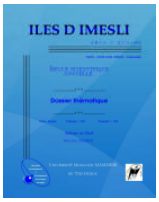


JOURNAL ILES D IMESLI, VOLUME 13, N°02 (2021), PP. 21-27

Handwritten Characters Recognition for Amazigh Script

Mokrane Kemiche[1], *Malika Sadou*[2]
[1]Centre de Recherche en Langue et Culture Amazigh, Algeria, k.muqran@gmail.com
[2]Centre de Recherche en Langue et Culture Amazigh, Algeria, malika.sadou142@gmail.com

Article information

History of the article		
Received: 27/01/2021	Accepted : 06/12/2021	Published : 31/12/2021

Abstract

Amazigh natural language processing faces several challenges, including the ununified Amazigh graphical system and the unavailability of large public databases. In this paper, we elaborate a deep learning architecture that can be effectively applied to recognizing Amazigh latin handwritten characters. It concerns the implantation of a CNN network, which is applied and tested on the BERBER-MNIST dataset. The experimental results showed that the proposed CNN achieved a very interesting recognition rate.

Keywords: BERBER-MNIST, CNN, Character Recognition, Amazigh script, Deep learning.

Resumé

Le traitement automatique du langage amazigh se heurte à plusieurs difficultés, notamment le système graphique Amazigh non unifié et l'indisponibilité de grandes bases de données accessibles au public. Dans cet article, nous élaborons une architecture d'apprentissage approfondie qui peut être appliquée efficacement pour la reconnaissance des caractères manuscrits Amazighs latins. Il s'agit de l'implantation d'un réseau CNN, qui est appliqué et testé sur la base de données BERBER-MNIST. Les résultats expérimentaux ont montré que le réseau CNN proposé a atteint un taux de reconnaissance très intéressant.

Mots-clés: BERBER-MNIST, CNN, Reconnaissance de Charactères, Les lettres Amazighes, Apprentissage profond.

Agzul

Tasenselkimt s tutlayt n tmaẓiɣt tettmagar-d uguren d imeqqranen aladɣa deg wayen icudden ɣer tira d wassaɣen yellan ger uselkim d tira s timmad-is, imi deg wayen yettwaxedmen yakan ar assa deg taɣult-a, ad naff ulac tarrayt s wayes i yezmer uselkim ad yeɛqel isekkilen n tmaziɣt imi ur ddukklen ara, yerna ur sɛin ara yiwen n talɣa. Deg umahil-a ad neɛreḍ ad neg yiwet n tmeskiwt i uselkim iwakken ad yuɣal ad iɛeqqel isekkilen-a. nexdem-d yiwen n uẓeṭṭa asenselkimay i nerra ɣef Berber-MNIST s yiwen n wammud alqayan. Deg wayen nwala deg umahil-a nerra lwelha-nneɣ d akken aẓeṭṭa-a yessaweḍ ad yerr ɣef yiswi-nneɣ imi yefka-d tifrat i waṭas n wuguren i tettmagar tsenselkimt, aladɣa deg waɛqal n yisekkilen, imi tuget deg-sen rran ɣef uselkim.

Awalen igejdanen: BERBER-MNIST, CNN, Aɛqal n yisekkilen, Tasensekelimayt.

Auteur correspondant : Malika Sadou, malika.sadou142@gmail.com
ISSN: 2170-113X, E-ISSN: 2602-6449,
Published by: Mouloud Mammeri University of Tizi-Ouzou, Algeria

Introduction

In recent years, human language technologies are showing more interest in the Amazigh language, and researchers have begun to give more attention to this language. Indeed, several deep learning techniques have been applied for promoting the Amazigh language. Handwritten Character Recognition (HCR) systems remain the most popular research area.

In fact, several character recognition methods have been applied for various languages; examples include English [1], Arabic [2], Chinese [3], Indian [4], etc. However, Amazigh character recognition remains a young field of research and few studies have been carried out the literature.

