Printed Arabic Character Classification Using Cadre of Level Feature Extraction Technique

S.Nouri *, M.Fakir **
Department of Mathematics and Computer Science, Beni Mellal, Morocco

***Abstract*—a Feature extraction technique is important in character recognition, because they can enhance the efficacy of recognition in comparison to other approaches. This study aims to investigate the novel feature extraction technique called the Cadre of Level technique in order to represent printed characters and digits. This technique gives statistic and morphologic information, i.e. the calculation is based on a statistical approach but in the positions which can give some information about the morphologic of character. The image of a character is divided into 100 zones, then for each zone we average 5 extracted values (one value for each level) to 1 value for each zone, which gives 100 features for each character. This technique was applied to 105 different characters and 10 different digits of Arabic printed script. K-Nearest Neighbor algorithm was used to classify the printed characters and digits.**

Keywords—Arabic Character; Cadre of Level; Recognition; K-Nearest Neighbor; Digits.

I. INTRODUCTION

Recognition of Arabic characters has been famously researched since of the 1970s [1] and this branch of research based in general on the analytic approach which proposes the systems for recognizing all type of Arabic characters has greatly contributed to the field. In actuality, there is not an ideal system to use for recognition of Arabic script. In our contribution we propose a system to recognize one type of Arabic printed script with new features to show if we can ameliorate the level of recognition. The systems propose three stages, first for processing and preparing the images for a second stage. In this stage the image is scanned and transformed to a binary image, and then the sides of the character are localized for cutting additional parts from the image in order to resize it to 100×100 pixels. Next, two morphological operations are applied, first to find a skeleton of the character, and second, to close the holes in the connected component. In this second stage, in order to extract the features from each character or digit, we use an algorithm, namely Cadre of Level. In the Cadre of Level technique, the matrix of the image is divided into 100 zones with each zone representing a matrix of 10×10. For each zone, a density of pixels is calculated for each level to get 5 values (1 for each level).Then averaging the values to obtain one value for each zone, we obtain finally a vector of 100 features. In the third stage, the Cadre of Level technique is applied to data sets including printed characters (105 classes) and Arabic printed digits (10 classes), then the data sets are classified by the k-Nearest Neighbors algorithm using the Spearman distance as a function for computing distances between data points.

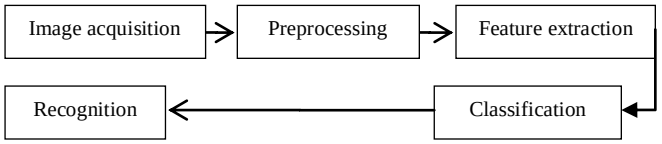
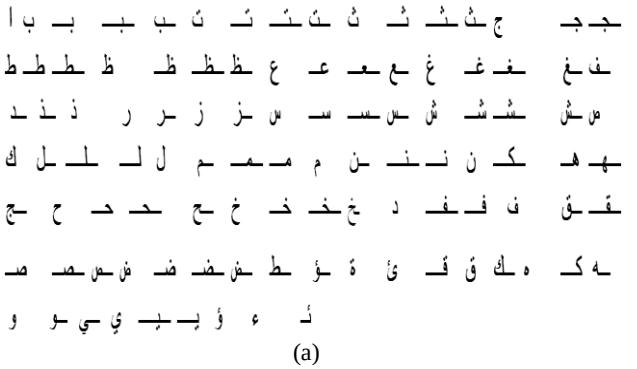


Fig.1. Stages of recognition system

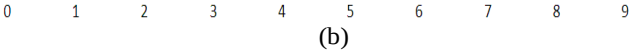
In this study, the classification rates obtained using Cadre of Level algorithms are compared.

II. CHARACTERISTICS AND PROBLEMS OF PRINTED ARABIC CHARACTERS [2]

- The Arabic script is written from right to left and it is cursive,
- Arabic script consists of 28 characters, but in cursive script each character can take 2, 3, or 4 different forms, and some isolated characters can form distinct ligatures (for example ﻻ,ﺋ), so in total there exists more than 105 different characters.
- No concept of upper or lowercase characters exists.
- The same Arabic printed script can be used in more than one font.
- Similarity of some characters like, ب ,ف ,ع ,غ ,ح ,ق ت...
- The noise, background of the image, overlapping of script, and the shape of the characters give us a small quantity of information to distinguish all characters from each other: therefore we experienced some difficulty in the classification stage.



