Arab Gulf J. Scient. Res., **8** (2), pp. 49-59 (1990)

Recognition of Handwritten Arabic Characters *via* Segmentation

H.S. Al-Yousefi and S.S. Udpa

Department of Electrical Engineering, Colorado State University, Fort Collins, CO 80523 U.S.A.

ABSTRACT. This paper introduces a new approach for segmenting handwritten Arabic characters to improve the recognition of these characters. The proposed approach involves, as a first step, digitization of the segmented characters. The dots and zigzags (secondaries) are then segmented and identified separately. This reduces the recognition issue from a 28 to a 20-class problem. The recognition of the primary part of the characters is achieved using features derived from the moments of the horizontal and vertical projections normalized with respect to the zero-order moment. Classification of the primary characters is accomplished using quadratic discriminant functions. Results showing considerable improvement in classification are presented.

The subject of handwritten character recognition has been of considerable interest to many researchers. The subject is of great challenge since even the best optical recognition system available now (human eye) is expected to make errors of about 4% (Suen, Berthod, and Mori 1980). However, a minimal performance expected of a practical system is 99.9% (Kahan, Pavilidis, and Baird 1987). This implies an average of 3 misrecognized characters for a typical page. Errors are caused by infinite variation in shapes and orientation resulting from writing habits, style, social and psychological state of the writer as well as other factors such as the writing instrument , writing surface and method of optical scanning.

Several algorithms have been proposed for the recognition of Latin, Kanji, and Chinese characters. However, the characteristics of the Arabic language do not allow a direct application of these algorithms since the recognition depends on the type of characters being recognized. Some of the characteristics of the Arabic characters are:

a) Arabic is written cursively from right to left.

b) There are 28 characters in the Arabic language, but each character has between two to four different forms depending on its position in the word or subword. This will increase the classes to be recognized from 28 to 100 classes as shown in Fig. 1.

Name	EF	MF	BF	IF
DHAD	ـض	ـضـ	ضـ	ض
TTA	ـط	ـطـ	طـ	ط
DHA	ـظ	ـظـ	ظـ	ظ
AIN	ـع	ـعـ	عـ	ع
GHAIN	ـغ	ـغـ	غـ	غ
FA	ـف	ـفـ	فـ	ف
GAF	ـق	ـقـ	قـ	ق
KAF	ـك	ـكـ	كـ	ك
LAM	ـل	ـلـ	لـ	ل
MEEM	ـم	ـمـ	مـ	م
NOON	ـن	ـنـ	نـ	ن
HA	ـه	ـهـ	هـ	ه
WAW	ـو			و
YA	ـي	ـيـ	يـ	ي

Name	EF	MF	BF	IF
ALIF	ـا			أ
BA	ـب	ـبـ	بـ	ب
TA	ـت	ـتـ	تـ	ت
THA	ـث	ـثـ	ثـ	ث
JEEM	ـج	ـجـ	جـ	ج
HHA	ـح	ـحـ	حـ	ح
KHA	ـخ	ـخـ	خـ	خ
DAL	ـد			د
THAL	ـذ			ذ
RA	ـر			ر
ZA	ـز			ز
SEEN	ـس	ـسـ	سـ	س
SHEEN	ـش	ـشـ	شـ	ش
SAD	ـص	ـصـ	صـ	ص

Fig. 1. Arabic alphabet in all its forms (beginning form BF, middle form MF, end form EF, and isolated form IF).

c) Characters may be joint to each other according to certain rules. Some characters can only appear on the beginning or the end of a word or subword. An Arabic word can have one or more subwords.

d) Sixteen of the Arabic characters have a dot, group of dots, or a zigzag (secondaries) either above or below or inside the primary character. The position of these secondaries is very significant for recognizing the characters.

Approaches to the recognition of handwritten Arabic characters has been mostly concentrated on the use of geometrical and topological features. Amin, Kaced, Haton and Mohr (1980) and Amin (1982) present a system called IRAC (Interactive Recognition of Arabic Characters) which recognizes isolated Arabic words written on a graphic tablet connected to a minicomputer with the use of a dictionary of limited words. Badi and Shimura (1982) show a method for the recognition of Arabic handprinted isolated script based on the extraction and identification of curves in the script. Recently, Almuallim and Yamaguchi (1987) introduced a method for the recognition of Arabic cursive handwriting.

