Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 73 (2015) 298 – 305

Procedia
Computer Science

The International Conference on Advanced Wireless, Information, and Communication Technologies (AWICT 2015)

Diacritical Language OCR based on neural network: Case of Amazigh language

Khadija EL GAJOUI[a], Fadoua ATAA ALLAH[b], Mohammed OUMSIS[c]

[a] *LRIT, Faculty of Sciences – Rabat, Mohammed V University, Rabat, Morocco*
[b] *CEISIC, The Royal Institute of Amazigh Culture, Rabat, Morocco*
[c] *Department of Computer Science, School of Technology-Sale, Mohammed V University, Rabat, Morocco*

Abstract

Document paper conversion into electronic format has become indispensable task in many areas, especially for digitizing and translating printed texts. In this context, several approaches have been studied focusing mainly on character recognition for diacritic-free languages. However in this paper, we are interested in the Amazigh language transcribed in Latin, distinguished by its diacritical characters. Thus, we propose to use a system based on neural networks, and to study its behavior against this type of characters.

Keywords: OCR; Amazigh; neural network.

1. Introduction

Optical character recognition (OCR) is a field of research in pattern recognition, artificial intelligence and computer vision. It has been applied as a recent technology across a spectrum of industries, revolutionizing the document management process. This technology allows scanned documents to be transformed into fully searchable documents with text content recognized by computers.

The Amazigh language is spoken by a significant part of the population in North Africa. Recently, it was recognized as an official language in Morocco, after being exclusively reserved for family and informal domains. Hence, the importance to develop an OCR system treating Amazigh writing transcribed in Latin, to contribute into its preservation by digitizing its literary heritage, especially that the most existing systems for Amazigh focus on Tifinagh writing[1, 2].

doi:10.1016/j.procs.2015.12.035

Generally, OCR systems are composed of different modules. The architecture of each system varies from one to another as needed. The studies undertaken, in this field, have proposed several approaches and techniques for each

module[3]. In the remaining of this paper, we introduce the OCR system architecture, in Section 2. In Section 3, we present the different approaches developed for classification module. In Section 4, we introduce the Amazigh language writing systems. In section 5, we present our proposed system. Then, we show, in section 6, the evaluation of the proposed system tested on a set of documents extracted from different books. Finally, in Section 7, we draw conclusions and suggest further related research.

2. Optical character recognition systems

An OCR system takes a text image as input and applies certain treatments through modules making up the system in order to output editable file with the same text[4, 5]. Generally, an OCR system is composed of the following phases[6]:

- Preprocessing phase: It consists on a set of treatments applied to the image in order to increase its quality. It prepares the sensor data to the next phase.
- Segmentation phase: It delimits document elements (line, word, character ...) with the purpose to increase the performance of OCR systems.
- Feature extraction phase: It defines features characterizing the delimited elements of the document.
- Feature extraction: It is one of the most important steps in the system. It describes various features characterizing the segmented elements.
- Classification phase: It is the most important phase of an OCR system. It has the aim to recognize and identify each element. It is performed based on the extracted features.
- Post-processing phase: It is an optional phase. It allows result's verification.

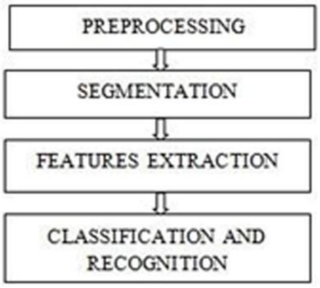


Fig 1. Steps of OCR Systems

3. State of Art

Each OCR module has an important role in the system's functioning and the success rate. However, the classification module is considered as the main phase of all systems. This module consists in identifying each character by assigning it to a correct character class. It helps to decide on the identity of a character from a learning form. In this context two kinds of approaches have been developed:

Statistical approaches:

They are based on the statistical study of measurements of the shapes to be recognized[3]. The study of their distribution in a metric space and statistical characterization of classes allow taking a decision on recognition of the type "highest probability of belonging to a class"[7]. For this kind of approaches, we mention four statistical methods among those most commonly used:

- **Bayesian method**

The Bayesian method consists in selecting from a set of characters one for which the following primitive extracted
has the highest posterior probability relative to the characters previously
learned.