(a)



(b)

Fig.2. (a) Arabic printed characters in different forms
(b) Digits used in Arabic printed script.

III. PREPROCESSING

To extract symbolic information from millions of pixels in document images, each component in the character recognition system is designed to reduce the amount of data .As the first important step, image and data preprocessing serve the purpose of extracting regions of interest, enhancing and cleaning up the images, so that they can be directly and efficiently processed by the feature extraction component. [10]

A. *Thresholding [9]*

Thresholding is a technique that aims to transform a matrix of three dimensions representing an R.G.B. image into a matrix of two dimensions representing a gray image where values are between 0 and 1. Then this matrix is transformed into a binary matrix where values are between 0 or 1, to obtain a binary image to use in the next step of processing.

B. *Noise reduction. [7] [8]*

Various techniques, such as morphological operations, are used to connect unconnected pixels, to remove isolated pixels, and to smooth boundaries. those techniques are also used to minimize the effects of noise on image quality or quantity of information that we use to extract features. In this approach we used thresholding to eliminate noise and unnecessary information and keep most significant information. After extracting the skeleton we used the techniques of mathematical morphology to eliminate the negative effects such as distortion and unconnected pixels, subsequently keeping the general shape of characters and digits.

C. *Normalization*

To process images of the same size we have resized the matrices representing the characters or digits to fixed dimensions, the choice of dimension depends on the nature of the object (character, digit,) and quantity of information that we need to extract features. We set a size of 100 × 100 to represent characters and 80 × 80 to represent digits.

D. *Skeletonization*

To minimize the influence from thick script, the image was normalized and thinned using a thin [3] algorithm to define the skeleton shape for localizing a character and using it to extract Cadre of Level features.

(a) (b) (c)

Fig.3. (a) Image of character before
(b) Thinned image
(c) Cropped image

IV. FEATURES EXTRACTION

Features extraction is a sensitive stage which directly influences the recognition rate. To provide a symbolic of information and to maximize the possibility of distinguishing each character or digit from each other, we have opted for a novel method of extracting 100 features which were obtained from the shape and distribution of pixels in the image. For that we have proposed a technique named Cadre of Level. In the Cadre of Level technique, the image of a skeleton character is used to divide it into 100 zones, each representing a matrix of 10×10 pixels. Each zone is used to statistically calculate values into 1 and 0 as following:

- The matrix is divided into 5 cadres, each cadre representing one level (Figure 3.).

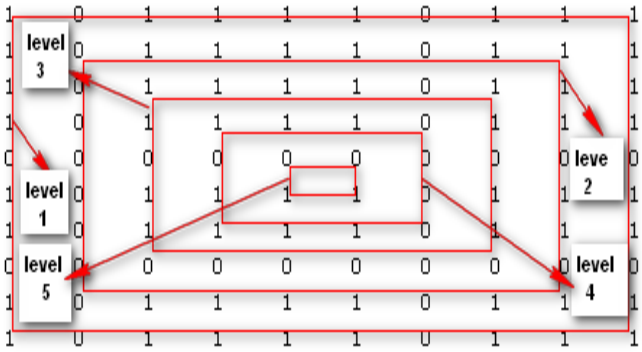


Fig.4. Example of cadre of level of one zone

For each level the following calculation was performed.

- r_1= densities of pixels in the upper row
- r_2= densities of pixels in the lower row
- cl =densities of pixels in the left column
- cr= densities of pixels in the right column

$$r = (r_1 + r_2)/2 \quad (1)$$

$$c = (cl + cr)/2 \quad (2)$$

$$l_i = (r + c)/2 \quad (3)$$

- l_i Represents features of the level number *i*.

The features of the five levels were divided by 5 and the result represents a feature for one zone.

Finally we obtained 100 features to represent one vector of (1×100) for each digit or character**.**

V. CLASSIFICATION

In the classification, to classify the data sets which represent Arabic characters and digits, we have used the k-Nearest Neighbor algorithm with the Spearman distance as a function for computing distances between data points.