In this paper a new segmentation approach for reducing the number of classes to be recognized is explained. Steps involved in the recognition process are shown in figure 2 and briefly described in the following sections.

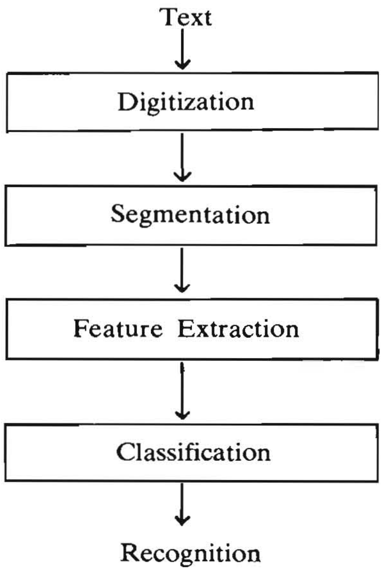

Fig. 2. A typical pattern recognition system

Digitization and Preprocessing

The images obtained from the text are converted to a binary matrix of "zeros" and "ones". The "zero" elements represent the background while the 'one' elements represent the character. Some assumptions have been made to insure proper identification of the secondaries:

a) Secondaries are located either below or above the character. Secondaries that are inside the character is assumed to be a part of the primary character.

b) Characters are vertical allowing for a reasonable tilt, not allowing secondaries that are not part of the primary to be included as part of the character.

The vertical and horizontal projects of the character are then obtained. The vertical and horizontal projections are defined as:

$$v(x) = \sum_{i} g(i,j) \qquad (1)$$

$$h(x) = \sum_{j} g(i,j) \qquad (2)$$

where g(i,j) is either 1 or 0. The vertical and horizontal projections of the letter (TTA) are shown in Fig. 3.

The secondaries are separated from the main characters by means of projections. The discontinuities in the functions v(x) and h(x) indicates a secondary to be separated. These secondaries are identified separately. A special program was written to separate, identify and eliminate the secondaries after the recognition process is completed. A more detailed discussion of this algorithm is presented in the section dealing with recognition.

Segmentation

The purpose of segmentation is to isolate the secondaries for the purpose of recognition. During the scanning of the segmented character, the number of dark points is added to find the total number of dark points. Another scan is made to find the horizontal and vertical projections. During the scanning of rows, the number of dark points is counted until an empty row is encountered. This indicates a separate segment. The process is continued until the entire image is scanned giving each segment a number (1,2,3 or 4). Each segment is scanned and the

