Proceedings of the 5th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases, Madrid, Spain, February 15-17, 2006 (pp194-198)

A hybrid approach for Multifont Arabic Characters Recognition

NADIA BEN AMOR[1], NAJOUA ESSOUKRI BEN AMARA[2]
[1]National Engineering School of Tunis, Tunisia
[2] National Engineering School of Monastir, Tunisia

Abstract: Pattern recognition is a well-established field of study and Optical Character Recognition (OCR) has long been seen as one of its important contributions. In this paper we describe the performances of a hybrid classification approach which combines both neural networks and hidden Markov models. This classification technique is dealing with features extracted through the wavelet transform method. Experimental tests have been carried out on a set of 85.000 samples of characters corresponding to 5 different Arabic fonts. Some promising experimental results are reported.

Key-words: Arabic Optical Character Recognition, Artificial Neural Network, Hidden Markov Models.

1. Introduction

Arabic is a language spoken by Arabs in over 20 countries, and roughly associated with the geographic region of the Middle East and North Africa, but is also spoken as a second language by several Asian countries in which Islam is the principle religion (e.g. Indonesia). However, non-Semitic languages such as Farsi, Urdu, Malay, and some West African languages such as Hausa, have adopted the Arabic alphabet for writing [1]. Arabic belongs to the group of Semitic alphabetical scripts in which mainly the consonants are represented in writing, while the markings of vowels (using diacritics) is optional.

The cursive nature of the Arabic writing makes recognition more difficult especially when we deal with multifont characters. Many researchers have been working on cursive script recognition for more than four decades. Nevertheless, the field remains one of the most challenging problems in optical character recognition.

The following figure (Fig. 1) shows examples of Arabic characters in the five considered fonts we have worked on.

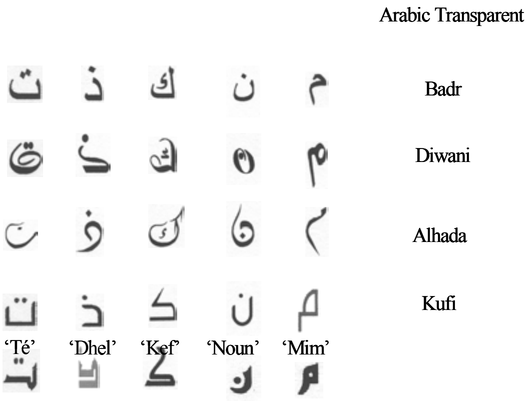

Fig. 1. Samples of different characters' shapes

We present in this paper a hybrid approach for Arabic characters recognition combining two different classifiers. This work belongs to the general field of Arabic documents recognition exploring the use of multiple sources of information. In fact, we have developed so far several methods which had proved the importance of the cooperation of different types of information at different levels (features extraction, classification…) in order to overcome the variability of Arabic and especially multifont characters[2,7,10,11].

In spite of the different researches realised in the field of Arabic OCR (AOCR), we are not yet able to evaluate objectively the reached performances since the tests had not been carried out on the same data base. Thus, the idea is to develop several single and hybrid approaches and to make tests on the same data base of multifont Arabic characters so that we can deduce the most suitable combination or method for Arabic Character Recognition.

In this paper, we present the performances achieved by combining two different methods of classification in an AOCR system based on wavelet transform for features selection. The two methods of classification used are Neural networks and Hidden Markov Models[4,8,12].

In the next section, the whole OCR system will be presented. The different tests carried out and obtained results so far are developed in the third section.

2. Characters Recognition System

The main process of the AOCR system we developed can be presented by the following diagram:

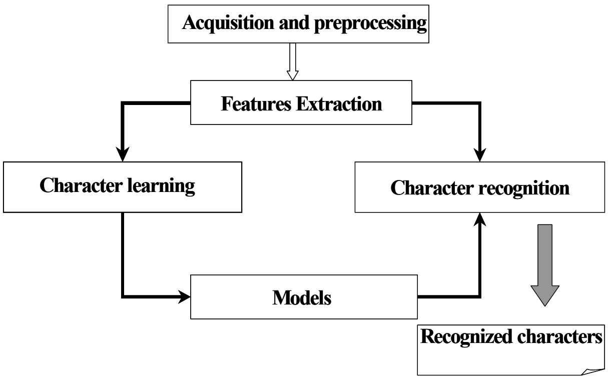

Fig 2. Block diagram of the OCR system

2.1 Pre-Processing

Pre-processing covers all those functions carried out prior to features extraction, to produce a cleaned up version of the original image. In our case, this step is limited to noise reduction which is a random error in pixel value, usually introduced as a result of reproduction and digitalization of the original image.