- **Nearest neighbor**

The KNN (K Nearest Neighbors) method compares the unknown form with forms stored in a reference class named prototype and assigns it to its closest class. This method has the advantage of being easy to implement and provides good results. Nevertheless, its main disadvantage is related to the low speed of classification due to the large number of distances to calculate.

- **Neural networks**

Artificial neural networks are composed of simple connected elements (or neurons). These elements were strongly inspired by the biological nervous system [7, 8]. The choice of the network architecture is a compromise between the computational complexity and the recognition rate. However, the strength of neural networks is their ability to generate a region of decision of any form, required by a classification algorithm, at the cost of integrating additional layers of cells in the network.

- **Hidden Markov Model**

The Hidden Markov Model (HMM) is a probabilistic method whose model is composed of a set of states, transition probabilities between these states and the observations made by the system on an image. These observations are represented by random variables, whose distribution depends on the state[9].The HMM is a sequential representation of the characteristics of the input image.

Structural approaches:

Structural methods are based on the physical structure of characters. They try to find simple or primitive elements and describe their relationships[4, 5]. Primitives are topological type as: a loop, an arc…. A relationship may be the relative position of a primitive compared to another [6, 7].
Among the structural methods, we can
mention:

- **Test methods**

They consist in applying tests on each character concerning the presence or absence of single elements or primitives to determine its class.

- **Chain comparison**

The characters are represented by chains of primitives. Comparison of character treated with the reference model
consists in measuring the similarity between two chains and decide on it. The measure of similarity can be done by calculating the distance or by the examination of the inclusion of all or a part of a chain in the other.

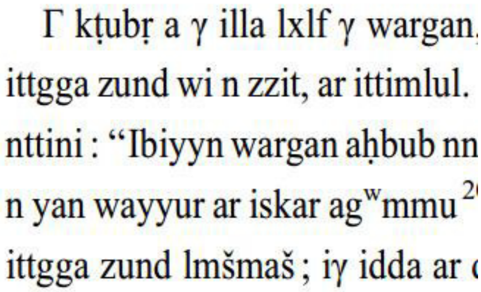
4. Amazigh writing

The Amazigh language, or Tamazight, is present in a dozen of countries across the Maghreb-Sahel-Sahara: Morocco, Algeria, Tunisia, Libya, Egypt, Niger, Mali, Burkina Faso and Mauritania. But Algeria and Morocco are by far the two countries with the largest Amazigh population.
Three writing systems are used[9, 10] to transcribe Amazigh language in
Morocco:

- Tifinagh is the authentic alphabet, attested in Libyan inscriptions since antiquity, and the official script in Morocco since 2003.
- Arabic alphabet used since the Arab arrival on the 6th century.
- Latin alphabet used since the end of the 19th century by colonial scholars, and later by national researchers. In this work, we focus on Amazigh language transcribed in Latin.

After exploring a set of Amazigh documents transcribed in Latin, such as “CHOICE OF BERBER TALK VERSION OF SOUTHWEST MOROCCAN” by Arsène Roux[11] “MOTS ET CHOSES BERBERES” by Emile Laoust[12] and “THE ARGAN TREE AND ITS TASHELHIYT BERBER LEXICON” by Harry Stroomer[13], we found that the Latin characters used in the transcription are represented in Latin, Extended-A Latin and Extended Additional Latin encoding blocks.

The figure and table below show respectively an example of text written in Amazigh language transcribed in Latin and an example of characters used in this transcription. These characters are composed of Latin alphabet and diacritics that represent a set of marks accompanying a letter or grapheme. Diacritics can be placed above (superscript diacritic), below (subscribed diacritic) or after (adscript diacritic).