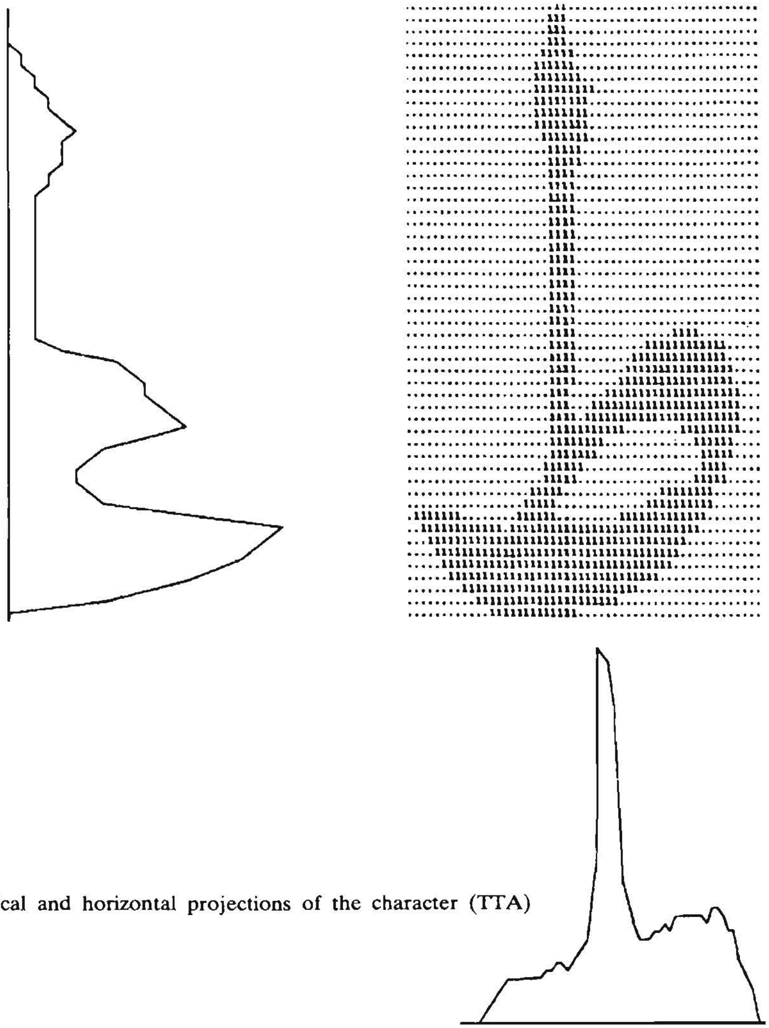

Fig. 3. Vertical and horizontal projections of the character (TTA)

number of dark points (nd) and the number of rows included in that segment (nr) are counted. The segment that has more than half the total points is classified as the primary character. The primary segment can only be the first or the last segment since Arabic characters can have secondaries either above or below but not both. The number of dots in each segment and the number of rows are very important factors in the recognition process of the secondaries. After getting this information about the secondaries, they are eliminated leaving only the primary character. Comparing the horizontal projections of the original character and the primary character, the number of segments, length of segments and the number of dark points in each segment is found.

Feature Extraction

The central moments up to the 4th degree are obtained from the horizontal and vertical projections of the primary character. The central moments are defined as

$$u_r = \int_{-\infty}^{\infty} (x-\bar{x}) \; g(x)dx \tag{3}$$

where g(x) can be either v(x) or h(x). The moments are then normalized with respect to μ_0. The features are extracted from the normalized moments. These features are:

1. $KV = \frac{\mu_4}{\mu_2^2}$ (for the vertical projection) (4)

2. $KH = \frac{\mu_4}{\mu_2^2}$ (for the horizontal projection) (5)

3. $SV = \frac{\mu_3}{(\mu_2)^{1.5}}$ (for the vertical projection) (6)

4. $SH = \frac{\mu_3}{(\mu_2)^{1.5}}$ (for the horizontal projection) (7)

5. $NV = \frac{\mu_3}{(\mu_4)^{.75}}$ (for the vertical projection) (8)

6. $NH = \frac{\mu_3}{(\mu_4)^{.75}}$ (for the horizontal projection) (9)

7. $LW = \frac{\mu_1^{(\text{vertical})}}{\mu_1^{(\text{horizontal})}} = \frac{\mu_{01}}{\mu_{10}}$ (10)

8. $VR = \frac{\mu_2^{(\text{vertical})}}{\mu_2^{(\text{horizontal})}} = \frac{\mu_{02}}{\mu_{20}}$ (11)

$$9.\ VV = \frac{\mu_4(\text{vertical})}{\mu_4(\text{horizontal})} = \frac{\mu_{04}}{\mu_{40}} \qquad (12)$$

Features 1 and 4 are measures of kurtosis which is a measure of flatness of the distribution. Features 3 and 4 are measures of skewness which is a measure of asymmetry of the distribution while 5 and 6 are called normalized ratio of skewness and kurtosis. Features 7 through 9 are obtained from the ratio of the vertical distribution v(x) to the horizontal distribution h(x).

Recognition

The recognition process is divided into two phases. The first phase is concerned with the recognition of the secondaries where the type of secondaries and their position with respect to the primary character are found. The second phase deals with the recognition of the primary character.

Recognition of Secondaries

From the information obtained during segmentation and elimination of the secondaries, it is possible to evaluate the secondaries. An important factor for evaluation is the line thickness (t). The line thickness is found by collecting statistics of the length of columns (number of dark points). The line thickness is equal to the length of the column appearing most frequently but less than m, where m is the average column length

$$m = \frac{1}{N} \sum_{N} cl \qquad (13)$$

where N is the number of columns

cl is the number of dark points for each column

Figure 4 illustrates the algorithm used to evaluate secondaries.