A. *K-Nearest Neighbor algorithm*

The k-Nearest Neighbor algorithm (*k*-NN) [4], [5] is a method for classifying objects based on closest training examples in the feature space. Conceder of X and Y

respectively test and training data. X is a vector of the size N features. Y is a matrix of $N_s \times N$ with N_s the number of samples in training data.

The first step of k-Nearest Neighbor algorithm (*k*-NN) involves calculating the distances between X and the N_s samples of the training data Y. In this paper, three distances were studied: City block D1, Spearman D2 and Correlation distance D3. All calculated features were positive. The second step involved sorting the distances obtained. Conceder D the vector contained the K-nearest neighbor distances. The final step is to determine the frequency of each 3 classes contained in the vector D. The minimum distance was adopted to classify the test data. The parameter K, the type of distance used, and features extracted from those three parameters can be influenced as a result of using k-Nearest Neighbor algorithm.

1) Distance Metrics [6]

Given an m-by-n data matrix X, which is treated as mx (1-by-n) row vectors x_1, x_2... x_m, and my-by-n data matrix Y, which is treated as my (1-by-n) row vectors y_1, y_2, ...,y_m, the various distances between the vector x_s and y_t are defined as follows:

a) Spearman distance

$$d_{st} = 1 - \frac{\left(r_s - \bar{r}_s\right)\left(r_t - \bar{r}_t\right)'}{\sqrt{\left(r_s - \bar{r}_s\right)\left(r_s - \bar{r}_s\right)'}\sqrt{\left(r_t - \bar{r}_t\right)\left(r_t - \bar{r}_t\right)'}} \qquad (4)$$

Where

- r_{sj} is the rank of x_{sj} taken over x_{1j}, x_{2j}, ...$x_{mx,j}$,
- r_{tj} is the rank of y_{tj} taken over y_{1j}, y_{2j}, ...$y_{my,j}$,
- r_s and r_t are the coordinate-wise rank vectors of x_s and y_t, i.e., $r_s = (r_{s1}, r_{s2}, \ldots r_{sn})$ and
- $r_t = (r_{t1}, r_{t2}, \ldots r_{tn})$

$$\bar{r}_s = \frac{1}{n}\sum_j r_{sj} = \frac{(n+1)}{2} \qquad (5)$$

$$\bar{r}_t = \frac{1}{n}\sum_j r_{tj} = \frac{(n+1)}{2} \qquad (6)$$

b) City block metric

$$d_{sj} = \sum_{j=1}^{n} \left| x_{sj} - y_{tj} \right| \qquad (7)$$

Notice that the city block distance is a special case of the Minkowski metric, where p=1.

c) Correlation distance

$$d_{st} = 1 - \frac{\left(x_s - \bar{x}_s\right)\left(y_t - \bar{y}_t\right)'}{\sqrt{\left(x_s - \bar{x}_s\right)\left(x_s - \bar{x}_s\right)'}\sqrt{\left(y_t - \bar{y}_t\right)\left(y_t - \bar{y}_t\right)'}} \qquad (8)$$

Where

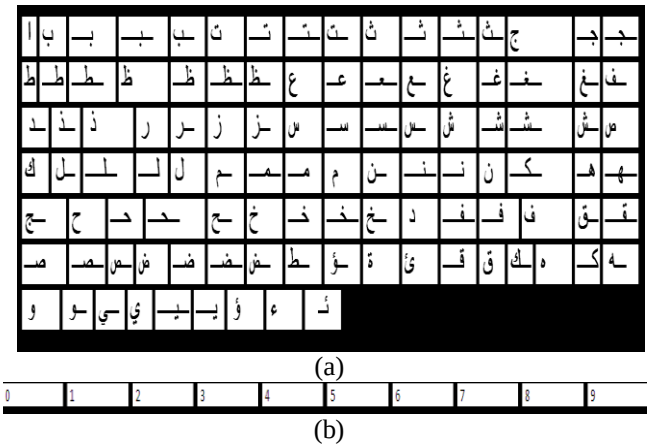
$$\bar{x}_s = \frac{1}{n}\sum_j x_{sj} \qquad (9)$$

$$\bar{y}_t = \frac{1}{n}\sum_j y_{tj} \qquad (10)$$

VI. Experimental Results

In our study we have adopted 'Simplified Arabic Fixed' as a printed font for digits and characters. The Cadre of Level was applied on the binary skeleton images as a method to extract 100 features, then 64 features, and k-Nearest Neighbor algorithm was applied to classify the data of characters and digits. We have analyzed 17,850 different images of characters using 16,590 images for training and 1,260 for testing. And we analyzed 480 images of numerals using 360 images for training and 120 digits for testing.

