

Classifiers Combination for Arabic Words Recognition: Application to Handwritten Algerian City Names

Soulef Nemouchi¹, Labiba Souici Meslati², and Nadir Farah³

¹ EPSECG, Ecole Préparatoire des Sciences Economiques,
Commerciales et Sciences de Gestion, Annaba, Algeria

² LRI Laboratory, Badji Mokhtar - Annaba University, Algeria

³ LabGED Laboratory, Badji Mokhtar - Annaba University, Algeria
soulef_inf@yahoo.fr

Abstract. In this paper, we present a global recognition system for Arabic handwritten words; we focus on the two phases of feature extraction and classification. In our system, we have retained three feature sets. The Zernike moments and the structural features of the word are extracted from the binary image, the Freeman code is established from the contour image of the word and the zoning is given from the skeleton image. These features, representing the words, are extracted to be used as input, in an individual or combined way, of the four classifiers used in our system: the Fuzzy C-Means algorithm (FCM), the K-Means algorithm, the K Nearest Neighbor algorithm (KNN) and a Probabilistic Neural Network (PNN). The system architecture is a parallel one where each expert (classifier) gives his point of view and we combine the results to make a final decision. The classifier results are combined using two methods: the simple vote and the weighted sum.

Keywords: Arabic handwriting recognition, Fuzzy C-Means (FCM), K Nearest Neighbor algorithm (KNN), K-Means algorithm, Probabilistic Neural Network (PNN), zernike moments, zoning, Freeman chain code.

1 Introduction

Communicate by writing has always been a first concern for humans who want to create an easy and direct interaction with a computer. This gives labor to the researchers in the “writing recognition” field, especially for handwriting recognition, which is the dream of all those who need to enter data in a computer.

The first research in this field was done more than thirty years ago. Nowadays, there are several applications in which the recognition of handwritten writing is required like the automatic mail sorting, the automatic processing of administrative documents, the investigation forms or the automatic reading of postal addresses and bank checks [1]. Contrary to Latin, the recognition of the handwritten or printed Arabic writing is, till now, in the research and experimentation level.

Unconstrained off-line handwriting recognition remains a challenging problem. Word recognition algorithms suffer from two major problems. One is the

segmentation error given by the word segmentation process, especially for cursive handwriting documents. The other is that the accuracy of recognition drops when the size of the lexicon increases [2].

On the other hand, given the number and variety of methods used in pattern recognition, there is no single method that can be called the *best*. Each approach has strengths and weaknesses, good ideas and bad. One way to take advantage of this variety is to build multiple sources of information based systems. This direction is given more attention in pattern recognition and more work is being done, especially for handwriting recognition applications. The reported results show the efficiency of such techniques including hybrid approaches and multiple classifier schemes, especially for Arabic recognition [3, 4].

The hybrid approaches are represented by those recognition systems that use different sources of information either at the feature extraction level, by using several types of primitives to better describe the input word/character, or at the classification stage by integrating two (or more) complementary classification paradigms, or at both feature extraction and classification levels. The multiple classifier approach is defined as a system consisting of a set of classifiers and a decision combination function. It applies a number of generally independent classifiers and combines their results to generate a single decision.

In our previous works, we dealt with the recognition of handwritten Arabic words in literal amounts [5] using single classifiers (structural, neural, statistical...) having as input several kinds of features. We focused later on multiple classifiers and hybrid systems; we were particularly attracted by the integration of neural and symbolic approaches. We built neuro-symbolic hybrid and multiple classifiers for Arabic literal amounts [6, 7, 8].

In this article, we focus on classifier combination, we propose a system (see figure 1) to recognize handwritten words in Algerian city names lexicon. This system combines several types of classifiers and features to enhance the recognition accuracy.

This article is structured into seven sections. We first describe the overall architecture of the proposed system in the Arabic handwritten words recognition. In the third section, we present the preprocessing operations performed on word images. The fourth section is devoted to the description of our choices for the feature extraction phase while section six is dedicated to the recognition phase. In the last sections, we present the results followed by some conclusions and perspectives.

2 Overview of the Proposed System

An operational pattern recognition system contains a set of processing modules where feature extraction and classification stages are the most important for its overall performance. The feature extraction methods are generally specific to each particular pattern recognition problem, whereas the same classification algorithms can be used in various applications.