Recognition of Primary Characters

Classification of the primary characters is accomplished using a quadratic Bayesian classifier which is explained in Al-Yousefi and Udpa (1988).

The generalized distance from an observation vector $\underset{\sim}{x}$ to a group w_i with a mean u_i and covariance matrix is Σ_i is given by

$$D_i^2\,(x) = (\underset{\sim}{x}-u_i)\;T\;\Sigma_i^{-1}\,(\underset{\sim}{x}-u_i) + \ell n\,|\,\Sigma_i\,| \tag{14}$$

and the posterior probability of $\underset{\sim}{x}$ belonging to group w_i is:

$$P(w_i\,|\,\underset{\sim}{x}) = \frac{\exp\,[-0.5\;D_i^2\,(x)]}{\sum_k \exp\,[-0.5\;D_k\,(x)]} \tag{15}$$

In other words $P(w_i\,|\,x)$ is largest for the smallest $D_i^2\,(x)$.

Results and Conclusions

The algorithm was applied to handwritten samples from a data base of 50 samples of Arabic character set written by 25 authors. The correct classification rate was found to be 98.50%. Applying the algorithm to the same characters without the segmentation of secondaries, a classification rate of only 95.8% was achieved. Isolating the secondaries and recognizing them separately enabled the reduction of the classification issue to a twenty class problem and improved the classification process by 2.7%.

The results obtained show that the algorithm used for segmenting the secondaries can be used successfully together with the method of moments for the recognition of handwritten characters.