Γ kṭubṛ a γ illa lxlf γ wargan,
ittgga zund wi n zzit, ar ittimlul.
nttini : “Ibiyyn wargan aḥbub nn
n yan wayyur ar iskar ag^{w}mmu^{2}
ittgga zund lmšmaš ; iγ idda ar

Fig 2. An example of text excerpt from the book “THE ARGAN TREE AND ITS TASHELHIYT BERBER LEXICON”

Table 1. An example of characters used in Amazigh transcription in Latin

Ā	ā	Ă	ă	A^{c}	a^{c}	B^{w}	b^{w}	B^{c}	b^{c}
Ḅ	ḅ	Ḍ	ḍ	D^{c}	d^{c}	Ĕ	ĕ	F^{w}	f^{w}
Ġ	ġ	G^{w}	g^{w}	Ḡ	ḡ	Ḥ	ḥ	Ḫ	ḫ
K^{w}	K^{w}	L^{c}	L^{c}	Ḷ	ḷ	M^{w}	m^{w}	Ŏ	ŏ
Ṛ	ṛ	Ṡ	ṡ	T^{c}	t^{c}	Ṭ	ṭ	Ů	ů
Ŭ	ŭ	Ẓ	ẓ						

5. The proposed system

With the aim to elaborate an OCR system, able to meet the needs of recognizing the Amazigh language characters containing diacritical marks, we proposed to adopt the neural network approach for classification phase. This step is considered as the most important phase of an OCR system. For this reason, we choose to use neural network classifier, known by its good ability to generalize and learn from data and examples, similarly to the human ability to learn from experiences.

Neural networks are based on two phases: learning and recognition. The first phase consists in descending iteratively the network layers and adjusting the weights at each passage.

The type of neural network chosen is Multilayer Perceptron (MLP), because it is the most used in literature. It has a general-purpose model, with a huge number of applications. MLP is characterized by its capacity of modeling complex functions and its robustness. Furthermore, it is good at ignoring irrelevant inputs and noise[14].

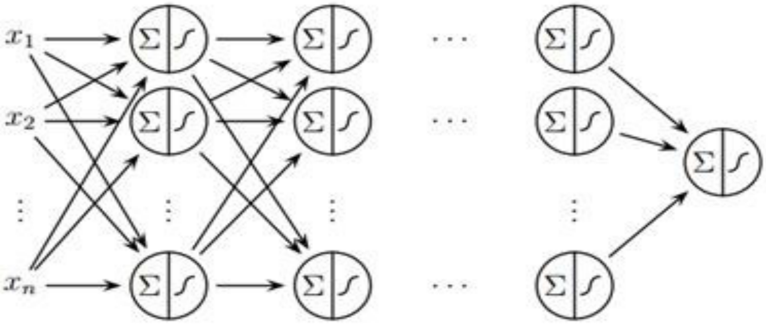


Fig 3. MLP architecture

Multilayer Perceptron function:

Considering the most classical case of a single hidden layer neural network, mapping a d-vector to an m-vector (e.g. for regression):

$$\mathbf{g(x) = b + W \tanh (c + Vx)}$$

Where:

x is a d-vector (the input)

V is an k * d matrix (called input-to-hidden weights)

c is a K-vector (called hidden units offsets or hidden unit biases)

b is an m -vector (called output units offset or output units biases)

W is an m * h matrix (called hidden-to-output weights).

With their remarkable ability to derive meaning from complicated or imprecise data, neural networks are an excellent solution for the OCR classification. They can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. The OCR process is characterized by noisy inputs, image distortion, and differences between typefaces, sizes, and fonts. Hence, the importance of the neural networks' approach in performing character recognition, due to their high noise tolerance.