There are three ways to approach word recognition problems [9]. The system recognizes the word as an entire and indivisible entity, it is the *global* or *holistic* approach, or it recognizes the word starting from its previously segmented characters, it is the *analytical* approach. The third way consists in the one used by *human reading based systems* which use only some properties and refine gradually, in loops, the word

description [10]. In our work, we have focused on the holistic approach because the considered vocabulary is limited to the Algerian city-names.

Many works show that the combination of classifiers (sequential, parallel or hybrid) improves significantly the performances of the recognition system compared with each classifier separately [7, 11, 12]. In our work, we focus on the parallel combination because it is the most used one in the pattern recognition problems and it proved its effectiveness in many classification tasks [8, 12, 13]. This success is due to its implementation simplicity and its capacity to explore the answers of different classifiers to be combined, by taking into account (or not) the behavior of each classifiers. In this parallel combination, each classifier has to recognize the entire word using its global features. The difference between these classifiers is their word processing manner according to their functional principles. The figure 1 gives the general scheme of the proposed system components which will be described in the following sections. This classifier fusion is made either by a democratic way where there is no superiority of any classifier compared to another, or in a directed manner where the answer of each classifier is weighted according to its performances.

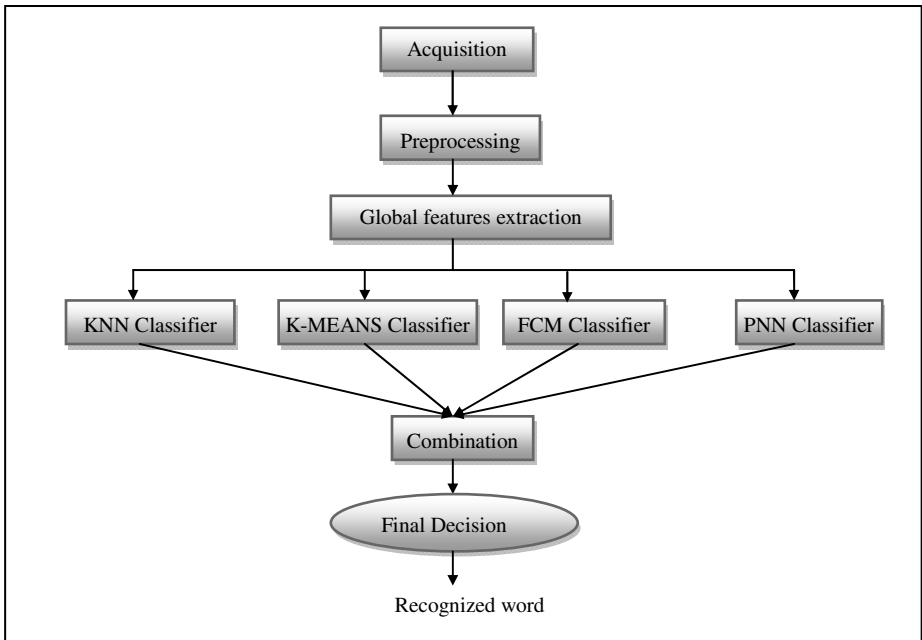


Fig. 1. Overview of the proposed handwritten word recognition system

3 Acquisition and Preprocessing

In our work, we use the Algerian city-names images database, built in the LRI laboratory at Annaba University [14]. The images acquisition was done a scanner and, before being analyzed, the images are submitted to some preprocessing operations.

