Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 6 (3): 647-650, 2006
ISSN 1812-5654

Arabic Words Recognition by Fuzzy Classifier

[1]Farah Lotfi, [1]Farah Nadir and [2]Bedda Mouldi
[1]Centre Universitaire, Souk Ahras, BP 1553 Souk Ahras 41000, Algeria
[2]Laboratoire d'Automatisme et Signaux, Annaba 23000, Algeria

Abstract: This research study on offline handwriting Arabic words recognition is presented, by the use of a fuzzy proximity measure. In recent years, fuzzy logic has been increasingly used to improve conventional methods especially in pattern recognition fields. The aim of this study was Arabic literal words amount recognition using a fuzzy classifier. We introduce briefly the technique for processing handwritten words, which begins with features extraction, then their classification. The purpose of the classifier was to allocate a class to the test word on a basis of a training set. The fuzzification was introduced in two stages. Firstly to reclassify the obtained K nearest neighbors by a crisp K nearest neighbor approach. Secondly in the classification of the tested word to a class among its K neighbors. The proposed system was tested on 1200 images and a 93.80% classification success rate was obtained.

Key words: Handwritten word recognition, fuzzy nearest neighbors, membership value

INTRODUCTION

Fuzzy logic has been applied to a wide range of applications, including pattern recognition since its introduction in 1965 by Lotfi Zadeh. Pattern recognition can be viewed as a transformation from Measurement Space to Feature Space, to finally the Decision Space[1]. The patterns processed include 2-D images (like faces, satellite pictures etc.) and 3-D objects for robotic manipulators.

The objective of this study is to develop an Arabic word recognition system. Handwritten word recognition is among the most widely studied fields. It supports not only statistical, structural information and semantic ones, but also some physiological and psychological state of the writer.

These characteristics make handwritten word recognition of consent especially in bank checks area.

Word recognition has become during the later decades almost universal. From that, many automatic systems have been developed and implemented. Most existing systems deal with some constraints such as limited lexicons or restricted writers number, resulting in interesting performances.

However, handwritten recognition deals with variability of scripts and noises generated by acquisition tools (scanner, camera, etc).

To recognize a word, a fuzzy K nearest neighbor[2] is implemented in the Arabic handwritten literal amount recognition system described in this study.

The proposed system we deal with consists of five parts: data acquisition, preprocessing, features extraction, recognition and post classification.

In data acquisition a handwritten literal amount is captured by a scanner, after which preprocessing techniques are used to prepare the image of words for features extraction.

The preprocessing stage begins by dividing the literal amount into words, based on vertical histogram analysis and a computed heuristic (the space between words is about 1.5 times greater than the spaces between sub-words). Then, binarisation is done on the obtained words; this consists of having a bimodal image from gray-levels one[3], then a smoothing is used to filter noises[4].

The third part of our system is features extraction; this part is used to reduce the input vector image by measuring (expressing) it, using certain properties or features of the word image.

The features used by our system are the holistic ones, which are ascenders, descenders, loops, etc. These features are quantitatively extracted from the image and used to recognize words.

For the recognition we use a fuzzy classifier to classify words. After feature extraction we use the vector obtained to compare it against a training set of feature vectors. The classifier tries to match these features to one of the 48 class's vectors.

The classifier generates a list of candidate words with maximum proximity, which will be used by a syntactic analyzer to make decision about the word which satisfies the grammatical rules designed for this problem.

Corresponding Author: Farah Lotfi, Cites des Peupliers B 9 N 37 Annaba, Algeria Tel: +0021371174736

Ideally the words using the structural features should be well classified. But this is not the case due to the poor features extracted and variability of the script. This is always a certain amount of overlap between classes in the feature space.

In the proposed system a fuzzy nearest neighbor posses advantages of both nearest neighbor and fuzzy systems and are particularly powerful in handling complex, non linear and imprecise problems[1] such as handwritten word recognition. Two membership functions are used, the first one is to reclassify the generated K nearest neighbors and the second one is to classify the tested word according to the K nearest neighbors.

ARABIC WRITING CHARACTERISTICS

The Arabic language has a very rich vocabulary. More than 300 million people speak this language and over 1 billion people use the Arabic language in several religion-related activities. Arabic script is written from right to left. As opposed to Latin one which start from left to right.

The Arabic script is cursive; its alphabet consists of 28 characters. Ten of them have one dot, three have two dots and two have three dots. Dots can be located above; ت ن, or below; ب ي. The shape of the character is context sensitive, depending on its location within a word. A letter can have up to four different shapes: isolated, beginning connection from the left, middle connection from the left and right and end connection from the right; example (ع ع ع). Certain character combination form new ligature shapes which are often font dependent. Some ligatures involve vertical staking of characters (محمد). These characteristics complicate the problem of automatically segmenting words into characters (known as analytic approach).

