,* ** *Communications Dept.,*
Faculty of Engineering, El-Mansoura University,
El-Mansoura, EGYPT

ملخص البحث

يعتبر تمييز الحروف والنصوص المطبوعة من اهم المجالا ت فى
الا بحاث الخاصة بادراك الا نماط ، وفى السنوات الا خيرة حدث تطور كبير
فى الا لا ت القارئه للنصوص اللا تينية والصينية وغيرها ، اما الا لا ت
القارئة للنصوص العربية فلم يبدا تطورها حتى الا ن وهذا يرجع الى
طبيعة مكونات اللغة وقلة الا بحاث فى هذا المجال حيث ان معظم الا بحاث
اهتمت بادراك وتمييز الحروف المفردة والقليل جدا اتجه الى تمييز
النصوص . لذلك ففى هذا البحث تم شرح وتحليل نظام التمييز الكامل
بجميع مراحلة وتم تنفيذ ثلا ثة من الطرق والتى يمكن تلخيصها كما وردت
فى اصول الا بحاث المنشورة كالا تى :- الطريقة الا ولى لا دراك النصوص
المطبوعة باستخدام سمات رياضيه وماسح يدوى - الطريقة الثانية
لا دراك الحروف المطبوعة وتستخدم خليط من السمات الرياضية والبنائية
مع المسح التخطيطى - الطريقة الثالثة لا دراك الحروف المكتوبة بخط
اليد باستخدام السمات البنائية مع المسح التخطيطى. ونظرا لا ن هذه
الطرق الثلا ث تتعامل بالطريقة اليدوية لذا فقد تم تعديل هذه الطرق
ليمكنها الا دراك الا وتوماتيكى للنصوص المناسبة لهذا العصر وهذا
بالا ضافة الى بعض التعديلا ت الا خرى فى الطريقة الثانية والثالثة ، و
قد تم اختبار هذه الطرق الثلا ث المعدله معمليا وتم تحليل ومقارنة
النتائج الواردة

ABSTRACT

This paper presents three efficient and reliable algorithms for the automatic recognition of printed arabic text. The first algorithm is based on mathematical features, the second uses structural and mathemtical features, while the third uses structural features. The performance of each algorithm is experimentally tested under different conditions. Comparisions between the results are also made.

Keywords : Pattern Recognition, Arabic text, Computer applications

1 INTRODUCTION

One of the major applications of pattern recognition is the characters recognition. Over the two past decades, in spite of the progress of machine recognition techniques of Latin and Chinese characters, machine recognition of Arabic text has remained almost untouched. This is due to the arabic language nature where the form of a cursive arabic character is changed according to its position within a word (beginning, middle, end or isolated). The size (height or width) of the character differs from one character to another. Also, the character has a main body and may include one of five different stress marks (such as dots and hamza). The main body shape of some characters may be the same, and the stress mark may be over or under the main body. So, arabic character can be recognized by classifying both its main body shape and its stress mark.

The system of arabic text recognition consists of four main stages; input, preprocessing, feature extractor and classifier. Firstly, via on-line system the physical image of arabic text can be transferred to the microcomputer memory through graphic tablet [4,6,7,8,14], video camera [5] or manual scanner [15], to convert the physical image into a matrix in a binary form. The preprocessing stage is concerned with the image enhancment to process the stored digtized image. In this stage, the input text is divided into separate lines by detection the blanks between the lines on their projection on a vertical axe. Words or subwords in a line are separated by enough spaces, so, they can be isolated (word or subword) through a projection on horizontal axe. Each word, can then be segmented into isolated characters. In general, a segmentation error will lead to an odd character shape not belonging to any character set, then misrecognition occurs. This stage may also include some additional operations, (such as thinning, smoothing, and normalization) to make the data more suitable for next stage. Thinning reduces the number of points which represent the image, and then reduces the processing time. Smoothing corrects any deformation and eliminates blurring which are resulted from poor sampling system or transmision channel. Normalization is a scaling method to convert the actual input characters to an standerd matrix [m x n], where m and n are the width and hight of the standerd matrix, the values of m and n may be adjusted by experiment [6]. Feature extraction plays a central role in this filed and it is an essential stage in the recognition system, where features are extracted from the boundary of each character. It is crucial to choose features that effectively discriminate between different characters and at the same time allow for small variations of different patterns of the same character. These two goals are, however, usually conflicting and this may be one of the reasons why feature extraction problem is still a challenging problem. Features of the arabic characters can be classified into two categories; Structural and Statistical (mathematical) features. Generally, mathematical features are suitable for printed characters recognition. Such features may be used alone, but it is preferable to use assist structural features to reduce the recognition time. In the last stage each character is classified and then recognized, using extracted features and the information given by any accompanied stress marks.