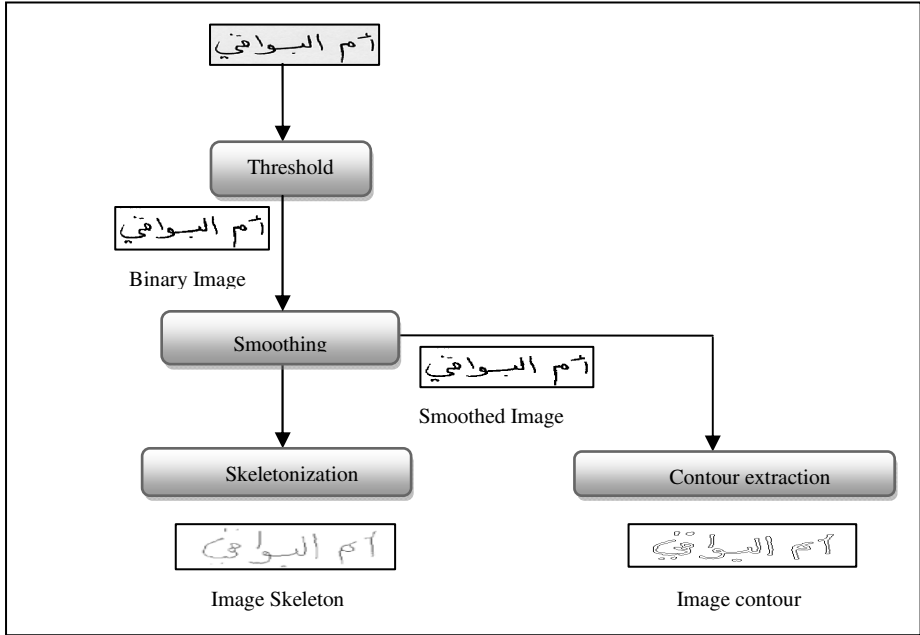


Fig. 2. Image preprocessing

Preprocessing poses major problems: some loops can be opened or not detected and some diacritic dots can be eliminated or confused with noise.

4 Extraction of Global Features and Word Description

The feature extraction problem consists in extracting, from data, the information which is most relevant for classification purposes, in the sense of minimizing the intra-class pattern variability while enhancing the inter class pattern variability [2, 4].

Handwriting recognition systems typically involve two steps: feature extraction in which the patterns are represented by a set of features and classification in which decision rules for separating pattern classes are defined.

The feature extraction phase must ensure a maximum of reliability, because the later phases will not handle the original image but use the results provided by this module.

In the literature, several works concern the elaboration of new features which have an increasing discriminative capacity while minimizing intra-class variability. These features are generally classified in two families: structural features (like strokes, concavities, end points, intersections of line segments, loops, stroke relations . . .) and the statistical features which derive from spatial measurements of the pixels (zoning, invariants moments, Fourier descriptors, Freeman chain code... etc).

We can represent image in different forms: gray-scale, binary, contour, skeleton... and different features can be extracted from each form. Our goal, in this work is to find

the best image representation for the considered Arabic words. Three sets of features are retained (see figure 3), our choice was based on the work done by Arrivault [10].

In our system, we retained the global features from the three image representation forms, for each one we choose the appropriate feature extraction methods. The Zernike moments are extracted from binary image, the Freeman chain code is extracted from the image contour, and zoning is done on the image skeleton. We have also retrained 9 global structural features:

- The number of ascenders in each connected component.
- The number of descenders in each connected component.
- The number of loops in each connected component.

And the diacritic dots which are:

- The number of a high single dot in each connected component.
- The number of two high dependent dots in each connected component.
- The number of three high dependent dots in each connected component.
- The number of a low single dot in each connected component.
- The number of two low dependent dots in each related component.
- And a statistical characteristic which represents the percentage of each component connected in the word (the number of the connected components).

