Application of Data Mining Tools for Recognition of Tifinagh Characters

M. OUJAOURA, R. EL AYACHI
Computer Science Department
Faculty of Science and Technology
Sultan Moulay Slimane University
Béni Mellal, Morocco

O. BENCHAREF, Y. CHIHAB
Computer Science Department
Higher School of Technology
Cadi Ayyad University
Essaouira, Morocco

B. JARMOUNI
Computer Science Department
Faculty of Science
Mohamed V University
Rabat, Morocco.

***Abstract*—The majority of Tifinagh OCR presented in the literature does not exceed the scope of simulation software such as Matlab. In this work, the objective is to compare the classification data mining tool for Tifinagh character recognition. This comparison is performed in a working environment using an Oracle database and Oracle Data Mining tools (ODM) to determine the algorithms that gives the best Recognition rates (rate / time).**

***Keywords*—OCR; Data Mining; Classification; Recognition; Tifinagh; geodesic descriptors; Zernike Moments; CART; AdaBoost; KNN; SVM; RNA; ANFIS**

I. INTRODUCTION

The Optical Character Recognition (OCR) is a rapidly expanding field in several areas where the text is the working basis. In general, a character recognition system consists of several phases [1, 2, 3, 4, 8]. The extraction is a phase that focuses on the release of attributes from an image. In this article, the geodesic descriptors and Zernike moments are two approaches used to calculate the parameters. The effectiveness of the system is based on the results given by the classification phase using data mining tools, which is the purpose of this document.

The rest of the paper is organized as follows. The Section 2 discusses the first primordial task of any recognition system. It's the features extraction problems in addition to a brief formulation for geodesic descriptors and Zernike moments as features extraction methods. The Section 3 is reserved for the second important task which is the classification problems using some classifiers based on several algorithms like ANFIS, ANN, SVM, CART, KNN and AdaBoost. Finally, the Section 4 presents the experimental results for the recognition system.

II. FEATURES EXTRACTION

Tifinagh is the set of alphabets used by the Amazigh population. The Royal Institute of Amazigh Culture (IRCAM) has normalized the Tifinagh alphabet of thirty-three characters as shown in Fig. 1.

Fig. 1. Tifinagh characters – IRCAM.

The Tifinagh alphabet has several characters that can be obtained from others by a simple rotation, which makes invariant descriptors commonly used less effective. For this reason, we used a combination of Geodesic descriptors [5] and Zernike moments [6].

A. Geodesic descriptors

A geodesic descriptor is the shortest path between two points along the spatial deformation of the surface. In the case of binary images; we used a Shumfer simplification which comprises those operations:

- Calculate the number of pixels traveled between the two points;
- Penalize the transition between horizontal and vertical pixels by 1 and moving diagonally by 1.5;
- Choose the optimal path.

We consider the preliminary processing that consists of two standard processes: (i) the noise elimination and (ii) the extremities detection (Fig. 2).

1) Extremities detection

In order to identify the extremities, we use an algorithm that runs through the character contour and detects the nearest points to the corners of the image.

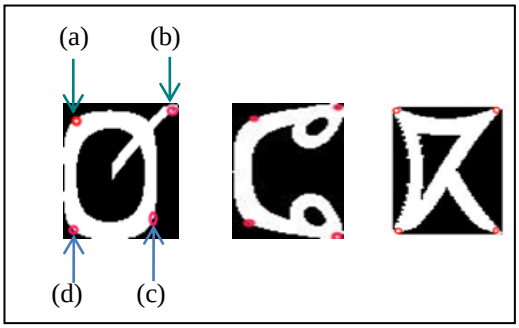

Fig. 2. Example of the extremities of three Tifinagh characters.

2) Geodesic descriptors calculation

We called "geodesic descriptors" the distances between the four detected extremities of the character divided by their Euclidean distances. We set:

- $Dl_M(xy)$: Geodesic distance between x and y;
- dxy: Euclidean distance between x and y;
- a, b, c and d: Detected extremities of each character.

And, we call:

- 1st metric descriptor: $D_1 = Dl_M(ab)/d_{ab}$
- 2nd metric descriptor: $D_2 = Dl_M(ac)/d_{ac}$
- 3rd metric descriptor: $D_3 = Dl_M(ad)/d_{ad}$
- 4th metric descriptor: $D_4 = Dl_M(bc)/d_{bc}$
- 5th metric descriptor: $D_5 = Dl_M(bd)/d_{bd}$
- 6th metric descriptor: $D_6 = Dl_M(cd)/d_{cd}$

To ensure resistance to scale change of the proposed descriptors, we divided each geodesic path by the corresponding Euclidean distances.