Recently, some algorithms [4-19], are published in the field of arabic recognition. However, the choice of an algorithm for practical application is difficult and requires careful analysis. This paper will trace the implementation of three recent algorithms. The first [15], is based on mathematical features to handle the printed arabic text. In the second algorithm [6], the recognition of isolated printed characters is carried out through structural and mathematical methodology. While the third algorithm is proposed in [14], to recognize isolated handwritten arabic characters using structural features. The above three algorithms are modified and implemented to recognize the printed arabic text, where, the segmentation stage is added to the second and the third algorithms. Moreover, data are introduced to all algorithms as printed text through on-line automatic scanner. The main contribution of this paper is to bring out the salient characteristics of these algorithms. So, the robustness of each algorithm is tested under different conditions. The results are obtained experimentally and the comparison between the performance of the algorithms is reported.

2 ALGORITHMS

In this section, three recent algorithms will be reformulated to recognize arabic text. Segmentation stage is added to the orginal form of the second and the third algorithms to recognize the printed arabic text instead of printed or handwritten arabic isolated characters.

2-1 ALGORITHM (A)

It was proposed in [15] for printed arabic text recognition, using manual scanner. It includes a segmentation technique to divided the text into lines using their projection on vertical axe as shown in figure (1), the line is then divided into words using their projection on horizontal axe as shown in figure (2). The word can then divided into isolated characters. Segmenting the word into isolated characters is a very difficult task in the recognition system. Segmentation of a cursive word is performed by obtaining the outer contour of the word. The boundaries between different characters of the word are consequently detected using basically the information contained in the obtained contour.

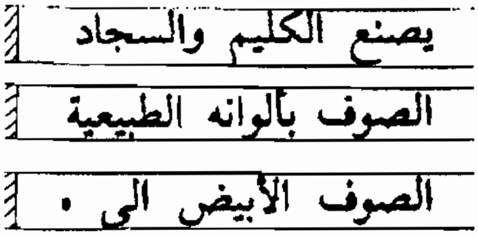

Fig. (1)

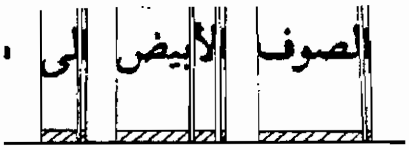

Fig.(2)

The segmentaion process is completely based on the calculation of the contour height (h) which is the distance between the two extreme points of the intersection of the contour with the vertical line. This contour height is usually very small at the boundaries between different characters. Regions, where the contour height is small, will be referred as "silent" regions. Once a closed curve contour is fed to the algorithm, a window of width (w) is moved in order to scan the contour horizontally from right to left. For each position of window, the average vertical distance (b_{av}) is calculated across the window. A "silent" region is detected if $h_{av} < T$, i.e.

when the average vertical distance over the window is less than a certain preset threshold distance T. The end of a character is detected if a silent region starts and the character width is greater than a certain minimum distance which determined in the training phase. On the other hand, the beginning of a new character is detected if a silent region, that has been previously detected after the end of the preceding character, ends. Following these two rules, the beginning and end of consecutive characters naturally alternate. The segmentation technique begins by assigning the right most pixel of the contour as the beginning of the first character and stops when the left most pixel is assigned as the end of the last character of a word or a subword. In general, a segmentation error will lead to an odd character shape not belonging to any character set, then misrecognition will be occured.
The algorithm uses a mathematical featuers extractor, which is suitable for printed characters, where the shape of the character is fixed. These features can be driven from the binary matrix of the character or from the transformation of the character coordinates of the input pattern vectors. These features may be correlation, variance, mean error probability or any suitable statistical relation. The Fourier series coefficients are evaluated for the two coordinates sequences x(m), y(m). The Fourier series expansion of x(m) and y(m) is defined as :

$$x(m) = \sum_{n=0}^{N-1} a(n)\exp(j2\Pi nm/N)$$

$$y(m) = \sum_{n=0}^{N-1} b(n)\exp(j2\Pi nm/N)$$

where N is the number of points on the contour. The sequences a(n) and b(n) are the copmlex Fourier coefficients of x(m), and y(m) respectively and can be obtained from :

$$a(n) = \left(\sum_{m=0}^{N-1} x(m)\exp(-j2\Pi nm/N)\right) / N$$

$$b(n) = \left(\sum_{m=0}^{N-1} y(m)\exp(-j2\Pi nm/N)\right) / N$$

the Fourier series coefficients of the complex sequence take the form :

$$z(m) = x(m) + j\,y(m)$$

or the forme :

$$c(n) = a(n) + j\,b(n)$$

It has been found that only the first three Fourier coefficients are sufficient to reconstruct arabic characters with reasonable accuracy.

