Arab Gulf J. scient. Res. **2** (1), pp. 329-347 (1984)

On Feature Extraction and Selection for Arabic Character Recognition

Adnan Nouh, Abobakr Sultan and Roshdi Tolba

College of Engineering, King Saud University, Riyadh, Saudi Arabia

ABSTRACT. An algorithm to select an optimum set of density features for the Arabic character set is developed and implemented. The method involves generation of all possible features by means of position-by-position matching between characters. The feature selected, each have a correlation coefficient greater than a certain threshold.

The obtained optimum feature set is utilized in the classification stage employing a sequential tree search technique. Computational results are presented and discussed.

1. Introduction

Any character recognition system must include two basic design steps, the feature extraction and selection procedures, and the decision rule formulation or classification stage. In the classification procedure, the properties of a character to be classified are measured, combined and evaluated with a decision rule to render a classification. The selection of properties that contain the most discriminating information is important because the cost of decision making is directly related to the number of properties used in the decision rule, and the number of sample points increases exponentially with this number (Mucciurdi and Gose 1971). The important requirements of these properties are low dimensionality, sufficiency of information, and discrimination ability (Kavel 1974).

The problem of selecting a high discriminating subset of properties or features has been approached in a number of ways. Statistical methods extract features by evaluating a number of features according to some performance measures such as error probability, information function and divergence criteria (Freedman 1974). Geometric features can be generated from the output of a raster sanners, and

density features are based on the density of binary 1's in predetermined regions (Vanderburg and Rosenfeld 1977). For binary patterns, sequential and parallel algorithms were developed to determine a minimum set of features which are common to a group of patterns. However, the problem of threshold selection and the number of features needed was not solved and required further investigation (Tou and Gonzalez 1974).

The unique characteristics of Arabic text and the problems relevant to computer processing of Arabic characters has been reported. A standard character set has been proposed for solving the problems arising from the diverse shape and size of Arabic characters. Consequently, in one system reported (Nouh *et al.* 1980) the features were selected simply by inspection and correlated to unknown characters using a sequential tree search technique.

In this work, the problem of automatic feature extraction and selection for Arabic characters is studied and solved. The method proposed represents a progression of previous work (Nouh *et al.* 1980), and could be generally applied to any set of Arabic characters. The method of feature extraction is described in section 2, along with a theoretical background. In section 3, feature allocation in recognition tree nodes is presented. Results are discussed in section 4, and in section 5 conclusions are presented.

2. Feature Extraction

2.1. *Theoretical Background*

The problem of extracting and selecting discriminating features has played a major role in pattern recognition studies. The most important features are not necessary easily measurable. Extracting and selecting binary features from binary patterns has not been solved in general (Freedman 1974). The theoretical difficulty is that the features must be evaluated in terms of the decision stage rather on their own (Nagy 1968).

Geometric features, such as line segments, line endings, relative position of line segments, ... etc., were also used. The character is represented by its binary feature vector, X,

$$X = (x_1, x_2, \ldots, x_i \ldots x_n) \tag{1}$$

where $x_i = 0, 1, i = 1, 2, \ldots n$ (2)

These were submitted to the computer in the form of a Boolean expression specifying the particular combination of black and white bits. A character U_k is said to be recognizable and of class U_A if $D_A \leqslant R$ and $D_B - D_A \geqslant C_A$, where D_A is the minimum overall distance of X from a particular reference,

$$D_A = \min_k\{D_k\},\ k = 1, 2, \ldots P \tag{3}$$

and D_k is the minimum distance of class U_k from the reference Y_{kj}

$$D_k = \min_j\{[D(X,Y_{kj})]\},\ j = 1, 2, \ldots S_k \tag{4}$$

$$\text{where } D(X,Y_{kj}) = \text{Constant} + \sum_{i=1}^{N} (x_i, Y_{ikj}) \tag{5}$$

and

$$(x_i, Y_{ikj}) = \left\{ \begin{array}{l} 1,\ x_i = 1\ \&\ Y_{ikj} = 0 \\ \quad x_i = 0\ \&\ Y_{ikj} = 1 \\ 0, \text{otherwise} \end{array} \right\} \tag{6}$$

R & T are positive integers with $R \leqslant T$, while CA is a constant associated with the reference yielding D_A. If $D_A \geqslant T$, the character is unrecognizable (Andrews *et al.* 1968).

The n-measured features of each member of a class can be represented by an n-dimensional vector. For all members of a class, a common set of properties in n-dimensional space are found. The task of assigning an unknown object to one class or another consists of finding the maximum correspondence in the n-dimensional vector (Harmon 1972).

The idea of using subtemplates has found extensive use in character recognition, where one could employ templates of strokes or other local features, rather than immediately using templates of entire character. The rest of the template is applied only when the subtemplate's degree of mismatch is less than a threshold. For the binary case, the mismatch measure is given by:

$$\text{Mismatch} = \frac{1}{n}\,[N_U(0) + N_Z(1)] \tag{7}$$

where U denotes the set of template points which are 1, Z denotes those points which are 0, $N_U(0)$ is the number of picture points in U which are 0, and $N_Z(1)$ is the number of points in Z which are 1, and n is the size of the template (Vanderburg and Rosenfeld 1977).