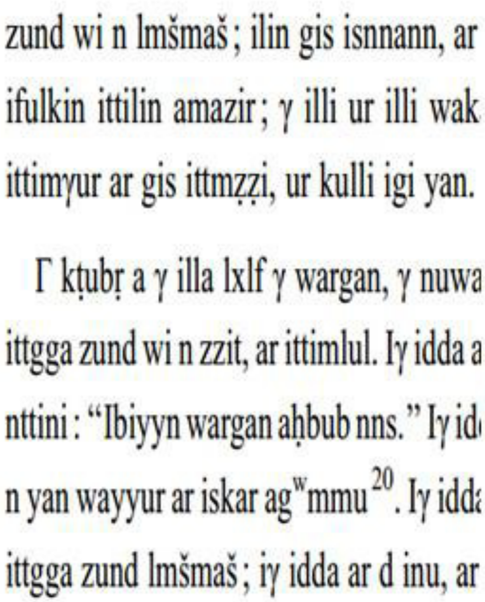
6. Experiments & Results

6.1. Working environment

For this study, we created first a corpus containing Latin characters used in the transcription of the Amazigh language. Then, we used OCRopus tool for experimentation.

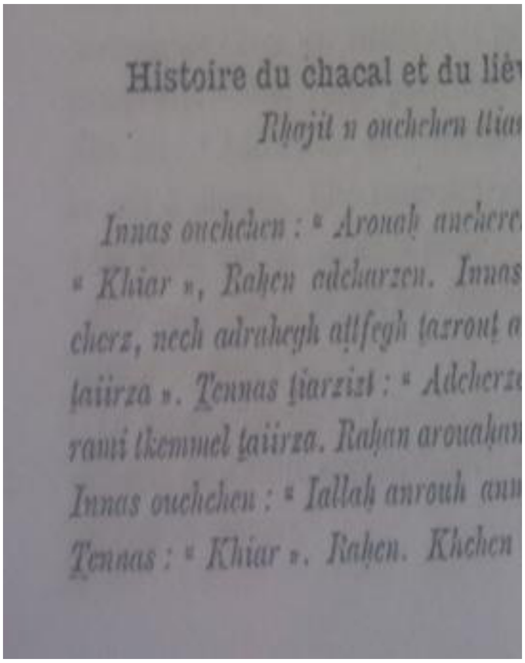
6.1.1. Data

To train our system, we created images containing a text line. The training corpus is composed of 10,000 images and the test corpus of 1,000 images. Learning is run in more than 20,000 iterations and gives birth to 20 different neural network models. The test phase shows that the best model gives a percentage of 97%. This model is then chosen and the system is tested on a set of documents taken from different books. The documents used are 220 pages collected from 4 different books[12,13,15,16, 17] written in Amazigh language transcribed into Latin.
Part of this collection has undergone a pretreatment to increase the image quality, while the other is kept with low quality in order to view the system behavior in both cases. The documents are divided into two parts:
Doc 1: documents in good quality.
Doc 2: Skewed documents in low quality.
Figure 3 and Figure 4 show an example of those documents.



zund wi n lmšmaš ; ilin gis isnnann, ar
ifulkin ittilin amazir ; γ illi ur illi wak
ittimγur ar gis ittmẓẓi, ur kulli igi yan.

Γ kṭubṛ a γ illa lxlf γ wargan, γ nuwa
ittgga zund wi n zzit, ar ittimlul. Iγ idda a
nttini : "Ibiyyn wargan aḥbub nns." Iγ id
n yan wayyur ar iskar ag^{w}mmu[20]. Iγ idd
ittgga zund lmšmaš ; iγ idda ar d inu, ar



Histoire du chacal et du liè
Rhajit n ouchchen ttia
Innas ouchchen : « Arouah anchere
« Khiar », Rahen adcharzen. Innas
cherz, nech adraheγh attfegh tazrout a
taiirza ». Tennas tiarziat : « Adcherz
rami tkemmel taiirza. Rahan arouahan
Innas ouchchen : « Iallah anrouh an
Tennas : « Khiar ». Rahen. Khchen

Fig 4. An example of Doc 1

Fig 5. An example of Doc 2

6.1.2. OCRopus

OCRopus is a free document analysis and optical character recognition (OCR) tool, featuring pluggable layout analysis, pluggable character recognition, statistical natural language modeling, and multi-lingual capabilities. It is based on statistical approaches using multi-layer perceptrons (MLPs) for character recognition [18].