In this paper, we are interested in the elaboration of a handwritten character recognition system for Amazigh scripts. Thus, we have elaborated a Convolutional Neural Technique (CNN), which we have applied and tested on BERBER-MNIST dataset [5]. The latter is composed of Tifinagh and Amazigh Latin character scripts, but in this study, we have focused only on Upper-case Latin characters.

The remainder of this paper is organized as follows. Section 1 gives a brief overview of previously proposed approaches for Amazigh handwritten character recognition. Then, a description of BERBER-MNIST dataset is detailed in Section 2. In Section 3, we present the elaborated CNN network for recognizing Latin Amazigh handwritten characters. Finally, Section 4 summarizes the conducted study by introducing the realized application for Latin Amazigh handwritten characters recognition.

1. Related works

In this part of the paper, we give an overview of the previous proposed research works in Amazigh handwritten character recognition.

CNN is considered to be the most widely used technique for Amazigh handwritten character recognition. In this setting, Benaddy et al. [6] have proposed an Amazigh handwritten Tifinagh character recognition system, based on a convolutional neural network (CNN). The latter is used especially in the feature extraction phase so that it uses directly the original character images and it doesn't require many preprocessing steps, which makes it flexible and reliable.

Aharrane et al. [7] have proposed an end-to-end system for automatic recognition of Amazigh scripts available in web images and natural scene images. The first step of this approach consists to detect text in images and perform its extraction. Then, the extracted text will be injected into a convolutional network as a trained input set. Finally, an OCR system is developed for Amazigh text script recognition, where a statistical method is used to extract character features in order to distinguish each character from the others.

Gajoui et al. [8] have chosen MLP network to elaborate an OCR system for recognizing the Amazigh language characters containing diacritical marks. To do so, the authors have built a corpus, which contains over 10,000 text line images of the Amazigh Latin scripts, which are trained and learned by the proposed model.

N. Aharrane et al. [9] have elaborated a handwriting character recognition system. While focusing on the features selection phase. The latter consists to distinguish each character from the others by finding a set of features of its description. To this end, the authors have developed a statistical set of features, by decomposing each character image into different overlapped zones. Then, a vector of components is extracted by calculating the density of each zone, and the total length of the histogram projection.
Another probabilistic approach that is applied to Amazigh character recognition is the Hidden Markov Model. In this scheme, Amrouch et al. [10] have proposed an automatic system, which combines Hidden Markov Models and the Hough transform. This approach consists to build a vector of observations by extracting directional information from the Hough transformation of each character. The obtained sequence of information is used for the learning phase of the Hidden Markov model. ELOUAHABI et al. [11] deal with speech Amazigh recognition. In this setting, they have proposed a speaker-independent system for Amazigh Isolated-Word speech, based on Hidden Markov Model Toolkit (HTK). The learning phase of this system consists to train a set of data collected from 60 speakers including both males and females in order to recognize the first 10 digits and 33 alphabets of the Amazigh language. Then, the features have been extracted using Mel frequency spectral coefficients (MFCCs).

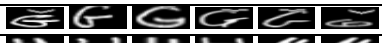
2. BERBER-MNIST Dataset

In this work, we have used the database BERBER-MNIST (Latin version). The latter is an extension of the MNIST data set [12], that we have extended with the Latin Amazight characters of the database AMCD [13].