Seven techniques for choosing good subsets of (N) properties from a set of (M) have been presented and compared. All the techniques resulted in lower error rates than did a random selection of properties. The selection of properties that contain the most discriminatory information is important because the cost of decision making is directly related to the number of properties used in the decision (Mucciurdi and Gose 1971).

The statistical properties selection technique possesses disadvantages that limit its utility in pattern classification problems. The information-theoretic approach is based on entropy or divergence.

For P pattern classes, the entropy of i*th* population is given by:

$$H_i = -\sum_{i=1}^{P} P_i(X) \log_2 P_1(X) \tag{8}$$

The divergence measure of discriminating class i from j is given by:

$$D_{ij} = \sum_{\substack{i=1,\\ j=1}}^{P} [P_i(X) - P_j(X)] \log_2 \frac{P_i(X)}{P_j(X)} \tag{9}$$

where $P_i(X)$ and $P_j(X)$ are the probability occurrence property density function of (X) when (X) ε i and j, respectively. This approach requires a considerable amount of computation (Tou and Gonzalez 1974).

The sequential analysis approach has the advantage of selecting the smallest number of properties (subsets of features) that minimizes the average decision-making risk (Fu and Min 1968).

2.2. *Feature Extraction Algorithm*

In this work, it is assumed that the unknown Arabic characters were scanned, isolated, and digitised into binary form ready to be processed by computer for feature extraction and recognition. The algorithm presented in this section generates the common features by matching each character with all other characters in every shift position. Basically, the procedure consists of establishing a variable threshold and extracting the feature being generated during a given iteration whenever the threshold is exceeded. The procedure starts with selecting an assumed range of threshold. The extraction of the first feature proceeds as follows. Let F(0) represent the initial value of the feature F, which may be any one of the characters under processing. Let also Max $F(0) \cap K_1$ represent the maximum similarity between F(0) and the first character K_1, which is defined as the maximum number of matched ones in the intersection F(0) and K_1. This is done by performing position-by-position matching (Fig. 1) of the feature with characters matrices using a logical multiplication technique (Nouh *et al.* 1980). The maximum similarity in this case is given by:

$$MS = \max_{n=1}^{(JK-JF+1)} \max_{m=1}^{(IK-IF+1)} \sum_{j=1}^{JF} \sum_{i=1}^{IF}$$

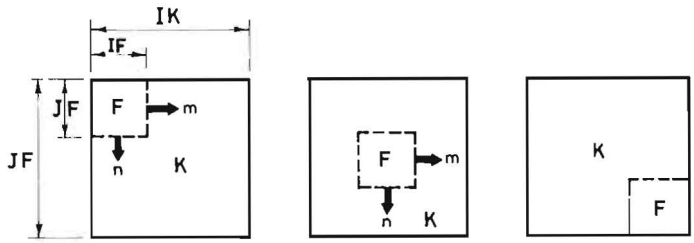

Fig. 1. Position-by-position matching between a feature F and a character K.

$$[F(i,j) \times K(i + m - 1, j + n - 1)] \tag{10}$$

where F(i,j) is the general term in the feature matrix of dimension IF × JF, and K(i,j) is the general term in the character matrix of dimension IK × JK. A correlation coefficient (C), representing the degree of presence of the feature in the character, is calculated as:

$$[C]_{MS} = (MS/FONE) \times 100 \tag{11}$$

where FONE is the number of binary ones in the feature. If the correlation coefficient falls within an assumed initial range of threshold T = 21-25%, then the common pattern of maximum matched ones is the first feature (F_1), otherwise the second character is considered, *i.e.*

$$F_1 = \left\{ \begin{array}{l} F(0) \cap \quad K_1 \ldots \text{if } C \geq T \\ \\ \text{Repeat for } K_2 \ldots \text{otherwise} \end{array} \right\} \tag{12}$$

Similar feature is disregarded using the following criterion. If the number of ones and the dimension of any two features are within an assumed tolerance (± 10% and ± 1 respectively) then the two features are considered similar and one of them is disregarded. To determine the other features, the procedure is repeated with all the characters.

A look-up table of the characters against the features is created, and if it is found that the features failed to be subsets of all the characters, the threshold is incremented and the procedure is repeated. The features fail to describe a character if in the corresponding row in the look-up table all the numbers fall below the range of the threshold.

The output result together with the threshold which fulfills the minimum number of features describing all the characters are selected. This procedure is iterated again considering the previously obtained features as new input patterns.

If the number of features does not decrease or the threshold could not be incremented then the procedure is terminated.

In this way, feature patterns, features-characters look-up table and the correlation coefficient of the maximum similarity $[CC]_{MS}$ are obtained.