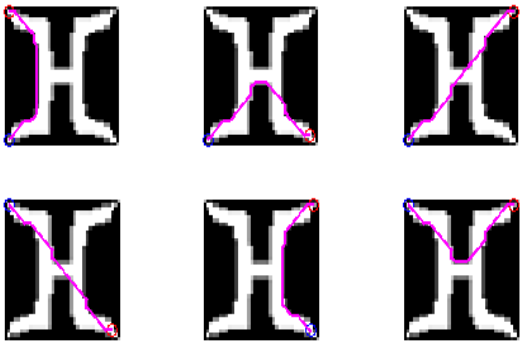

Fig. 3. Example of Geodesic descriptors calculation.

B. Zernike moments

The Zernike moments are a series of calculations that converts an image into vectors with real components representing moments A_{ij}.

By definition, the geometrical moments of a function f(x, y) is the projection of this function on the space of polynomials generated by $x^p y^q$ where $(p, q) \in N^2$. Zernike introduced a set of complex polynomials which form an orthonormal basis inside the unit circle as $x^2 + y^2 \leq 1$. The form of such polynomials is [7]:

$$V_{nm}^*(x,y) = V_{nm}(\rho,\theta) = R_{nm}(\rho).\exp(jm\theta) \quad (1)$$

Where:

n: a positive or null integer;

m: an integer such that $|m| \leq n$;

r: length of the vector from the origin to the pixel (x, y);

θ: angle between the vector x and p;

R_{nm}: radial polynomial.

V * (x, y): complex polynomial projection of f (x, y) on the space of complex polynomials.

Such polynomials are orthogonal since:

$$\int_{x^2+y^2<=1} \int [V_{nm}^*(x,y)] V_{pq}(x,y)dxdy = 0 \text{ or } 1 \quad (2)$$

The geometrical Zernike moments are the projection of the function f (x, y) describing an image on a space of orthogonal polynomials generated by:

$$A_{nm} = \frac{n+1}{\pi} \sum_x \sum_y f(x,y).V_{nm}^*(\rho,\theta)dxdy \quad (3)$$

For identification of the image, the Zernike moments modules are used:

$$|A_{nm}| = \sqrt{\mathrm{Re}^2(A_{nm}) + \mathrm{Im}^2(A_{nm})} \quad (4)$$

III. CLASSIFICATION

The choice of the classifier is primordial. It is the decision element in a pattern recognition system. In this context, we compared the performance of six Data Mining algorithms.

A. CART Algorithm

CART (Classification And Regression Tree) builds a strictly binary decision tree with exactly two branches for each decision node. The algorithm partitions or divides recursively the training set using the principle of "divide and conquer" [9].

B. KNN (k-nearest neighbor)

The k-nearest neighbors algorithm (kNN) [10] is a learning method based instances. To estimate the associated output with a new input x, the method of k nearest neighbors is taken into account (with the same way) the k training samples whose entrance is nearest to the new input x, according to distance measurement to be defined.

C. SVM (Support Vector Machines)

Support Vector Machines (SVM) [11] are a class of learning algorithms that can be applied to any problem that

involves a phenomenon f that produces output y=f(x) from a set of input x and wherein the goal is to find f from the observation of a number of couples input/output. The problem is to find a boundary decision that separates the space into two regions, to find the hyper-plane that classifies the data correctly. This hyper-plane must be as far as possible of all the examples in order to maximize the margin which is the distance from the nearest point of the hyper-plane.

D. ADaBoost Algorithm

AdaBoost [12] is a meta-algorithm to boost the performance of a classifier based on neural network. The algorithm combines weak assumptions made by a classifier trained on a small training subset which the distribution of elements is reinforced, iteration after iteration, so that learning can be focused on examples that pose the greatest difficulties to the trained classifier.

E. ANFIS

The ANFIS (Adaptive Network Fuzzy Inference System based) [13] where the adaptive network fuzzy inference system uses a hybrid learning algorithm to identify the parameters of the association function of the single output type systems Fuzzy Inference of Sugeno (FIS) [13]. A combination of the square and back-propagation gradient descent methods are used for the parameters of the training of the FIS and functions to model a given set of input / output data. The program ANFIS is available Matlab fuzzy toolbox.