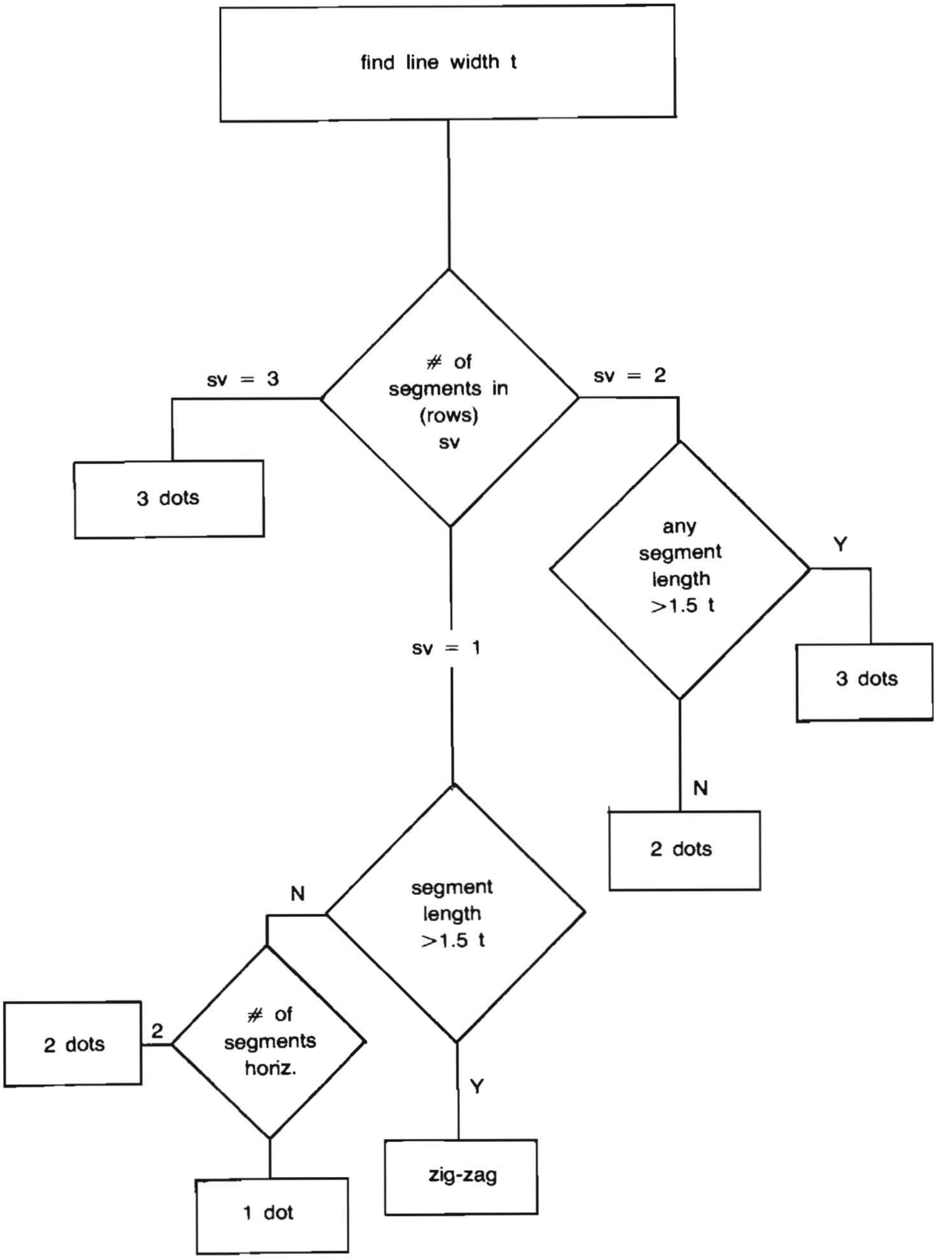

Fig. 4. Recognition algorithm of the secondaries (dot, group of dots or zig zag)

References

Almuallim, H. and **Yamaguchi, S.** (1987) A method of recognition of Arabic cursive handwriting, IEEE Trans. on Pattern Analysis and Machine Intelligence *PAMI* **9:** 715-722.

Al-Yousefi, H. and **Udpa, S.S.** (1988) Recognition of handwritten Arabic characters, SPIE's 32nd Annual Technical Symposium on Optical and Optoelectronic Applied Science & Engineering, San Diego, California.

Amin, A. (1982) Machine recognition of handwirtten Arabic words by IRAC II, 6th International Conf. on Pattern Recongition, Munich, Germany, 34-36.

Amin, A, Kaced, A., Haton, J.P. and **Mohr, R.** (1980) Handwritten Arabic character recognition by IRAC system, 5th International Conf. on Pattern Recognition, Miami, Florida, 729-731.

Badi, K. and **Shimura, M.** (1982) Machine recognition of Arabic cursive scripts, *Trans. Inst. Electron. and Commun. Eng.* **E65:** 107-114 Japan.

Kahan, S., Pavilidis, T. and **Baird, H.S.** (1987) On the recognition of printed characters of any font and size, IEEE Trans. on Pattern Analysis and Machine Intelligence *PAMI* **9:** 274-287.

Suen, C.Y., Berthod, M. and **Mori, S.** (1980) Automatic recognition of handprinted characters - the state of the art, *Proc. of IEEE,* **68:** 469-487.

(Received 31/10/1988;
in revised form 25/04/1989)

تمييز الحروف العربية المكتوبة باليد بواسطة التجزيء

اتش. اس. اليوسفي و اس. اس. يودبا

قسم الهندسة الكهربائية ـ جامعة ولاية كولورادو
فورت كولينز ـ كولورادو ٨٠٥٢٣ ـ الولايات المتحدة الأمريكية

نوجد في هذا البحث وسيلة جديدة لتجزيء الحروف العربية المكتوبة باليد بغية تمييزها والتعرف عليها بشكل أفضل. إذ يتم تجزيء النقاط والحركات بشكل منفصل ويتم التعرف عليها وكذلك بشكل منفصل. وبالتالي فإن هذا يقلص عدد المسائل من ٢٨ مسألة (وهو عدد حروف الأبجدية العربية) إلى ٢٠ مسألة. يتم بعدها تمييز الجزء الرئيسي من الحروف بإستخدام معالم مستقاة من فترات الإسقاطات الشاقولية والأفقية بالنسبة لفترة الأمر ـ صفر. ويتم تصنيف الحروف الرئيسية بإستخدام توابع مميزة تربيعية. ونبين في النهاية كيف أن النتائج الحاصلة تحسن بشكل جيد تصنيف الحروف.