The 'Nuqtah' and 'Hamzah' are excluded from this routine since they are known to be common features in some Arabic characters, and it should be added

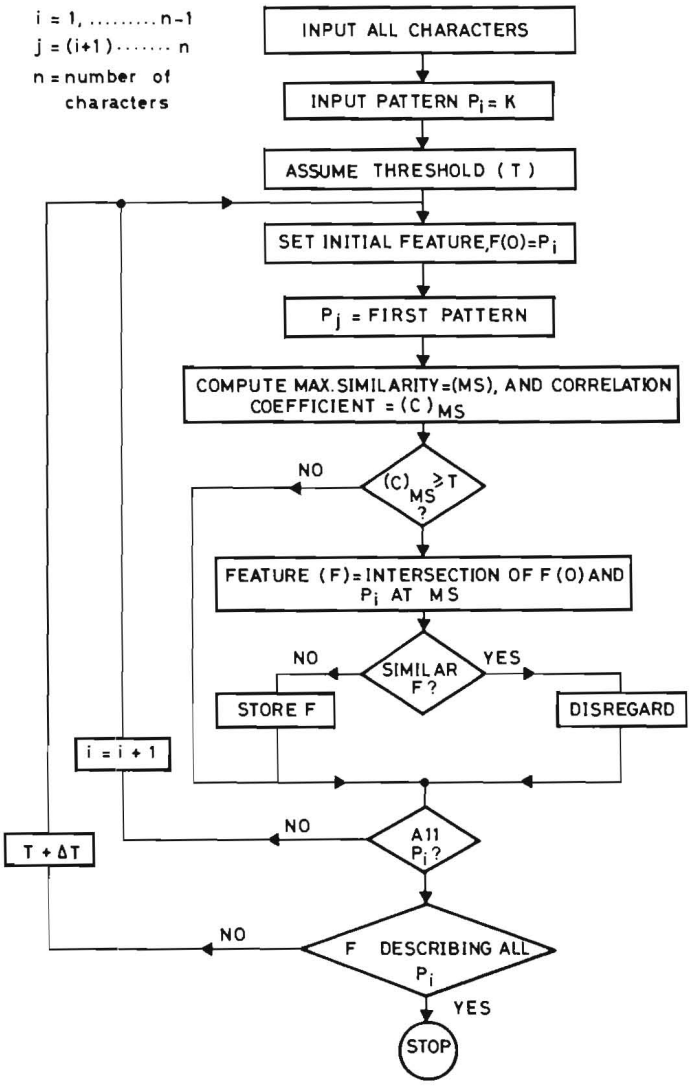

Fig. 2. Simplified flow chart for feature extraction algorithm.

to the above obtained features. Figure 2 illustrates a simplified flow chart for the above algorithm.

3. Feature Allocations in Recognition Tree

In order to determine the power of a feature in distinguishing between character classes, it is necessary to determine the divergence in the measurements obtained when a feature is correlated with the average of the character classes. This can be seen by measuring the ability of a feature to distinguish between characters.

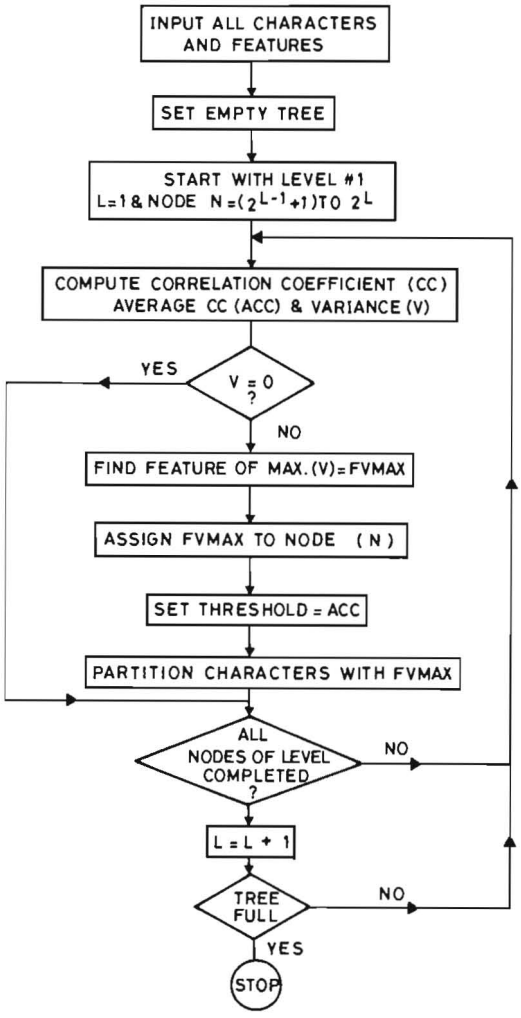

Fig. 3. Simplified flow chart for assigning features to recognition tree nodes.

One criterion of this distinction is the variance of the correlation coefficients, defined as:

$$v = \frac{1}{n} \sum_{i=1}^{n} (C_i - \bar{C})^2 \tag{13}$$

$$\bar{C} = \frac{1}{n} \sum_{i=1}^{n} C_i \tag{14}$$

where C_i is the *ith* correlation coefficient, $\bar{C}$ is the average correlation coefficient and n is the number of characters.

The algorithm presented in this section assigns features to the recognition tree nodes. The feature with the highest variance for all the characters is chosen to be first in the recognition tree. In every iteration step, the variance is calculated for the characters incident to a certain node point in the recognition tree. The feature with the highest variance is assigned to that partitioning node. In Table 5 and Table 6, the feature with highest variance was found to be feature number 7 in both cases.

Two branches result when partitioning the characters with this assigned feature, one containing the characters identifiable and the other not identifiable by the feature. This procedure is repeated to every branch and new partitioning nodes are formed. The average of the correlation coefficients of the incident characters is calculated and used as a threshold. The algorithm stops when the calculated variance is zero. Thus, a tree is formed with partition nodes assigned with certain features in a sequence giving the highest possible distinction. Figure 3 illustrates a simplied flow chart for the above algorithm.