Most of the letters can be connected from both sides; the right and the left. However, there are six letters which can be connected from one side only; the right. These letters are (و ر ز د ا ذ), this characteristic implies that each word may contain from one unit or more (sub-words).

Example: اربعون (Forty). This word is composed of four sub-words.

FEATURE EXTRACTION

We have been inspired by the human reading process that considers the global high level words shape[5,6]. For holistic paradigm there is a wide range of methods to word recognition. They can be basically classified in two categories: Statistical and Structural The statistical method is expressed in terms of partitioning the word feature space. The features are statistics based such as spatial distribution of black pixels, number of black pixels etc.

The structural method is expressed as a composition of structural units and a word is recognized by matching its structural representation with that of a reference words

The feature extraction step is carried out to determine word's structures which may be used for recognition. These features are the observables, where the observation provides a value for each of the set of properties.

The main concept is to calculate the number of ascenders, descenders, loops, etc.

Base line detection[4] is the most important information that permits us to situate diacritical dot position and the main part of the word.

The considered vocabulary is composed of 48 words that can be written in an Arabic literal check amount (Table 1).

The boundary following algorithm of the word's image permits to detect different constituents such as: sub words, loops, ascenders, descenders and diacritical dots[7].

Table 1: Vocabulary of Arabic literal amounts

احد	تسعة	ستون	اربعمائة	ألفا	مليارات
اثنان	عشر	سبعون	خمسمائة	الفان	ملايير
ثلاثة	عشرة	ثمانون	ستمائة	مليون	سنتيم
اربعة	اثنا	تسعون	سبعمائة	ملايين	و
خمسة	عشرون	مائة	ثمانمائة	مليونا	دينار
ستة	ثلاثون	مائتا	تسعمائة	مليونان	دنانير
سبعة	اربعون	مائتان	ألف	مليار	سنتيمات
ثمانية	خمسون	ثلاثمائة	الاف	مليارا	جزائري

The structural features used in our approach are high level ones, which are numbers (Fig. 1) of: descenders, ascenders, loops, one dot above, two dots above, three dots above, one dot below, two dots below, Sub words.

The features extracted are corresponding to 9 structural (Fig. 1) ones according to their possible occurrence numbers in the lexicon's word:

Three for ascenders, 2 for descenders, 2 for a one dot above, 2 for two dots above, 2 for three dots above, 1 for one dot below, 2 for two dots below, 3 for loops, 4 for sub words.

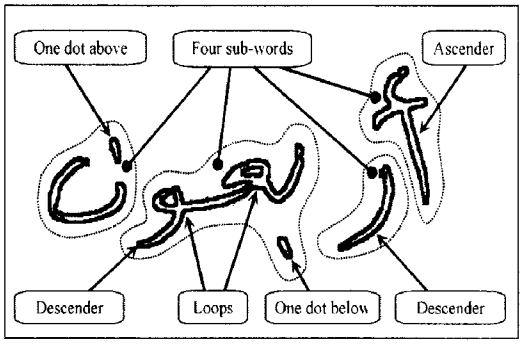

Fig. 1: Word's structural features (Forty)

Example: For the word ثلاثمائة Three hundred), we have: 3 ascenders, 1 double dot above, 2 triple dots above, 2 loops, 3 sub words.

FUZZY K NEAREST NEIGHBOR CLASSIFIER (FUZZY K-NN)

The classifier used in our system is a Fuzzy K-NN, which consist on proximity measures. It has been suggested by Pal and Majumder[8].

Fuzzy nearest neighbor classifiers are ideally suited for modeling the non parametric distribution on handwritten word recognition data.

For the purpose of our system the data were divided as of training and test type.

For a given word X, the fuzzy classifier computes the membership X in different classes $C_1,..,C_j,..C_m$. The membership of X in class C_j can be expressed as $\mu_j(X)$. The test word is allocated to a class for which the membership function yields the maximum value.

After having generated the K nearest neighbors for a test word (by distance similarity), the fuzzification principle is used in two stages. Firstly, it is used in reclassification of the K nearest neighbors obtained by the classical K-NN system. This resignation tries to redefine class boundaries. Formally we express it by: looking for memberships (by distance calculation) of each neighbor (noted y_j) with training classes (noted i class), for every training class we have p_i prototypes noted Z_p, this memrbership function is given in (1):

$$\mu_i(y_j)=\left[1+\left(\max_{p=1..p_i} d\left(y_j,Z_p\right)/Fd\right)^{Fe}\right]^{-1} \quad (1)$$

This function permits to introduce fuzziness, which permits to reclassify y_j in classes where it presents the highest membership value. When neighbor's membership value has been tested with the training set, we compute the membership of test word X noted $\mu_i(X)$ calculated for each of the K nearest neighbor classes, using formula (2):

$$\mu_i(X)=\left\{\mu_i(y_j)*\exp(-a*d(X,y_j)/d_m\right\} \quad (2)$$

dm represents the average distance between words of the same class in the reference set.

a, Fe, Fd are constants that determine the degree of fuzziness in membership space, which has been fixed experimentally to the following values: a = 0.45, Fd = 1, Fe = 1. We have used a threshold S that has been fixed to 0, 7.