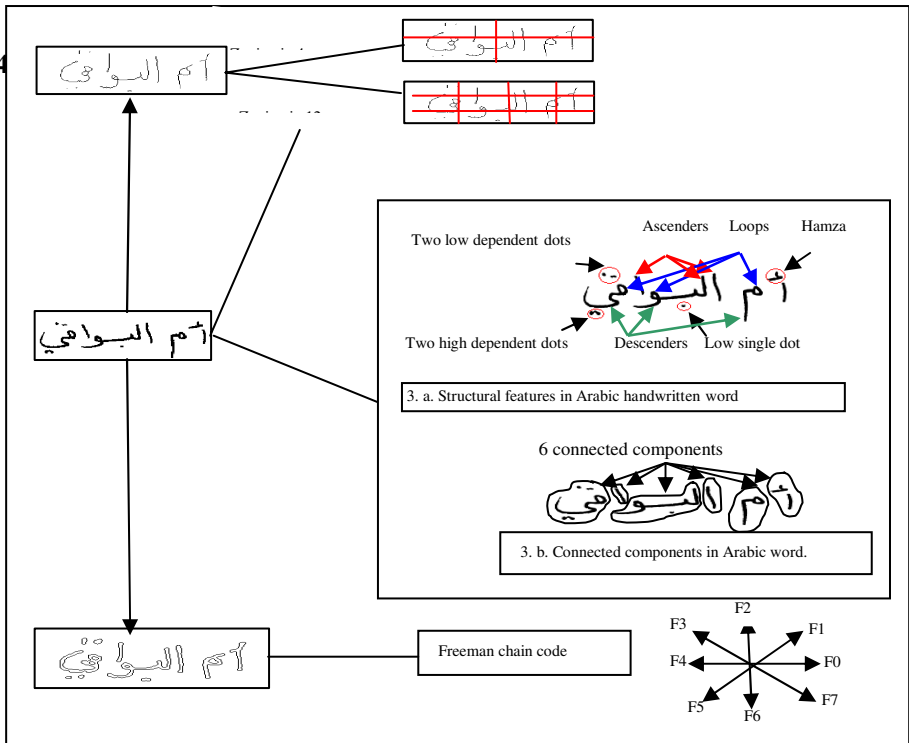


Fig. 3. Chosen features for Arabic handwritten word description

In our work, the extraction of these structural features consists in processing on connected component detected in the preprocessing phase, on the labeled image, from right to left according to the orientation of Arabic writing. The figure 3 illustrates these features. At the end of the feature extraction phase, a word is described by a features vector containing 133 elements:

- 9 structural features.
- 100 Zernike moments.
- 8 Freeman chain code.
- 16 zoning features.

5 Recognition

There are broadly two main approaches for classification: the statistical approach consisting in representing a pattern as an ordered, fixed-length list of numerical values and the structural approach describing the pattern as an unordered, variable-length list of simple shapes. The statistical approach relies on firmly established elements of statistical decision theory, even though viewing a pattern in an n -dimensional space is rather difficult. On the contrary, the structural approach is intuitively appealing because it appears closer to human recognition strategy. Unfortunately, this approach is usually difficult to implement in a fast, trainable and robust way over a large variety of shapes.

As a structural classifier is naturally well-suited to the use of structural features but cannot easily handle statistical features, we have chosen the frame of statistical classification to investigate the combination of both structural and statistical features in a one-shot classifier. In the other way we have chosen the neural networks because several neural network models have been proposed for various difficult problems, especially classification ones. Traditional classifiers test the competing hypothesis sequentially, whereas neural network classifiers test the competing hypothesis in parallel, thus providing high computational rates [15].

As each classification method has its advantages and shortcomings, we can deduce that the performance of system can be increased significantly by combining multiple classifiers. Thus, as shown in figure 1, we propose a system which combines, with two parallel combination methods, four classifiers:

- A K Nearest Neighbor (KNN) classifier.
- A K-Means classifier.
- A Fuzzy C-Means (FCM) classifier.
- A Probabilistic Neural Network (PNN).

We have chosen these four types of classifiers for several reasons. We used the FCM method because it doesn't require prior information about the classes; this algorithm has been already used successfully in image segmentation. In addition, FCM is a

classification method which allows a data sample belonging to two classes or more (with a membership degree for each class) contrary to the K-MEANS algorithm where the data sample must belong or not to one class. We thought to implement these two methods for comparing performances between the obtained results to deduce the influence of fuzzy logic in the area of images classification (in our case, the images represent handwriting Arabic words). The third implemented algorithm is KNN, opposing to FCM, it requires a reference base (word images already classified). The latter chosen approach is the neural one because neural networks have a large capacity of classification and showed their skills in handwriting recognition.

6 Results and Discussion

In this paragraph, we give some classification results with different pairs of (features types/ classifiers). The aim of our work is to find the most interesting combinations at feature and classification levels. The results are summarized in tables 1 and 2.

Note that we have divided our base into two parts, one for training (270 images for each city name) and the second for the test (30 images for each city name).

Table 1. Recognition rates for individual classifiers

Features \ Classifiers	FCM(%)	KMeans(%)	KNN (%)	PNN (%)
Zoning features	65.24	67.32	76.76	78.00
Zernike moments	59.95	58.91	61.80	74.50
Freeman chain code	69.50	66.63	74.87	77.21

The table 1 gives the recognition rates for the four individual classifiers using different feature sets. From these results, we can conclude that the most interesting features are the structural and zoning features. In addition, we note that Gaussian parametric evaluation (PNN) gets good results with zoning. Overall, the two classifiers KNN and PNN reach comparable results which are more interesting than the two others.

At the end of our work, we combined the four classifiers. This combination is made either in a democratic way, in the sense that it does not prefer any classifier compared to another, or by attributing, to the answer of each classifier, a weight according to its performances generally based on the rate obtained in training phase. According to the obtained results (Table 1) we have given priority to the PNN and KNN classifiers while the two others have the same priority. The results are summarized in the following table:

Table 2. Classifiers combination results

Features \ Combination	Without priority between classifiers (%)	With priority between classifiers (%)
Zoning features	70.80	72.09
Zernike moments	61.00	62.03
Freeman chain code	69.65	70.18
All features	77.50	79.80

We have evaluated the performance of our system; we have tested it on database containing 1440 words images. Approximately 80% of these words were properly assigned to the correct class when using all features. This result is very encouraging in handwriting word recognition.