4. Results and Discussion

The two algorithms described in sections 2 and 3 were implemented with a computer program. The 31 isolated Arabic characters were represented by binary patterns in the form of the proposed Arabic character set presented in (Noah *et al.* 1980). These characters were used as input for the feature extraction program. Figure 4 shows the density features of the computer output for the first run. The number of features is 13 for the 31 input isolated characters. It is noticed that the features obtained from the first run are more or less the stem of the isolated main characters. For example, the group (ب ت ث) is represented by one feature (ـبـ) and the group (ج ح خ) is represented by one feature (ـجـ). A feature-character look-up table was constructed and it is given in Table 1. This look-up table gives the correlation coefficient between the features and the characters, and it is used later in the final recognition stage.

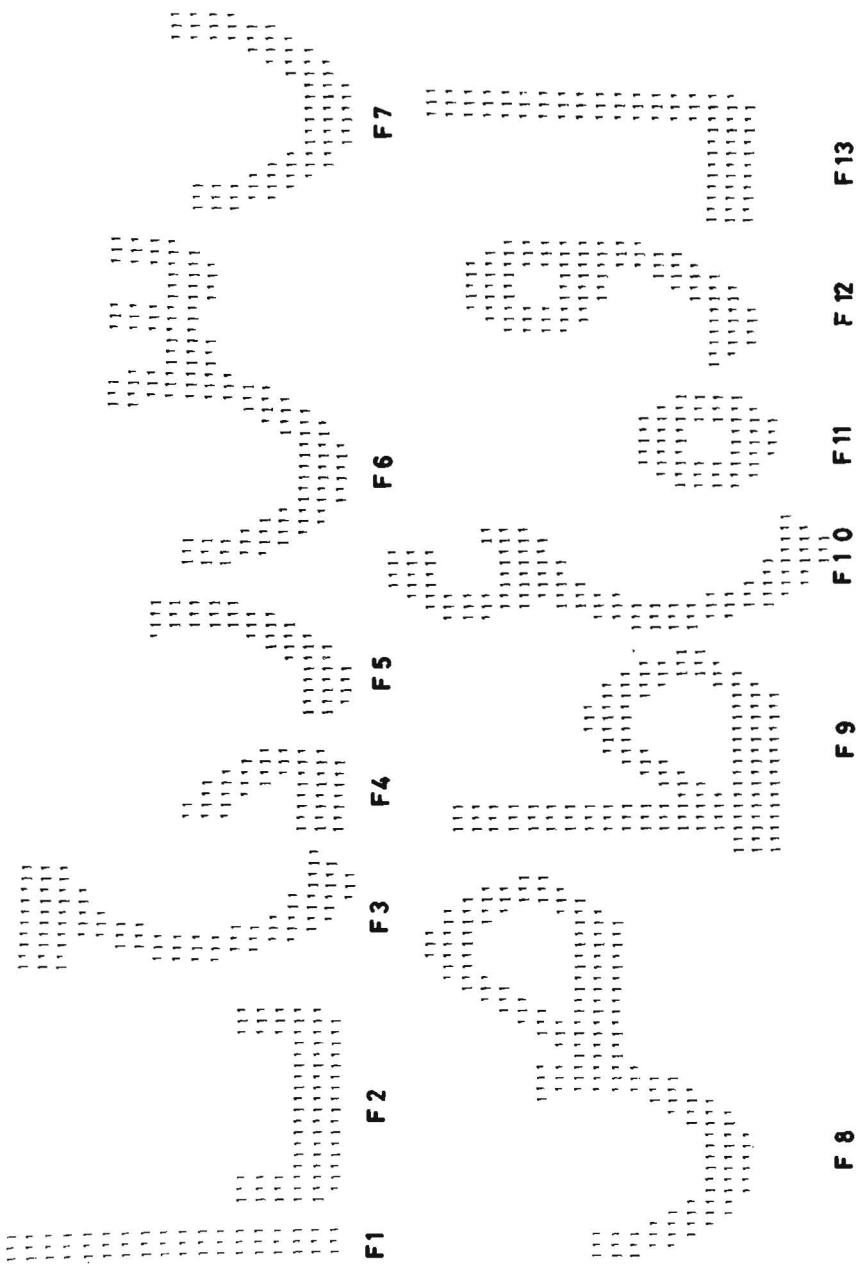

Fig. 4. Features obtained from the first run.

Table 1. Look-up Table for the 13 features case.