F. ANN (Artificial Neural Networks)

Multi-layers Artificial Neural Networks (ANN) are typically built according to a normalized model that includes 3 or 4 layers in total (i.e. 1 or 2 hidden layers).

The method of changing weight is easy with the algorithm of Rosenblatt, but it involves some learning limitations [14]. In the case of multilayer perceptrons, since the desired output of hidden layers are not known, only those of the last layer are known, it is necessary to propagate the errors responsibility of the last layer to the first layer in the opposite direction of network execution, hence the name back-propagation. Multilayer neural perceptrons use the sigmoid activation function; it allows the necessary nuances for proper use of back-propagation.

IV. RESULTS & DISCUSSIONS

In this paper, the recognition approach of Tifinagh character is tested on the database [15], it is composed of 2175 Tifinagh printed characters with different sizes and styles.

The feature vector of each character is the combination of geodesic descriptors and the seven first Zernike moments. Also, the recognition rates and the execution time have been calculated for the used algorithms: ANFIS, ANN, SVM, CART, KNN and AdaBoost.

The following table (Table 1) shows the obtained results for a test of 2000 isolated characters.

TABLE I.
OBTAINED RESULTS: RCOGNITION RATES AND EXECUTION TIME
(PC WITH 2GHZ PROCESSOR AND 4GO OF RAM)

	Zernike	Geodesic	Zernike + Geodesic	Time (s)
CART	41	62	68	4.9
KNN	40	73	81	6.05
ANFIS	63	68	77	23.7
SVM	78	76	93	15
ANN	79	73	94	21
AdaBoost	73	67	93	16

An overview in Table 1 shows that the combination of the two methods (geodesic descriptors and Zernike Moments) used together to calculate the character image parameters in the extraction phase increases the recognition rate. One might also note that the ANN present the best recognition rate, but with a large computation time. The SVM and AdaBoost algorithms show similar results, but with a less hard learning time. The CART algorithm has the best computation time, but with a lower recognition rate.

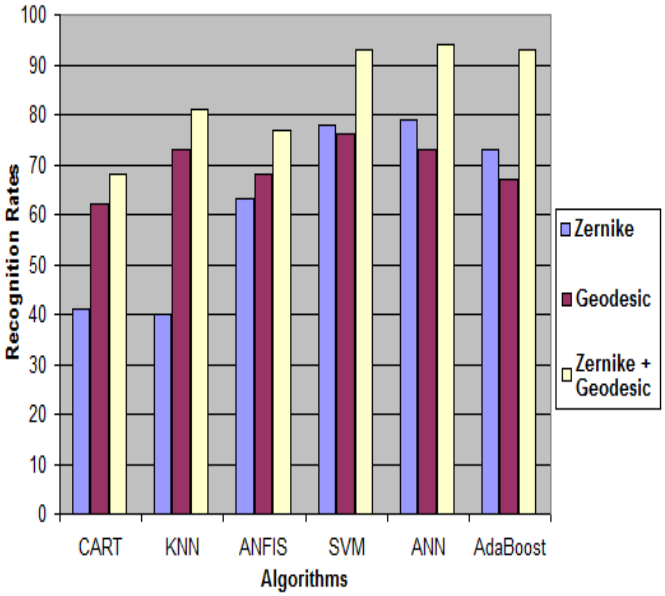

Fig. 4. Comparison of recognition rates for the used algorithms: ANFIS, ANN, SVM, CART, KNN and AdaBoost.

The Figure 4 and 5 gives respectively a comparison of recognition rates and execution time for the used algorithms that are ANFIS, ANN, SVM, CART, KNN and AdaBoost.

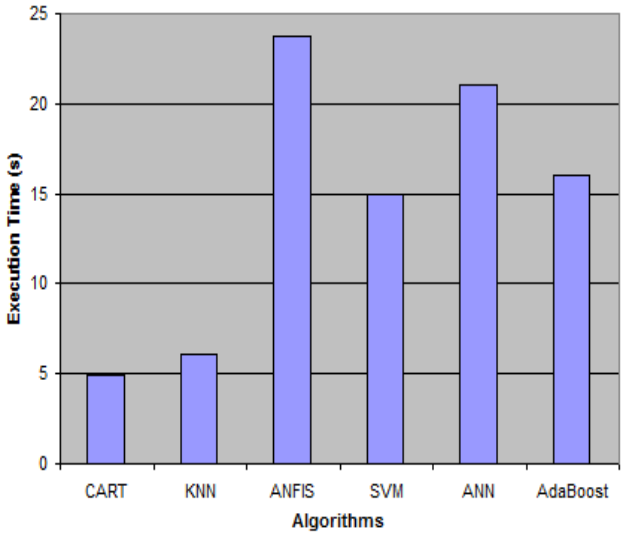

Fig. 5. Comparison of the execution time for the used algorithms: ANFIS, ANN, SVM, CART, KNN and AdaBoost.