To compare the results obtained, we have used three different distances, City block, Spearman and Correlation distance.



(a)

(b)

Fig.5. (a) Sample of Arabic printed characters database used
(b) Sample of printed digits database used

The following tables illustrate the results obtained.

TABLE I. Results Of Character Recognition Using Cadre Of Level With The K-Nearest Neighbor Algorithm.

Distance	64 features	100 features
Spearman	96.82%	98.65%
Correlation	87.69%	96.03%
City block	91.58%	96.98 %

TABLE II. RESULTS OF DIGITS RECOGNITION USING CADRE OF LEVEL WITH K-NEAREST NEIGHBOR ALGORITHM.

Distance	100 Features	64 Features
Spearman	95%	99.16%
Correlation	92.5%	88.33%
City Block	90%	90.83%

In the first table (see Table 1), we have obtained 98.65% as the highest rate percentage recognized from printed characters using Spearman distance, then 96.98 % using City block distance, and 96.03% using Correlation distance. These statistics represent analysis on 100 features. But when we use 64 features, we obtain 96.82% as the highest rate percentage using Spearman distance.

In the second Table (see Table 2), we have obtained 99.16% as the highest rate percentage recognized from printed digits using Spearman distance for 64 features.

Preprocessing and distance metrics chosen with k-Nearest Neighbor algorithm *k*-N.N. were influenced directly by the performance of recognition system.

VII. CONCLUSION

In this paper, a new type of feature extraction, namely Cadre of Level extraction, is proposed to elaborate a new offline system for printed Arabic digits and characters. Two approaches using 64 features and 100 features have been chosen to elaborate data for training and testing using k-Nearest Neighbor algorithm. To compare the recognition efficiency of the proposed Cadre of Level method of feature extraction, the k-Nearest Neighbor is used with three different distances, City block, Spearman, and Correlation distance. Experimental results reveal that 100 features gives better recognition accuracy than 64 features for all distances using Cadre of Level feature extraction. From the test results it has been identified that the Spearman distance yields the highest recognition of character accuracy of 98.65% for 100 features and 96.82% for 64 features, and the highest recognition of digits accuracy of 99.16% for 64 features and 95% for 100 features. This study has proved that the field of recognition of Arabic printed characters needs more precision and specificity in all stages to obtain a stronger system of recognition. To improve it is a necessity to continue research with an alternate vision.

References

[1] M. Eden and M. Hall, '' The characterisation of cursives writing '' , proc.4th symp.Informatics Theory, London 1961, pp:287-299

[2] M. Fakir, C. Sodeyama, " Machine recognition of Arabic printed scripts by dynamic programming matching, " Transaction on informatics Systems, vol. 76, No 2, pp. 235-242, 1993.

[3] Lam, L., Seong-Whan Lee, and Ching Y. Suen, "Thinning Methodologies-A Comprehensive Survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 14, No. 9, September 1992, page 879, bottom of first column through top of second column.

[4] Kartar Singh Siddharth , Mahesh Jangid, Renu Dhir, Rajneesh Rani, "Handwritten Gurmukhi Character Recognition Using Statistical and Background Directional Distribution Features", International Journal on Computer Science and Engineering (0975-3397), Vol. 3 No. 6 June 2011.

[5] Puneet Jhajj, D. Sharma, "Recognition of Isolated Handwritten Characters in Gurmukhi Script", International Journal of Computer Applications (0975-8887), Vol. 4, No. 8, 2010.

[6] http://www.mathworks.com/help/stats/pdist.html

[7] Kong, T. Yung and Azriel Rosenfeld, Topological Algorithms for Digital Image Processing, Elsevier Science, Inc., 1996.

[8] Pratt, William K., Digital Image Processing, John Wiley & Sons, Inc., 1991.

[9] Otsu, N., "A Threshold Selection Method from Gray-Level Histograms," IEEE Transactions on Systems, Man, and Cybernetics, Vol. 9, No. 1, 1979, pp. 62-66.

[10] Book : "Character recognition systems a guide for students and practioners"