2.2 Feature Extraction

There is a great number of potential features that one can extract from a finite 2D character shape. However, only those features that are of possible relevance to classification need to be considered. This entails that during the design stage, the expert is focused on those features, which, given a certain classification technique, will produce the most and efficient classification results.

Obviously, the extraction of suitable features helps the system reach the best recognition rate [3]. For this step, we have tested some of well known methods such as Fourier transform and Gabor filter yet, the obtained results using wavelet are far better. Besides, in a previous work, we have used Hough Transform in order to extract features and we have obtained very encouraging results[10]. In this paper, we present Wavelet Transform based method for features extraction. In fact, thanks to its efficiency, wavelet transform is more and more implemented in writing recognition systems and signature verification [6].

In fact, wavelet transform which are widely used in image and signal compression [5] seems to be very interesting as far as features extraction are concerned. We tested several kinds of wavelets such as Haar, Symmlet and Daubechies.

We retained the Daubechies 3 wavelets since we reached better results than when using the other ones.

In addition to features obtained from wavelet transform, we kept the black pixels density of the image as an interesting criteria especially in the case of a multifont context.

Thus, the carried out parameters are as follow:

- the black pixels density
- the mean absolute deviation and the standard deviation of the matrix corresponding to the approximate image
- the mean absolute deviation and the standard deviation of the matrix associated to horizontal details
- the mean absolute deviation and the standard deviation of the matrix associated to vertical details
- the mean absolute deviation and the standard deviation of the matrix associated to diagonal details

2.3 Characters classification

In the following section we describe Neural Networks and Hidden Markov Models based classifier that we have already implemented.

2.3.1 Hidden Markov Models Classifier

Hidden Markov Models or HMMs are widely used in many fields where temporal (or spatial) dependencies are present in the data [12].

During the last decade, HMMs, which can be thought of as a generalization of dynamic programming techniques, have become a very interesting approach in character recognition.

The power of the HMMs lies in the fact that the parameters that are used to model the signal can be well optimized, and this results in lower computational complexity in the decoding procedure as well as improved recognition accuracy. Furthermore, other knowledge sources can also be represented with the same structure, which is one of the important advantages of the Hidden Markov Modeling [9].

- *HMM's Topology*

The retained HMMs use a left-to-right topology, in which each state has a transition to itself and the next state. HMM for each character have 4 to 7 states, but we have noticed that 5 is approximately the optimal number of states .

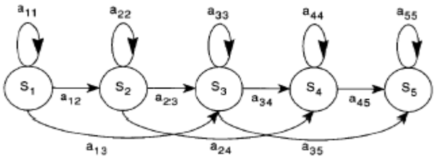

Fig 3. HMM left to right topology

The standard approach is to assume a simple probabilistic model of characters production whereby a specified character *C* produces an observation sequence *O* with probability *P*(*C;O*). The goal is then to decode the character, based on the observation sequence, so that the decoded character has the maximum *a posteriori* probability.

- *Recognition :*

Since the models are labeled by the identity of the corresponding characters, the task of recognition is to identify, among a set of L models λ_k , k=1,…..L the one (the character) which gives the best interpretation of the observation sequence to be decoded i.e:

$$\text{Car}= \underset{1<=car<=L}{\arg\max}[P(O \mid \lambda)]$$

2.3.2 Artificial Neural networks classification

Artificial Neural Networks classifiers (ANN) have been used extensively in character recognition [8,14]. These networks can be used as a feature extractor, when the inputs are scaled or sub–sampled input image, as a "pure" classifier when the inputs are features already extracted or combined with another classifier which is the case in our system .

Two models of neural network were tested : multilayer perceptrons (MLP) and radial basis function (RBF) networks but the best results were achieved with the MLP network.

The system contains three layers of neurons. One layer of neurons for the input, one intermediate layer, and finally an output layer.

- *Multilayer Perceptron Network per Character*

This structure implies the creation of twenty eight networks corresponding each to the HMM observation of an Arabic character in its isolated form. Every network is characterized by:

- an input layer formed by two neurons
- only one hidden layer
- an output layer formed by a neuron corresponding to the considered character

During the learning phase, we present to the system not only the results of the HMM observation of extracted features corresponding to the considered character but also those of others characters. Thus, it will not only learn good observations but also the bad ones.

3. Experimental Results

3.1 Test vocabulary and results

The different tests have been carried out on isolated Arabic characters.

Due to the absence in AOCR of a common data base, we have created our own corpus which is formed by 85 000 samples in five different fonts among the most commonly used in Arabic writing which are: Arabic transparent, Badr, Alhada, Diwani, Koufi (Figure 1). The achieved results for the MLP/HMM approach are shown in table 1. Overall recognition rate is of 98.04% .