Since the value of $\mu_i(X)$ increases while the distance value (d) decreases, therefore, for a tested word using a threshold S, a decision rule is stated as follows: let N be the number of classes where the membership function is greater than S, then:

- If N = 0, X is rejected, membership function too low.
- if N = 1 or N>1 and $\mu_i(X)$is unique, X is recognized
- If N>1 and $\mu_i(X)$ is not unique, there is ambiguousness.

SYNTAX BASED POST CLASSIFICATION

The classification phase has generated a list of candidate words pondered by confidence values, which are the membership values. We will consider from this point that a candidate is a couple of information, the word class and its confidence value.

When obtaining the list of candidate words by the recognition stage, we first sort it by word's confidence value and then we can consider two cases:

- If there is a word which confidence value is greater than the other and if this word succeeds the syntactic analysis (Table 2), the word is kept and will be part of the resulted literal amount. If on the other hand, the word doesn't satisfy the syntax, it is rejected and the next word of the list will be analyzed.
- If at the head of the list, two words have the same confidence value and satisfy the syntactic analysis, we consider this case like an ambiguity, which can be solved with the use of high level information, the courtesy (numeric) amount for example.

RESULTS

For the purpose of the fuzzy K-NN we have constructed four (04) reference sets of different sizes (Table 3), in order to determine the best value of the K parameter and recognition rates.

Table 2: A part of the grammatical rules used

<Hundreds> ::=
<Hund>+و+<Less_Hund> |
<Hund> |
<Less_Ten>+ مائة +و+<less_Hund> |
<Less_Ten>+ مائة
<Hund > ::=
مائتان | ثلاثمائة | اربعمائة | خمسمائة
| ستمائة | سبعمائة | ثمانمائة | تسعمائة
< less_Hund > ::=
< Less_Ten > |
<Comp_Nbr>
< Less_Ten > ::=
احد |
اثنان |
ثلاثة | اربعة | خمسة | ستة
| سبعة | ثمانية | تسعة

Table 3: Reference sets used

Reference set	Set 1	Set 2	Set 3	Set 4
Number of tested words	1200	1200	1200	1200
Number of words in reference set	96	144	240	480

Table 4: Word recognition rates

	Recognition rates		
K	1	3	8
Set 1	85.00	85.00	87.86
Set 2	91.20	92.10	82.10
Set 3	92.30	93.10	90.13
Set 4	92.60	93.80	89.47

Recognition rates obtained according to the reference sets and the value of the parameter K, are represented in Table 4.

From these results the value of the parameter K has been fixed to 3, which represents the K classes with highest membership values.

CONCLUSIONS

The Arabic literal amount considered in our case are composed with 48 words. In this study we have used structural features to perform recognition and we have tested our system on a basis of 1 200 words (the 48 words of the lexicon written by 25 different writers). The word shapes are analyzed with a fuzzy classifier obtaining an average recognition rate of 93.80%.

We conclude that this performance is very interesting and represents a promising platform, on which more investigations and/or improvements may be done.

REFERENCES

1. Pal, S.K., 1992. Fuzzy sets in image processing and recognition. IEEE International Conference on Fuzzy Systems, San Diego, pp: 119-126.
2. Singh, S. and A. Amin, 1999. Fuzzy recognition of Chinese characters. Proc. Irish Machine Vision and Image Processing Conference (IMVIP'99), Dublin, 8-9 September, pp: 219-227.
3. Pavlidis, T., 1982. Algorithms for Graphic and Image Processing. Rockville, MD: Computer Science Press.
4. Belaid, A. and Y. Belaid, 1992. Pattern Recognition: Methods and Applications. International Editions.
5. Madhvanath, S. and V. Govindaraju, 2001. The Role of Holistic Paradigms in Handwritten word Recognition. IEEE Trans. Pattern, Analysis and Machine Intelligence, 23: 149-164.
6. Steinherz, T., E. Rivlin and N. Intrator, 1999. Off-line cursive script word recognition: A survey, pp: 90-110.
7. Ameur, A., K. Romeo-Pakker, H. Miled and M. Cheriet, 1994. Holistic approach for Arabic handwritten words recognition. CNED'94, 3rd National Colloque on Writing and Document, pp: 151-156.
8. Pal, S.K. and D.D. Majumder, 1986. Fuzzy Mathematical Approach to Pattern Recognition. John Wiley, New York.