The classification stage is based on clustering technique and K-Nearest Neighbour classifier. The K-Nearest Neighbour rule classifies the sample character U by assigning it to the class most frequently represented among the K nearest samples. The performance of this rule is expected to improve as the number of samples increased. When K = 1, the rule reduced to one of the most widely used non-parametric techniques, (i.e. the nearest neighbour classifier). First, classification depends on the nearest neighbour. Next, clustering technique is used to find the natural grouping in a set of data. Natural grouping means that the samples in one cluster are closer to another than samples in other clusters. Similarity between samples can be measured by calculating distances between them using any suitable distance metric. If the chosen distance is a good measure of dissimilarity, then one would expect the distance between samples in the same cluster to be significantly less than the distance between samples in different clusters. To partition a set of samples into clusters, a criterion function, J_E is usually calculated. One generally seeks the clustering that optimizes the criterion function. The criterion function is usually chosen such that the between-cluster distances are large and the within-cluster distances are small. The most widely used criterion function is the sum of the squared error and défined by the following equation :

$$J_E = \sum_{i=1}^{z} \sum_{u \in S_i} | U - \mu_i |^2$$

where z is the number of clusters, and μ_i is the mean of samples in cluster S_i. The value of J_E depends on how the samples are grouped into clusters and an optimal partition is defined as the one that minimizes J_E. Figure (3) shows a clustering problem involving six samples. Level 1 shows the six samples as belonging to a single cluster. partitioning of this cluster is performed in the higher levels by successively splitting clusters until each sample is included in a separate cluster. Once a training set is partitioned into clusters, an unknown sample may be classified by tracing the corresponding group. Such procedure is called the hierarchical classifier and it usually leads to very fast and accurate classification results. At the end of this stage the unknown sample is recognized.

2-2 ALGORITHM (B)

This algorithm was proposed in [6] for isolated printed arabic character recognition. The data was introduced to the system using graphic tablet. In order to recognize printed text, segmentation stage is added to this algorithm The segmentation technique is similar to that explained in the above algorithm (A). This algorithm uses both the strcurual features and mathematical features. Suitable primitives are chosen together with relationships among them in order to describe the profile, topological and geometrical properties of the characters. Therefore, these features are used for printed and handwritten characters. The most common parameters of these features for Arabic character recognition are :

- Number of strokes.
- Number of cusbs in main stroke.
- Presence of closed loop.
- Presence of hamza and it's position.
- Size of main stroke.
- Zone location.
- Number and position of dots.

The features for this algorithm are selected as follow :

- Character zone (G).
- Number of dots (ND).
- Availability of Hamza(H).
- Correlation coefficient (R).
- Availability of dots (D).
- Zone of dots (ZD).
- Zone of Hamza (ZH).

Where the value of G, D, ND, ZD, H and ZH are equal to one (true) or zero (false), while the value of R is ranged from 0 to 1.

The classification stage is done through partitioning the feature space into classes (groups), one class for each category. The selection of classification rule and it's parameters depend upon the type of characters. Thus, it is important to know how the classifier is close to the performance of the best discrimination of any class. This algorithm uses a linear classifier (K-Nearest Neighbor) where a satisfied test is made to compare between the input characters and prestored characters to identify the unknown one. This test is based on the following equation :

$$P = (W_1.G + W_2.R + W_3.D + W_4.ND + W_5.ZD + W_6.H + W_7.ZH)/N$$

where W_1, W_2, ... W_7 are weight factors to be adjusted by experiments according to the importance of each parameter and N is the number of identification parameters. The identfied character is defined as the character satisfying the following test; $P > TH$; where TH is a certain threshold, which can be adjusted by experiment according to the perfection of the written character. If more than one character, in a given group, satisfies this test, then, the most probable character is the one with the greatest value of P. If no character satisfies this test, then a search should be performed for all characters in all groups.