OCRopus is becoming a powerful tool for optical character recognition, capable to analyze a complex layout (containing columns and boxes…). OCRopus does not reconstitute the page layout after processing, but performs the recognition in a logical order after analyzing the layout. Although, its use on the command line is very simple, OCRopus is not yet available as a GUI, or integrated into an existing graphical tool (like gscan2pdf or XSane ...).

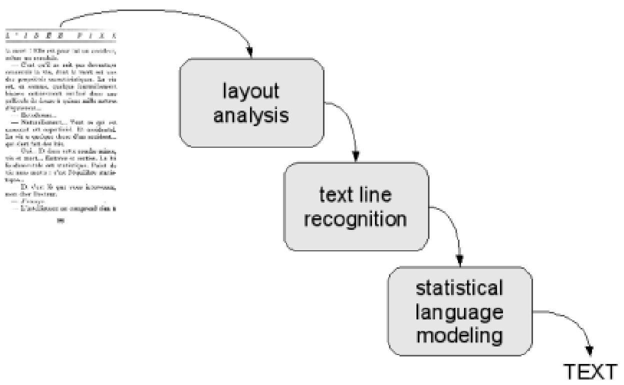


Fig 6. Architecture of the OCRopus tool

The overall architecture of OCRopus consists of three major components:

- Layout analysis: Identifies text columns, text blocks, text lines, and determines the reading order.
- Text line recognition: Separates text line of images into a collection of characters. Then, it performs the character recognition based on a hypothesis graph.
- Statistical language modeling: Integrates alternative recognition hypotheses with prior knowledge about language, vocabulary, grammar, and the domain of the document.

6.2. Results & Analysis

We test our system with documents of different qualities containing text in Amazigh transcribed into Latin alphabet and English. The purpose of this test is to study the behavior of this based neural network system against diacritical language (Amazigh) and diacritic-free language (English). Moreover, to visualize the impact of variation in the quality of the image on the proposed system.
The results obtained are shown in the following table:

Table 2. Recognition rates

Recognition rates		Nature of language	
		Diacritical language	Diacritic-free language
Document quality variation	Doc 1	96%	99%
	Doc 2	60%	65%

We note that the recognition of good quality documents (Doc 1) provides important percentages for diacritical and diacritic-free language. The recognition rate for Amazigh language reaches 96%. However there are some misclassifications errors for examples:

- The capital letters are confused with lowercase in some cases.
- The character "G" is confused with "Ḡ", "Ů" with "U" and "ε" with "s".
- "W " is generally not recognized.
- Spaces are sometimes missed.

In the low-quality documents case (Doc 2), the rate varies remarkably compared to good quality documents. Several recognition errors appear in both cases. Those errors are usually due to breaks in characters that are either broken or losing a part of their body.
The errors remarked on diacritical language are:

- Characters that are recognized as two characters, such: "ů" is recognized as "ii" or "u" as "ṛr".
- Confusion between characters: "ḏ" and "ṯ" with "ḷ", "g" with "ġ", ….

The same types of errors are found for diacritic-free language. We remark confusion between "e" and "c", "a" and "u", "nn" and "m", "h" and "lr", ….

7. Conclusion

In this work, we described OCR systems and their architecture composed of several modules. We also presented a state of the art of the most important module which is the classification phase.

We created a corpus containing over than 10,000 text line images of the Amazigh language transcribed into Latin alphabet. Using the OCRopus system, we compared the behavior of our system based on neural networks to a diacritical language and a diacritic-free language with a different quality paper.

The results showed that good quality document provides interesting recognition rate for the Amazigh language but the percentage for diacritic-free language remains more important. For poor quality documents the rate decreases due to some problems related to image's preprocessing.