LOOK-UP TABLE

NO. OF FEATURE 13

1	ALIF HAMZA	*	98.15	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
2	ALIF	*	98.15	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
3	BEH	*	.00	100.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
4	TEH	*	.00	100.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
5	THEH	*	.00	100.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
6	GEEM	*	70.37	.00	100.00	57.78	34.78	.00	.00	.00	.00	.00	56.60	39.78	25.00
7	HAH	*	70.37	.00	100.00	57.78	34.78	.00	.00	.00	.00	.00	56.60	39.78	25.00
8	KHA	*	70.37	.00	100.00	57.78	34.78	.00	.00	.00	.00	.00	56.60	39.78	25.00
9	DAL	*	.00	.00	.00	100.00	.00	.00	.00	.00	.00	.00	66.04	.00	.00
10	ZAL	*	.00	.00	.00	100.00	.00	.00	.00	.00	.00	.00	66.04	.00	.00
11	REH	*	.00	.00	.00	53.33	100.00	.00	.00	.00	.00	.00	47.17	.00	.00
12	ZEN	*	.00	.00	.00	53.33	100.00	.00	.00	.00	.00	.00	47.17	.00	.00
13	SEEN	*	.00	71.43	.00	66.67	97.83	100.00	100.00	.00	.00	.00	58.49	.00	.00
14	SHEEN	*	.00	71.43	.00	66.67	97.83	100.00	100.00	.00	.00	.00	58.49	.00	.00
15	SAD	*	42.59	85.71	37.93	73.33	97.83	88.37	100.00	100.00	37.86	.00	66.04	52.69	58.75
16	DAD	*	42.59	85.71	37.93	73.33	97.83	88.37	100.00	100.00	37.14	.00	66.04	52.69	57.50
17	TAH	*	100.00	82.86	64.37	84.44	73.91	.00	84.51	.00	100.00	.00	75.47	51.61	58.75
18	ZAH	*	100.00	82.86	64.37	84.44	73.91	.00	84.51	.00	100.00	.00	75.47	51.61	58.75
19	AIEN	*	77.78	.00	73.56	75.56	54.35	.00	.00	.00	.00	100.00	77.36	39.78	35.00
20	GHYN	*	77.78	.00	73.56	75.56	54.35	.00	.00	.00	.00	100.00	77.36	39.78	35.00
21	FEH	*	.00	100.00	.00	95.56	89.13	.00	74.65	.00	.00	.00	100.00	.00	.00
22	KUF	*	.00	62.86	.00	77.78	100.00	.00	100.00	.00	.00	.00	100.00	100.00	.00
23	KAF	*	98.15	.00	27.59	75.56	86.96	.00	.00	.00	.00	.00	64.15	60.22	100.00
24	LAM	*	100.00	62.86	36.78	55.56	97.83	.00	100.00	.00	37.14	33.33	49.06	62.37	80.00
25	MEEM	*	98.15	.00	47.13	71.11	69.57	.00	.00	.00	.00	62.86	100.00	52.69	23.75
26	NOUN	*	.00	62.86	.00	55.56	78.26	.00	100.00	.00	.00	.00	49.06	.00	.00
27	WAW	*	.00	.00	.00	82.22	100.00	.00	.00	.00	.00	.00	100.00	100.00	.00
28	HEH	*	.00	.00	.00	71.11	.00	.00	.00	.00	.00	.00	100.00	.00	.00
29	YEH	*	.00	62.86	.00	75.56	78.26	.00	91.55	.00	.00	.00	69.81	.00	.00
30	TEH MARBOTA	*	.00	.00	.00	73.33	.00	.00	.00	.00	.00	.00	100.00	.00	.00
31	LAM ALIF	*	98.15	.00	42.53	91.11	95.65	.00	.00	.00	.00	.00	75.47	74.19	100.00

Table 2. Look-up Table for the 8 features case.

LOOK-UP TABLE

NO. OF FEATURE 8

1	ALIF HAMZA	*	100.00	100.00	100.00	.00	.00	.00	.00	.00
2	ALIF	*	100.00	100.00	100.00	.00	.00	.00	.00	.00
3	BEH	*	.00	.00	.00	100.00	100.00	.00	.00	.00
4	TEH	*	.00	.00	.00	100.00	100.00	.00	.00	.00
5	THEH	*	.00	.00	.00	100.00	100.00	.00	.00	.00
6	GEEM	*	90.00	100.00	100.00	100.00	80.65	100.00	.00	76.60
7	HAH	*	90.00	100.00	100.00	100.00	80.65	100.00	.00	76.60
8	KHA	*	90.00	100.00	100.00	100.00	80.65	100.00	.00	76.60
9	DAL	*	100.00	85.00	95.00	.00	61.29	.00	.00	.00
10	ZAL	*	100.00	85.00	95.00	.00	61.29	.00	.00	.00
11	REH	*	80.00	100.00	90.00	63.33	51.61	46.15	.00	44.68
12	ZEN	*	80.00	100.00	90.00	63.33	51.61	46.15	.00	44.68
13	SEEN	*	95.00	100.00	100.00	90.00	93.55	100.00	100.00	100.00
14	SHEEN	*	95.00	100.00	100.00	90.00	93.55	100.00	100.00	100.00
15	SAD	*	90.00	100.00	95.00	100.00	93.55	89.74	100.00	91.49
16	DAD	*	90.00	100.00	95.00	100.00	93.55	89.74	100.00	91.49
17	TAH	*	100.00	100.00	100.00	100.00	100.00	89.74	100.00	87.23
18	ZAH	*	100.00	100.00	100.00	100.00	100.00	89.74	100.00	87.23
19	AIEN	*	95.00	95.00	95.00	80.00	100.00	92.31	.00	100.00
20	GHYN	*	95.00	95.00	95.00	80.00	100.00	92.31	.00	100.00
21	FEH	*	100.00	100.00	100.00	100.00	100.00	76.92	71.93	78.72
22	KUF	*	100.00	100.00	100.00	86.67	96.77	58.97	100.00	63.83
23	KAF	*	100.00	100.00	100.00	100.00	70.97	41.03	.00	38.30
24	LAM	*	100.00	100.00	100.00	86.67	87.10	43.59	100.00	42.55
25	MEEM	*	100.00	100.00	100.00	93.33	93.55	84.62	.00	87.23
26	NOON	*	70.00	85.00	75.00	86.67	87.10	43.59	100.00	.00
27	WAW	*	100.00	100.00	100.00	83.33	96.77	64.10	.00	65.96
28	HEH	*	90.00	85.00	100.00	83.33	90.32	.00	.00	.00
29	YEH	*	85.00	85.00	90.00	86.67	87.10	69.23	100.00	70.21
30	TEH MARBOTA	*	90.00	85.00	100.00	73.33	90.32	.00	.00	.00
31	LAM ALIF	*	100.00	100.00	100.00	100.00	90.32	79.49	.00	78.72