All the tests are performed using a database containing a set of 2000 images of isolated characters [15]. The proposed system has been implemented and tested on a core 2 Duo personnel computer with 2 Ghz processor and 4Go of RAM using Matlab software.

V. CONCLUSION

This work presented a Tifinagh character recognition system that uses Data mining tools (ANFIS, ANN, SVM, CART, KNN and AdaBoost) in the classification phase. Also the geodesic descriptors and Zernike moments are adopted to extract the attributes in the features extraction phases. It can be concluded that each one of the tested algorithms has advantages and disadvantages regarding to the recognition rate or execution time. In the future work, the performance of the combination of these classification Data mining algorithms will be studied.

REFERENCES

[1] Y. Es Saady, A. Rachidi, M. El Yassa, D. Mammass. (2010). Printed Amazigh Character Recognition by a Syntactic Approach using Finite Automata, ICGST-GVIP Journal, Volume 10, Issue 2, June 2010.

[2] O. Bencharef, M. Fakir, N. Idrissi, B. Bouikhalen et B. Minaoui. (2011). Application de la géométrie riemannienne à la reconnaissance des caractères Tifinaghe, Agadir-Maroc, 06-07 Mai 2011. pp : 179 :188.

[3] R. El Ayachi, M. Fakir et B. Bouikhalen. (2012). Transformation de Fourier et moments invariants appliqués à la reconnaissance des caractères Tifinaghe, Revue eTI, Numéro 6. 2012.

[4] .El Ayachi, K. Moro, M. Fakir and B. Bouikhalene. (2010). On The Recognition Of Tifinaghe Scripts, JATIT, vol. 20, No. 2, pp: 61-66, 2010.

[5] O. Bencharef, M. Fakir, B. Minaoui et B. Bouikhalene. (2011). Tifinagh Character Recognition Using Geodesic Distances, Decision Trees & Neural Networks. (IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence, 1-5.

[6] A. Prata, W.V.T. Rusche, Algorithm for computation of Zernike polynomials expansion coefficients, Appl. Opt. 28, pp. 749–754, 1989.

[7] Leo Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. Classification and Regression Trees Wadsworth, 1984.

[8] Y. Es Saady, A. Rachidi, M. El Yassa et D. Mammass. (2011). Amazigh Handwritten Character Recognition based on Horizontal and Vertical Centerline of Character. International Journal of Advanced Science and Technology Vol. 33, Août, 33-50.

[9] C.W. Chong, P. Raveendran, R. Mukundan, A comparative analysis of algorithms for fast computation of Zernike moments, Pattern Recognition 36 (3) , pp. 731–742, 2003.

[10] Oren Boiman, Eli Shechtman and Michal Irani, In Defense of Nearest-Neighbor Based Image Classification, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2008.

[11] K.-B. Duan, S.S Keerthi, Which is the best multiclass SVM method? An empirical study. Technical Report CD-03-12, Control Division, Department of Mechanical Engineering, National University of Singapore, 2003.

[12] Paul Viola and Michael Jones. Robust real-time object detection. In International Journal of Computer Vision, 2001.

[13] J.-S.R. Jang, ANFIS: Adaptive-Network-Based Fuzzy Inference Systems, IEEE Trans. Systems, Man & Cybernetics 23 (1993), pp: 665—685.

[14] O. Lezoray. Segmentation d'images par morphologie mathématique et classification de données par réseaux de neurones : Application à la classification de cellules en cytologie des séreuses. Thèse de doctorat, Université de Caen (2000).

[15] Y. Ait Ouguengay, M. Taalabi, "Elaboration d'un réseau de neurones artificiels pour la reconnaissance optique de la graphie amazighe: Phase d'apprentissage", Systèmes intelligents-Théories et applications, Paris: Europia, cop. 2009 (impr. au Maroc), ISBN-102909-285553, 2009.