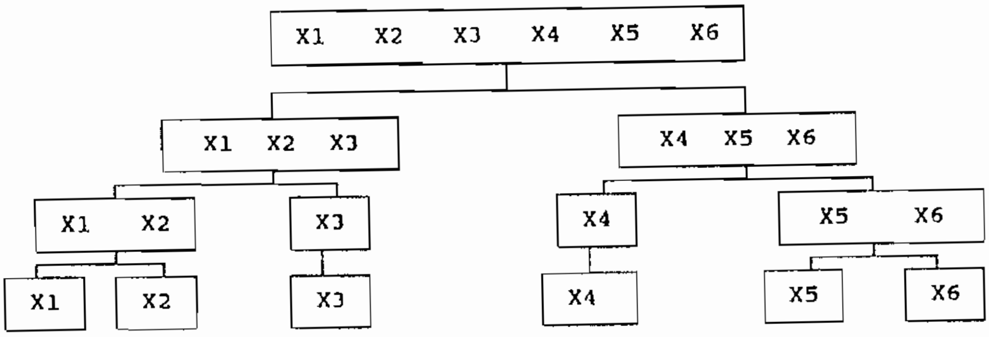

Figure (3): Culstering technique

2-3 ALGORITHM (C)

It was proposed in [14] for handwritten isolated arabic character recognition. It uses graphic tablet. The character is already come isolated, therefore the segmentation stage does not exist in the original alogorithm. While in our work the segmentation stage is added as in algorithms (A) and (B) to recognize printed arabic text.

The feature extractor in this algorithm, is based on feature vector FV of a character. Where FV consists of four features as given below :

FV = [ND, PD, NS and SS]

Where

ND = number of dots of the character,
= 0 for DAL, WAW, and RAA
= 1 for BAA, KHA, and GHAIN
= 2 for QAF, TAA, and YAA
= 3 for SHEEN and THAA

PD = the relative position of the dot (or dots) with respect to the character,
= 1 means dot above as; KHAA and GHAIN
= 2 means dot within as; GEEM and NOON
= 3 means dot below as; BAA and GEEM
= 4 means dot within or above as; TAA and NOON

NS = the number of secondary strokes.
= 0 (zero secondary stroke) as; WAW, DAL
= 1 (one secondary stroke) as; ZAH , LAMALEF

SS = the slope of secondary stroke.
= 0 (from ≥ 0° to ≤ 90°) such as; KAF.
= 1 (from ≥ 90° to ≤ 180°) such as; LAMALEF.

A hierarchical classifier (tree) is used in this algorithm. In this scheme the characters will be divided into a separate classes, each class contain some or many characters which have the same featuers. The characters in each class are sub-divided into sub-classes according to another common featuers, and the sub-classes may be divided into sub-sub-classes, and so on, as shown in figure (3). The string representation of the unknown character is matched against the prototype characters of the matched F.V.(F1, F2, F3, F4). In this stage an inexact matching classification algorithm (Minimum Distance Classification Approach) is used. It does not need a more powerful algorithm since the distance between the characters in the prototype level are relatively far from each other. Finally the unknown character is recognized from the characters having the same matched prototype, a K-Nearest Neighbor Recognition Rule is used. In this stage a powerful algorithm is needed since the distances between the characters in this level are relatively small.

3 EXPERIMENTAL RESULTS

Input data is transfered through the on-line automatic scanner (IBM model 3117) , with high resolution graphic mode of (640 x 320), and the test is performed using IBM personal computer. In this case, complicated processes are needed to convert the physical image into a binary matrix and stored for manipulation. This way will save effort and time, and give more accurate results. The sample were introduced to the three systems as a printed arabic text as shown in figures (1), (2). The accuracy and recognition time which resulted from the exprimental results are reported in figures (4), (5) respectivly. Comparison between the performance of the algorithms shows that the second algorithm is not suitable for parctical text application, and the first algorithm is the most suitable one for on-line arabic text recognition, where it takes one fifth of the recognition time taken by the third algorithm for the same accurecy.