Table 3. Look-up Table for the 4 features case.

LOOK-UP TABLE

NO. OF FEATURE 4

1	ALIF HAMZA	*	100.00	100.00	.00	.00
2	ALIF	*	100.00	100.00	.00	.00
3	BEH	*	100.00	100.00	100.00	100.00
4	TEH	*	100.00	100.00	100.00	100.00
5	THEH	*	100.00	100.00	100.00	100.00
6	GEEM	*	100.00	100.00	100.00	100.00
7	HAH	*	100.00	100.00	100.00	100.00
8	KHA	*	100.00	100.00	100.00	100.00
9	DAL	*	100.00	100.00	100.00	73.68
10	ZAL	*	100.00	100.00	100.00	73.68
11	REH	*	92.31	92.31	94.74	63.16
12	ZEN	*	92.31	92.31	94.74	63.16
13	SEEN	*	100.00	100.00	100.00	100.00
14	SHEEN	*	100.00	100.00	100.00	100.00
15	SAD	*	100.00	100.00	100.00	100.00
16	DAD	*	100.00	100.00	100.00	100.00
17	TAH	*	100.00	100.00	100.00	100.00
18	ZAH	*	100.00	100.00	100.00	100.00
19	AIEN	*	100.00	100.00	100.00	100.00
20	GHYN	*	100.00	100.00	100.00	100.00
21	FEH	*	100.00	100.00	100.00	100.00
22	KUF	*	100.00	100.00	100.00	94.74
23	KAF	*	100.00	100.00	100.00	78.95
24	LAM	*	100.00	100.00	100.00	89.47
25	MEEM	*	100.00	100.00	100.00	100.00
26	NOON	*	100.00	100.00	100.00	89.47
27	WAW	*	100.00	100.00	94.74	94.74
28	HEH	*	92.31	100.00	89.47	94.74
29	YEH	*	100.00	100.00	100.00	94.74
30	TEH MARBOTA	*	92.31	100.00	89.47	94.74
31	LAM ALIF	*	100.00	100.00	100.00	94.74

Table 4. Summary of feature extraction and distinction of the three runs.

Result	First run	Second run	Third run
1. Number of features	13	8	4
2. Maximum size of features	[24 × 11] [18 × 35]	[12 × 10] [9 × 17]	[6 × 3] [5 × 9]
3. Range of correlation coefficient	96-100%	34-45%	61-65%
4. Maximum variance of all the characters	2053.7	2197.0	637.7
5. First feature of maximum variance	F.7	F.7	F.4
6. Average correlation Coefficient (ACC)	33.39%	34.57%	87.09%

Table 5. Recognition Table for the 13 features case.

RECOGNITION TREE

NODE	FEAT. NO.	ACC %	VAR.	K'S IN
1	7	33.39	2053.7	31
2	0	.00	.0	0
3	6	34.24	2065.0	11
4	1	42.87	1918.1	20
5	8	50.00	2500.0	4
6	1	42.85	2448.9	7
7	10	26.28	1704.1	10
8	5	30.00	2100.0	10
9	13	58.12	.3	2
10	0	.00	.0	2
11	9	79.04	878.0	3
12	12	25.00	1875.0	4
13	10	87.61	306.5	3
14	3	52.87	1857.2	7
15	12	33.33	2222.2	3
16	2	42.85	2448.9	7
17	0	58.75	.0	1
18	0	57.50	.0	1
21	0	58.75	.0	2
22	0	80.00	.0	1
23	0	.00	.0	1
24	11	72.95	437.4	3
25	0	35.00	.0	2
26	0	23.75	.0	1
27	0	25.00	.0	3
28	13	50.00	2500.0	4
29	0	.00	.0	1
30	0	.00	.0	2
31	0	.00	.0	3
32	11	83.01	288.3	4
47	0	.00	.0	1
48	11	59.43	107.6	2
55	4	83.33	60.4	2
56	0	.00	.0	2
63	4	72.22	1.2	2
64	0	.00	.0	2
95	0	.00	.0	1
96	0	.00	.0	1
109	0	100.00	.0	1
110	0	100.00	.0	1
125	0	.00	.0	1
126	0	.00	.0	1

Table 6. Recognition Table for the 8 features case.