By comparing the recognition rates carried out trough these hybrid method with each of these two methods considered in a singular way [7,11], we notice that the hybrid approach provides better results than the HMM models but not as good results as the pure neural solution . However, this is not enough to conclude about the performances of this method.

3.2 Results of using the HMM/MLP classifier :

Characters	ا	ب	ت	ث	ج	ح	خ	د	ذ	ر	ز	س	ش	ص
Recognition rate db3/HMM-MLP	**96.54**	**96.59**	**99.8**	**97.26**	**97.11**	**97.31**	**96.87**	**95.9**	**96.72**	**97.58**	**97.12**	**96.98**	**95.68**	**97.15**

Characters	ض	ط	ظ	ع	غ	ف	ق	ك	ل	م	ن	ه	و	ى	ة
Recognition rate db3/HMM-MLP	**97.59**	**96.14**	**98.11**	**97.34**	**95.73**	**95.73**	**99.61**	**99.34**	**99.97**	**99.85**	**99.98**	**100**	**100**	**100**	**99.97**

Table1 : Recognition rate per character

4. Conclusion

A wide variety of techniques are used to perform Arabic character recognition.

In this paper we presented A hybrid technique based on wavelet decomposition for features extraction and both neural networks and the Hidden Markov Models for classification.

As results show, designing an appropriate set of features for the classifier is a vital part of the system and the achieved recognition rate is indebted to the selection of features especially when we deal with multifont characters.

We are intending to carry out other hybrid approaches on the level of features extraction such as combining features extracted from wavelet decomposition and hough transform in order to take advantages of their both characteristics besides testing other hybrid classifiers such as the neuro-fuzyy one on which we are working.

References

[1] A. Amin. "Arabic character recognition" . In H. Bunke and P. Wang, editors, *Handbook of Character Recognition and Document Image Analysis*, pages 397–420. World Scientific Publishing Company, 1997.

[2] N. Ben Amor, S. Gazeh, N. Essoukri Ben Amara: "Adaptation d'un système d'identification de fontes à la reconnaissance des caractères arabes multi-fontes" *Quatrièmes Journées des Jeunes Chercheurs en Génie Electrique et Informatique*, GEI'2004, Monastir, Tunisia, 2004

[3] E. W. Brown, "Character Recognition by Feature Point Extraction", *Northeastern University internal paper*, 1992

[4] T. Klassen "*Towards Neural Network Recognition Of Handwritten Arabic Letters*" Dalhousie University 2001

[5] N. Ben Amor, N. Essoukri Ben Amara : DICOM Image Compression By Wavelet Transform . *Proc. IEEE International Conference on Systems, Man and Cybernetics*, Vol. 2, 6-9 October 2002.

[6] A Fadhel et P. Bhattacharyya, "Application of a Steerable Wavelet Transform using Neural Network for Signature Verification'', *Pattern Analysis & Application*, Springer-Verlag, London, pp. 184-195,1999.

[7] N. Ben Amor , N. Essoukri Ben Amara: "Applying Neural Networks and Wavelet Transform to Multifont Arabic Character Recognition" *International Conference on Computing, Communications and Control Technologies* (CCCT 2004), Austin (Texas), USA, on August 14-17, 2004.

[8] M Altuwaijri , M.A Bayoumi , "Arabic Text Recognition Using Neural Network" *ISCAS 94. IEEE International Symposium on Circuits and systems*, Volume 6, 30 May-2 June 1994

[9] N. Ben Amara, A. Belaïd and N. Ellouze:"Utilisation des modèles markoviens en reconnaissance de l'écriture arabe : État de l'art" *CIFED 2000*

[10] N. Ben Amor, N.Essoukri Ben Amara "Multifont Arabic Character Recognition Using Hough Transform and Hidden Markov Models" ISPA2005 IEEE *4th International Symposium on Image and Signal Processing*

and Analysis September 15-17, 2005, Zagreb, Croatia

[11] N. Ben Amor, N. Essoukri Ben Amara : "Hidden Markov Models and Wavelet Transform in Multifont Arabic Characters Recognition", *International Conference on Computing, Communications and Control Technologies* (CCCT 2005), July 24-27, in Austin, Texas, USA (Silicon Hills) 2005.

[12] R.-D. Bippus and M. Lehning. "Cursive script recognition using Semi Continuous Hidden Markov Models in combination with simple features". In *European workshop on handwriting analysis and recognition*, Brussels, July 1994.

[13] Patrick K. Simpson, pp. 112-123, *Artificial Neural Systems*, 1990, Pergamon Press, NewYork, NY