Optical Character Recognition For Arabic language using neural network

Abdelkarim Mars

Sience&Technique loboratory
Time Higher School
abdelkarim.mars@gmail.com

Abstract. As part of the DocumentToText project, which is being piloted by TIME University and Horizon data society, we are working to develop an Optical Character Recognition engine to convert all scanned books that exist in Tunisian University into editable textual documents . In this article we present our approach for the development of an OCR system as well as the presentation of the utility of using the artificial neural networks for Arabic characters. We will present the realization context, our point of view on the particularity of the Arabic language with regard to the literature and finally the reasons that have governed the decisions taken in the steps of the realization.

Keywords: OCR; Artificial neural network; Arabic character.

1 Introduction

The purpose of the recognition of writing is to transform a written text into a machine-readable representation easily reproducible by word processor. This task is not easy because the words have an infinity of representations because each person have his own writing, and because each writing can be represented by many fonts and many styles (bold, italic, underline, shaded) and each writing have a different layouts. Depending on the type of writing that a system must recognize (manuscript, cursive or printed), the operations to be carried out and the results vary significantly.

The Optical Character Recognition technology (OCR) knows several practical applications in several fields of activity Among which we can cite:

- Banks and insurance for the authentication of bank checks (Correspondence between amounts and wording on the one hand and correspondence between the identity of the signatory and its signature, on the other), and the verification of clauses contracts for insurance.
- Mail for address reading and automatic mail sorting.
- Police and security for the recognition of mineralogical numbers for the control, authentication and identification of manuscripts and identification of the writer.

The Arabic character recognition is a large problem [1]. This problem is due to the characteristic of the Arabic language[2]. In this project we will work with Modern Standard Arabic (MSA) wich is a standardized version used for official communication across the arab world [3].

Earlier surveys presented both printed character and handwriting, with more discussion about machine-print [4][5]

Our project uses a neural networks approach to recognize the Arabic characters. We will use the Multi Layer Perceptron to learn our OCR system [6]. MLP use backpropagation network to minimize the errors[7] in trainer model. It is simply a gradient descent method to minimize the cost of the total squared error of the result computed by the MLP network.

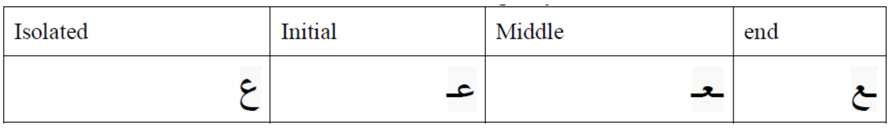
2 Characteristics of the Arabic language

Arabic is the 8th most spoken language in the world, with more than 400 million speakers [8]. Yet, with respect to character recognition technologies, even those of the leaders in dematerialization, performances are low of those for Latin characters.

The most important step on the OCR engines rely on graphical analysis of the image to identify shapes and characters and use their reconciliation method to reconcile characters.

Arabic writing, on the other hand, has its own characteristics which pose difficulties for the engines: [9]

- Arabic is a Semitic language: it uses three-letter roots where vowels are not always written. The engine has difficulties to reconstruct the words.
- The diacritical signs (compulsory signs or which facilitate reading) accompany each word. Thus, during the preprocessing step, these signs can be suppressed by the automatic image enhancement functions (which is particularly preferred for old and / or damaged documents) and thus alter the expected result.
- Graphically, the shape of the characters is lying on the line and not vertical like most other writings. Moreover, it is a cursive writing and the continuity of the characters weakens the segmentation necessary for the identification of characters.
- Arabic letters change their shapes depending on their position in the word; isolated, initial, middle, end (Table 1).



Isolated	Initial	Middle	end
ع	عـ	ـعـ	ـع

Table 1. Change of the shape of a letter according to its position, example of variation of the letter ع "Ayn".

There are, therefore, a number of obstacles inherent in the specificities of the language which may alter the recognition of Arabic characters. However, technologies have evolved and are still evolving to bring more performance and improve the quality of results.

3 Pretreatment steps

Because of the high granularity of the sampling and the various problems lighting and seizure, the image of the character may suffer defects. These problems should be corrected, if possible, before any analysis. Moreover, it is not always useful to use all the points of the image Character to extract the characteristic properties. A reduction step eliminates redundant points. The pretreatment techniques are as follows:

3.1 Smoothing

The image of the character may be tainted by noise due to artifacts acquisition and often to the quality of the document, leading either to absences from points (holes) either to impasto or excrescences and therefore to an overload of points. Smoothing techniques solve these problems.

3.2 Standardization of size

The size of a character can vary from one writing to another, which can cause an instability of the parameters. A natural pretreatment technique consists in bring the characters to the same size. The normalization algorithm we used is imported from OpenCV[1] python library.

3.3 Thinning

The goal of slimming a character is to simplify the image of the character into an easier image to be treated, for example by reducing it to one dimension, that is, the thickness of the character is reduced to one pixel.

4 Neural network approach

There are many methods to train and model an OCR system. Among the existing methods, we mention the neural networks, the neighbor k-nearest, hidden Markov model (HMM), expert systems.