RECOGNITION TREE

NODE	FEAT. NO.	ACC %	VAR.	K'S IN
1	7	34.57	2197.0	31
2	0	.00	.0	0
3	8	73.88	810.8	11
4	6	42.30	1738.6	20
5	7	95.98	96.4	7
6	8	44.14	754.5	4
7	6	80.51	406.5	10
8	4	55.66	2133.4	10
9	8	92.90	28.1	6
10	0	78.72	.0	1
11	2	92.50	56.2	2
12	8	21.27	452.6	2
13	8	86.17	109.7	6
14	5	72.58	444.8	4
15	3	50.00	2500.0	6
16	5	30.64	939.1	4
17	0	100.00	.0	2
18	1	95.00	25.0	4
21	0	63.82	.0	1
22	0	70.21	.0	1
23	0	42.55	.0	1
24	0	.00	.0	1
25	4	84.44	39.5	3
26	0	76.59	.0	3
27	4	91.66	69.4	2
28	0	44.68	.0	2
29	6	13.67	374.0	3
30	0	.00	.0	3
31	0	.00	.0	2
32	0	.00	.0	2
35	0	87.23	.0	2
36	0	91.48	.0	2
49	0	87.23	.0	1
50	0	100.00	.0	2
53	0	78.72	.0	1
54	0	65.95	.0	1
57	0	38.29	.0	1
58	4	78.33	25.0	2
115	0	.00	.0	1
116	0	.00	.0	1

Table 7. Recognition Table for the 4 features case.

RECOGNITION TREE

NODE	FEAT. NO.	ACC %	VAR.	K'S IN
1	4	87.09	637.7	31
2	0	.00	.0	0
3	4	97.80	11.3	24
4	3	69.92	1960.5	7
5	0	100.00	.0	16
6	3	96.71	20.3	8
7	4	70.52	39.8	5
8	0	.00	.0	2
11	4	92.63	6.6	5
12	1	94.87	13.1	3
13	4	75.43	6.1	3
14	0	63.15	.0	2
21	0	94.73	.0	3
22	0	89.47	.0	2
23	0	94.73	.0	1
24	0	94.73	.0	2
25	0	78.94	.0	1
26	0	73.68	.0	2

The output features of the first run were considered as input patterns for the second run. The features found by the second run is shown in Fig. 5. The range of correlation coefficient which gives complete descriptive features was found to be 34 to 45%. Other ranges did not give acceptable results. The number of features decreases from 13 to 8 for the same 31 input characters. The shapes of the features obtained could not be recognised as the stem of the characters no more and their size has decreased. The feature-character look-up table for the case of 8 features is given in Table 2.

The 8 features obtained from the second run were used as input patterns for a third run and produce 4 output features for the same 31 input characters. The four features are shown in Fig. 6. The correlation coefficient range which gives the required common features is found to be 61 to 65%. The feature-character look-up table for this case is given in Table 3. The different results of the three runs are summarized in Table 4. It is clear from this table that the maximum variance of all the characters is almost the same for the first and second runs, while it decreases appreciably for the third run. At the same time, the average correlation coefficient increases for the third run which means that the distribution of the correlation coefficient values is very narrow when compared with the first and second runs. It is expected that the features obtained from the third run would have lower distinction. This is confirmed by the shape of the features as shown in Fig. 6.

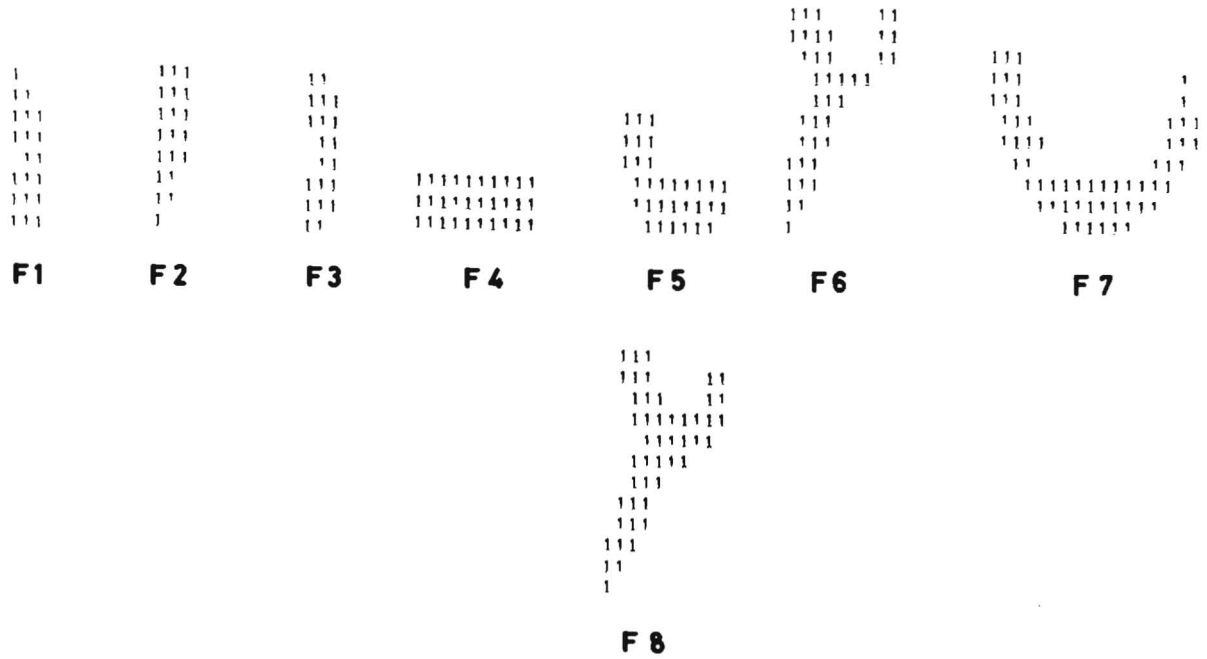

Fig. 5. Features obtained from the second run.