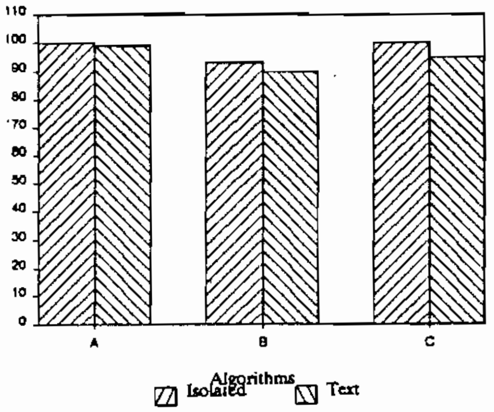

Fig.(4) Recognition accuracy

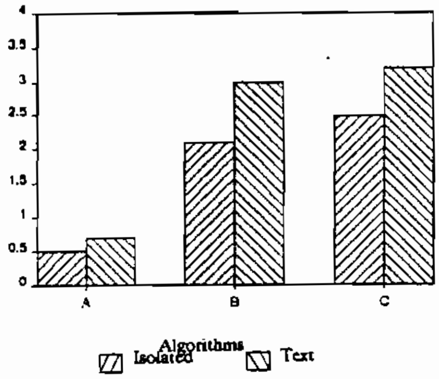

Fig. (5): Recognition time

4 CONCOLUSION

This paper traced the implementation of three recent algorithms. The first, is based on mathematical feature to handel the printed arabic text. While the second algorithm is to recognize isolted handwritten arabic characters using structural features. The third algorithm recocnizes isolated printed characters through strctural and mathematical methodology. In order to recognize printed arabic text, the second and the third algorithms are modified by adding segmentation technique. The experimental results indicate that, the first algorithm produces the highest accuracy in smallest recognition time.

REFERENCES

[1] Richard O. Dva, Peter E. Hart, " Pattern Classification and scene analysis ", Wiley-inter. Publication, New York, 1973.

[2] Tazy Y. Young, " Classification, estimation and patteren recognition ", Amrican Elsevier Publishing Co., New York. 1974.

[3] Rafael C. Gonzalez, Pavl wintz, " Digital image processing ", addison-wesley Publishing Co., 1987.

[4] Adnan Amin, " Arabic handwriting recognition and understanding ", WorkShop on computer processing and transmission of arabic language, Kuwait, pp 1-37, April,1985.

[5] Hussin Almuailim and Shoichiro Yamaguchi, " A Method of recognition of arabic cursive handwriting ", IEEE Transactions on computer, vol. 9, No. 5, pp 5-722, sep. 1987.

[6] M.F. Tolba, S.A. Wahab, A. Salem, " A Recognition algorithm for printed Arabic Character ", computer arabization, Egyption Computer Society, pp 5-8, April 1988.

[7] Adnan Amin, Azzdine Kaced, J. Haton, " Handwritten arabic character Recognition By the I.R.A.C. System", pp 9-11, ibid.

[8] Adnan Amin, "Machine recognitionof handwritten arabic words by the IRAC II system", pp 12-14, ibid.

[9] F. Haj Hassan, " Arabic character recognition", pp 15-20, ibid.

[10] A. Nouha, N. Ula, and A. Sharaf, " A Boolean recognition technique for typewritten arabic character set ", pp 21-28, ibid.

[11] A. Sharaf, A. Nouha and A. Uin, " Algorithms for features extraction : A Case study for the arabic character recognition ", pp 29-42, ibid.

[12] M. Sharkawy, M.F. Tolba and E. Shaddad, " Fourier descriptors for printed arabic character", pp 43-52, ibid.

[13] Suhail Saad Aiiah and Sad G. Yacu, " Design of an arabic character Reading Machine ", pp 53-69, ibid.

[14] M. El-Wakil and Shoukry, " On-line recognition of handwritten Isolated arabic characters", pp 70-82, ibid.

[15] Talaat S. El-Sheikh and Ramez M. Guindi, " Computer recognition of arabic cursive scripts",Pattern Recognition, Vol.21, pp 293-302, 1988.

[16] Tolba and E. Shaddad," Segmentation and recognition of printed arabic character " , Second conference of arabic computational linguistics, pp 490-508, kuwait, nov. 1989.

[17] Yossry M. Youssef ", An Expert knowledge-based system for arabic OCR", 14' th International Conference for Statistics, Computer since, Social and Computer since, Social and demographic Research, vol. 4, pp 227-245, Cairo, Egypt, Mar. 1989.

[18] Salwa H. El-Ramly, Mohamed A. El-Hamalaway ", A Language depandant arabic character recognition approach", pp 247-254, ibid.

[19] Salwa H. El-Ramly and Mohamed A. EL-Hamalaway ", A new font for arabic character semplities recognition procedure " , pp255-261, ibid.