This work opens us interesting prospects, especially for OCRopus system that gives the opportunity to add new modules. Hence, the interest to develop a particular module for diacritical languages likes Amazigh.

References

1. El Ayachi R., Fakir M., Bouikhalene B. Recognition of Tifinaghe Characters Using Dynamic Programming & Neural Network : Recent Advances in Document Recognition and Understanding; October 21 2011.
2. Aharrane N., EL Moutaouaki, K., Satori kh. Recongnition of handwritten Amazigh characters based on zoning methods and MLP : Wesas transactions on computer, volume 14; 2015.
3. El Gajoui K., Ataa Allah F. Optical Character Recognition for Multilingual Documents: Amazighe-French. : The 2nd World Conference on Complex Systems, Agadir, Morocco; November 2014.
4. Eikvil L. OCR, Optical Character Recognition : Norsk Regnesentral; 1993.
5. Belaïd A. Reconnaissance automatique de l'écriture et du document, Campus scientifique, Vandoeuvre-Lès-nancy ; 2001.
6. Charles P., Harish V., Swathi M., Deepthi C.H. A Review on the Various Techniques used for Optical Character Recognition. In: International Journal of Engineering Research and Applications; 2012.
7. Bousslimi R. Système de reconnaissance hors-ligne des mots manuscrits arabe pour multi-scripteurs : Mémoire de mastère ; 2006
8. Noor N. Bangla Optical Character Recognition, Thesis; 2005.
9. Muaz A. Urdu Optical Character Recognition System, Thesis; 2010.
10. Skounti A., Lemjidi A., Nami M. Tirra aux origines de l'écriture au Maroc : Publications de l'Institut Royal de la Culture Amazigh, Rabat; 2003.
11. Roux A. Choix de Version Berbères Parler du Sud-Ouest Marocaine, France; 1951.
12. Laoust E. Mots et Choses Berbères, Paris ; 1920.
13. Stroomer H. The argan tree and its tasheliyt berber lexicon, Université de Leyde, Etudes et document berbères; 2008.
14. Riedmiller M. Machine Learning: Multi Layer Perceptrons : Albert-Ludwigs-University Freiburg AG Maschinelles Lernen.
15. Justinard C. Manuel de berbère Marochain (Dialecte Rifain), Librairie Paul Geuthner, Paris ; 1926.
16. Lasri Amazigh B. Ijawwan n tayri, Marrakech, Imp Ima; 2008.
17. Leguil A. Conte berber grivois du haut atlas, L'Harmattan, Paris ; 2000.
18. Breuel T. M. The OCRopus open source OCR system, Proc. IS&T/SPIE 20th Annu. Symp. ; pp.1 -15 2008.
19. GACI Z. Quel système d'écriture pour la langue berbère (le Qabyle), Mémoir de magister ; 2011.
20. Charles P., Harish V., Swathi M., Deepthi C. A Review on the Various Techniques used for Optical Character Recognition : International Journal of Engineering Research and Applications; Jan-Feb 2012..
21. Chaker I., Benslimane R. Nouvelle approche pour la reconnaissance des caractères arabes imprimés : Revue Méditerranéenne des Télécommunications, VOL.1 No.2 ; 2011.
22. Choudharya A., Rishib R., Ahlawatc S. Off-Line Handwritten Character Recognition using Features Extracted from Binarization Technique : AASRI Conference on Intelligent Systems and Control, AASRI Procedia 4 306 – 312; 2013.
23. Supriana I., Nasution A. Arabic Character Recognition System Development : The 4th International Conference on Electrical Engineering and Informatics (ICEEI); 2013.
24. Mithe R., Indalkar S. and Divekar N. Optical Character Recognition : International Journal of Recent Technology and Engineering (IJRTE); 2013, ISSN: 2277-3878, Volume-2, Issue-1.
25. Jacob R. B. Mathematical Expressi on Detection an Segmentation in Document Images, Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of Master of Science In Computer Engineering; February 12 2014 Blacksburg, VA.