Fig. 6. Features obtained from the third run.

The second algorithm developed in section 3 was also implemented with a computer program for feature allocation in recognition tree nodes. The feature allocation program generates a general recognition tree. The tree is terminated whenever all the characters are distinguished and no further partition is possible. In other words, whenever the variance of the tested characters becomes zero. Tables 5-7 illustrate the computer result of this program using the obtained 13, 8 and 4 features, respectively. The average correlation coefficient is employed as threshold for partitioning the incident characters. A complete search tree was obtained in the cases of 13 and 8 features, whereas it is failed in the case of 4 features and the recognition tree for this case is incomplete. This is explained by the fact that the variance between those features is generally small, and even becomes still smaller as the nodes progress further until it becomes zero and the incident characters cannot be distinguished. For example, Table 7 at node number 3, the number of incident characters is 24 and the maximum variance is 11.3. On the other hand, for node number 13 the variance is zero although the number of characters is 16, which results a tree stop.

Comparing the two schemes of 13 and 8 features, it can be concluded that using the 8 features will result in less computational time in the case of the longest recognition route characters. These characters are (ه هـ لا ى ن ك) and (ة م) and it was found to give rise to CPU times 12 and 4 seconds when using 13 and 8 features, respectively.

Generally, the CPU in the case of 8 features is about 80 seconds on HP 3000 computer which is much less than previous method (Nouh *et al.* 1980).

5. Conclusion

The problem of extracting and selecting an optimum set of density features for the isolated Arabic characters is presented. Two algorithms were implemented for the extraction and selection of Arabic features and the result was used in experimental recognition for the 31 characters. It is found that the optimum number of primary features is 8 for the 31 characters with the addition of 'Nuqtah' and 'Hamzah' as secondly features. The problem of selecting the optimum value of threshold

was solved by using the average correlation coefficient of the incident characters to each node with the feature having maximum distinction. Using these features and their allocation in the recognition tree results in an improvement in computational time when compared with earlier work (Nouh *et al.* 1980).

Acknowledgement

The authors would like to express their appreciation for the research facilities and support offered by the Research Centre, College of Engineering, King Saud University.

References

Andrews, D.R., Atrubin, A.J. and **Hu, K.C.** (1968) The IBM 1975 optical page reader, P. III. *IBM J. Res. Dev.* **12**: 364-371.

Freedman, M.D. (1974) Optical character recognition, *Int. elect. Electron. Engrs Spectrum* **11** (3): 44-52.

Fu, K.S. and **Min, P.J.** (1968) *On Feature Selection in Multiclass Pattern Recognition*, Tech. Rep. TR-EE68-17, Purdue University, School of Elect. Eng., Lafayette, IN, U.S.A.

Harmon, L.D. (1972) Automatic recognition of print and script, *Proc. Inst. elect. Electron. Engrs* **60**: 1165-1176.

Kavel, L. (1974) Patterns in pattern recognition, *Inst. elect. Electron. Engrs Trans. on Infor. Theory, IT-20*, **6**: 697-719.

Mucciurdi, A.N. and **Gose, E.E.** (1971) A comparison of seven techniques for choosing subsets for pattern recognition properties, *Inst. elect. Electron. Engrs Trans. on Computer, C-20*, **9**: 1023-1031.

Nagy, G. (1968) State of the art in pattern recognition, *Proc. Inst. elect. Electron. Engrs*, **56**: 836-862.

Nouh, A., Sultan, A. and **Tolba, R.** (1980) An approach for Arabic character recognition, *J. Eng. Sci., Riyadh Univ.* **6**: 185-191.

Tou, J.T. and **Gonzalez, R.C.** (1974) *Pattern Recognition Principles*, Addison-Wesely Publ. Co., MA, 377 p.

Vanderburg, G.J. and **Rosenfeld, A.** (1977) Two stage template matching, *Inst. elect. Electron. Engrs Trans. on Computer, C-26*, **4**: 384-389.

(Received 04/07/1982;
in revised form 30/11/1983)

عن استخلاص واختيار الملامح للتعرف على حروف اللغة العربية

عدنان نوح، أبو بكر سلطان و ورشدى طلبة

كلية الهندسة - جامعة الملك سعود - الرياض - المملكة العربية السعودية

في هذا البحث، تم إنشاء وتحقيق (ألجوريثم) لاختيار المجموعة المثلى لملامح حروف اللغة العربية.

وتتلخص الطريقة المستعملة في إيجاد جميع الملامح الممكنة بوساطة التواؤم الموضعي بين الحروف. وقد اختير كل من هذه الملامح بحيث يكون له معامل ارتباط أكبر من حد معين.

واستخدمت الملامح المثلى التي تم الحصول عليها في تصنيف الحروف العربية الأساسية بوساطة تقنية البحث المتتابع الثنائي التشعب.

وتم في نهاية البحث تقديم ومناقشة تطبيق ونتائج هذه الطريقة بوساطة الحاسب الآلي.