7 Conclusion and Perspectives

In this paper, we have presented a system for holistic (global) handwritten Arabic words recognition, which combines the strengths of both statistical and structural feature extractors thanks to a combination of four complementary families of features (ranging from pure structural to pure statistical and including both local and global features). In the classification phase, our system combines four different types of classifiers in order to get a better recognition rate. The obtained results are interesting and encouraging. They experimentally confirm the assumption that the combination of multiple classifiers decisions and the use of different feature types enhance the overall accuracy of a recognition system. However, for each new pattern recognition systems, problems still remain: How many classifiers and what kind of classifiers should be used? For each classifier, what types of features should be chosen?. Moreover, this involves multiple tedious learning steps for both the chosen features, classifiers and combination rules. We have several possibilities for the evolution of our work: we think that the performance of the system can be increased by combining an analytical approach with the holistic proposed one. We can also increase the performance of our system by automatically selecting the most relevant features with the use of feature selection methods. We can also investigate the field of classifiers ensembles and dynamic classifier selection.

References

1. Bunke, H.: Recognition of Cursive Roman Handwriting - Past, Present and Future. In: International Conference on Document Analysis and Recognition, ICDAR, Edinburgh, Scotland (2003)
2. Steinherz, T., Rivlin, E., Intrator, N.: Offline cursive script recognition: a survey. IJDAR, International Journal on Document Analysis and Recognition 2, 90–110 (1999)
3. Essoukhri Ben Amara, N., Bouslama, F.: Classification of Arabic script using multiple sources of information: state of the art and perspectives. IJDAR, International Journal on Document Analysis and Recognition 5, 195–212 (2003)

4. Lorigo, L.M., Govindaraju, V.: Offline Arabic Handwriting Recognition: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28(5), 712–724 (2006)
5. Souici, L., Aoun, A., Sellami, M.: Global recognition system for Arabic literal amounts. In: *ICCTA 1999*, Alexandria, Egypt (1999)
6. Souici-Meslati, L., Sellami, M.: A hybrid approach for Arabic literal amounts recognition. *AJSE* 29(2B), 177–194 (2004)
7. Farah, N., Ennaji, A., Khadir, T., Sellami, M.: Benefit of multiclassifier systems for Arabic handwritten words recognition. In: *International Conference on Document Analysis and Recognition, ICDAR*, Seoul, Korea, vol. 1, pp. 222–226 (2005)
8. Farah, N., Souici, L., Sellami, M.: Classifiers combination and syntax analysis for Arabic literal amount recognition. *Engineering Applications of Artificial Intelligence* 19(1), 29–39 (2006)
9. Vinciarelli, A.: A survey one off-line cursive Word recognition. *Pattern Recognition* 35(7), 1433–1446 (2002)
10. Arrivault, D.: *Apport des graphes dans la reconnaissance non-contrainte de caractères manuscrits anciens*. PhD Thesis, University of Poitiers, France (2002) (in French)
11. Al-Ohali, Y.: *Handwritten Word Recognition – Application to Arabic Cheque Processing*. PhD Thesis, Concordia University, Montreal, Canada (2002)
12. Azizi, N., Farah, N., Sellami, M.: Off-line handwriting word recognition using ensemble of classifier selection and features fusion. *JATIT, Journal of Theoretical and Applied Information Technology* 14(2), 141–150 (2010)
13. Zouari, H., Heutte, L., Lecourtier, Y., Alimi, A.: Building Diverse Classifier Outputs to Evaluate the Behavior of Combination Methods: The Case of Two Classifiers. In: Roli, F., Kittler, J., Windeatt, T. (eds.) *MCS 2004*. LNCS, vol. 3077, pp. 273–282. Springer, Heidelberg (2004)
14. Souici-Meslati, L., Sellami, M.: Toward a generalization of neuro-symbolic recognition: an application to arabic words. *KES, International Journal of Knowledge-Based and Intelligent Engineering Systems* 10(5), 347–361 (2006)
15. Prema, K.V., Subba Reddy, N.V.: Two-tier architecture for unconstrained handwritten character recognition. *Sadhana* 27, Part 5, 585–594 (2002)