[1] http://docs.opencv.org/2.4/modules/refman.html

On our OCR engine we used a technology based on the neural networks that have been present in the Machine Learning community for decades, winning each year in maturity and answering ever more challenges.

Learning an array of artificial neurons involves the following steps:

- Acquisition of data forming the learning base.
- Pre-processing: it consists of locating, segmenting and normalizing representations.
- Choice of attributes: after the pre-processing, we must extract attributes that define the data. These attributes serve as network entries of neurons.

Before the processing of the data, we have to make the choice of the objects, the definition of the attributes characterizing the objects and the construction of the base learning. At the end of this phase we obtain a table of numbers at two inputs: data and attributes that characterize them.

Learning consists of presenting the examples sequentially and modifying the synaptic weights according to an equation called the learning equation.

Artificial Neural Network consist of simple processing elements and a very high degree of interconnection [10]. The weights of the network are learned from training data. The weight are intialized into the initialized input layer, hidden layers and on the final output layer. We have user the cross entropy function to compute the error rate. The extracted information from data will be processed from input layer to output layer gives a character in this task.

We have developed this algorithm for learning our artificial neural network [11]:

Definition and allocation of the ANN

```
Initialization of all weights
Construction of a standardized training base
For each example e from learning base
Resampling of Example e
Standardization of example e
Extraction and saving all the features of the example e
Saving the label of example e
While the stop condition is not satisfied
For each example e from to the learning base
Propagation of example e
Calculate the local error
Checking the stop condition
End while
// Calculate the criterion
For each example e from the learning base
Propagation of example e
Calculate the local error
Calculating the cumulative error
```

To calculate the errors we use the following function:

$$E_{CE} = -\Sigma \, [d_j \log(y_j) + (1 - d_j) \log(1 - y_j)]$$

5 Evaluation of the OCR System

In the acquisition step, a database should be acquired representing the different Arabic characters. For this reason, we have used a book already scanned from the library (Le cahier de la Tunisie) for the segmentation of words into characters. We were able to extract all the characters that compose the book using the OpenCV library. At the end of this operation, we were able to obtain 100,000 clean characters. The algorithm used in this step uses the Python language.
Once the data are ready we have applied the following pretreatments:

- Cleaning and thinning the image of each character in the database.
- Normalization of the size of a character.
- Centering the image.
- Extraction of attributes.

The classification is done by a network of multilayer perceptron neurons, using the Back propagation algorithm [11].

In order to test our system we used 80% of data for the training of our system and 20% for the test.

Once we finished learning our model, we passed the test data to our system and we got an accuracy of 92%. The results obtained are encouraging, but still require some improvements to begin the conversion of scanned books into texts.

6 Conclusion

We proposed in this article an Optical character recognition system for Arabic language based on neural networks approach. We suggest a method bases on Multi Layer Perceptron classifier, which allow an effective results and a high accuracy. In addition, a neural networks approach allows us to reduce the computational complexity by exploiting the redundancy of the scanning letter. Otherwise, our OCR system still require some improvements. We need to increase the size of our dataset and maybe using deep learning approach to train a new model.

References

1. S. F. Saleh, "An On-Line Automatic Arabic Document Reader", MSc. Thesis, University of Basrah, Iraq, 1998.
2. M. Elleuch, N. Tagougui and M. Kherallah. "Deep Learning for Feature Extraction of Arabic Handwritten Script" Computer Analysis of Images and Patterns: 16th International Conference, CAIP 2015, Valletta, Malta, September 2-4, 2015, Proceedings, Part II. Pages 371--382.
3. Ethnologue: languages of the world, 14th ed: SIL international 2000.
4. B. Al-Badr, and S. A. Mahmoud, "Survey and bibliography of Arabic optical text recognition" Signal Processing, vol 41, 2001.
5. A. Amin, "Off-line Arabic character recognition : the state of the art", Pattern Recognition, vol. 31, pp 517-530 1998.
6. A. Ashiquzzaman and A. Tushar. " Handwritten Arabic Numeral Recognition using Deep Learning Neural Networks ". Computer Science and Engineering Department, University of Asia Pacific, Dhaka, Bangladesh. 15 February 2017.
7. R. Sarkhel, N. Das, A.K. Saha, and M. Nasipuri, A multi-objective approach towards cost effective isolated handwritten Bangla character and digit recognition, Pattern Recognition, 58, 2016, pp. 172-189.
8. “Languages spoken in each country of the world,” http://www.infoplease. com/ipa/A0855611.html, accessed: 2016-12-25.
9. R. Duda, P. E. Hart and D. G. Stork, Pattern Classification, Second ed: John Wiley & sons, Inc., 2001
10. A, Mars. & G, Antoniadis (2015) "Handwriting recognition system for Arabic language learning". International Journal of Engineering and Advanced Technology Studies Vol.3, No.7, pp.55-63, September 2015.
11. Mars, A. and Antoniadis, G. 2016 "Arabic on-line handwriting recognition for Arabic using neural network ". International Journal of Artificial Intelligence and Applications (IJAIA), Vol. 7, No. 5, September 2016