2.1. Overview of BERBER-MNIST Latin version

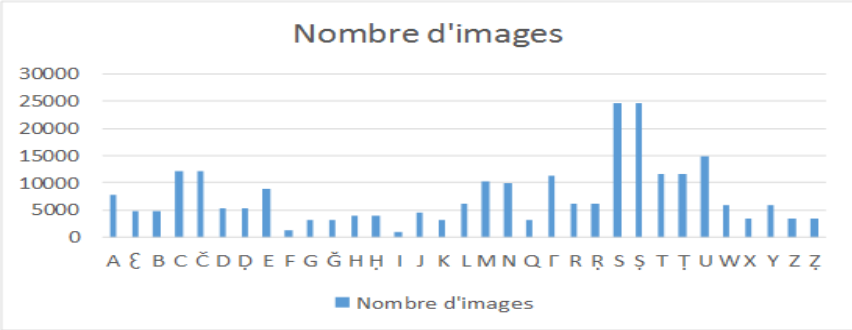
The Latin version of BERBER-MNIST dataset is composed of 247985 images, which are organized into 32 classes of characters. Each image size is 28×28 pixels. The characters specific to Latin Amazigh writing are represented in the following table, illustrated by a few examples in the form of images.

Table N°1. Some examples of Dataset images

Extended letter for Amazigh Script	
Ɛ	
Č	
Ḍ	
Ǧ	
Ḥ	
Γ	
Ṛ	
Ṣ	
Ṭ	
Ẓ	

The following figure represents the number of images for each character stocked in the dataset.

Fig. 1. Distribution of the number of images by character



3. CNN Architecture

In the next section, we present the architecture of the proposed convolutional network.

3.1. Followed process

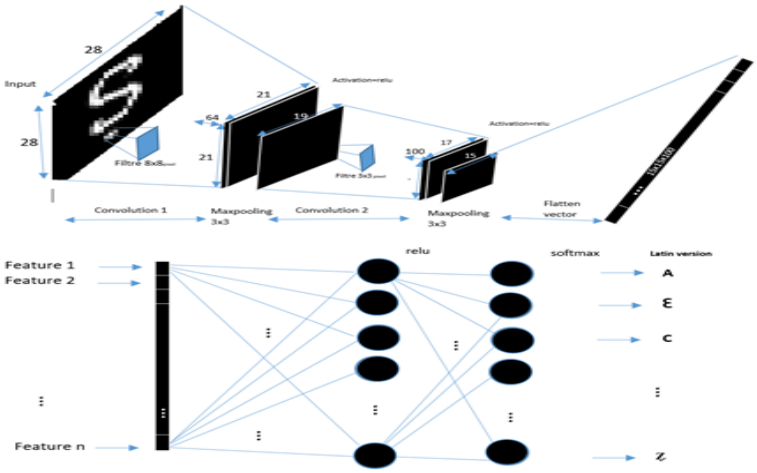
We have firstly partitioned the dataset into two subsets, which are **Train** and **Test**. The first is used in the training phase, while **Test** set relevant to test the accuracy of our CNN's prediction.
In order to normalize the data of this network, a pre-processing was carried out on all the image vectors. Such that computing simplification using standard deviation.

The table below gives the sample sizes for both **Train** and **Test** set.

Train test	Test Set
223185 (90%)	24799 (10%)

The CNN operates in two steps; the first one consists to filter each input image in order to extract specific features. For example, S and Ṣ, the dot is a characteristic that differentiates the two characters. The filter is then responsible for extracting and selecting this dot to distinguish each character from another.
Train set images is analyzed by two convolutional layers, with different filter sizes, which are empirically defined.
Thereafter, we obtain a vector for each image class, which will be injected in the next step into a full-connected neural network. The number of the output of this network is equal to the number of classes to predict. In another hand, this system executes a determined number of iterations to adjust the weights associated with the neuron to achieve a better prediction.
The following figure illustrates the CNN architecture used in our model

Figure 2 : CNN Architecture



3.2. Test and Results

The following table gives the obtained results in 50 epochs.
Using the following formula,

$$accuracy = \frac{correct\ predictions}{total\ predictions} * 100$$

we have reached a 99% of success prediction in the Test set. This leads to a 1% of error prediction.

	Latin version
Train_Accuracy	99.78%
Test_accuracy	99%
CNN Error	1.00%

In order to illustrate the recognition of characters interactively, an appropriate application has been developed. The following figure shows recognition of the character Ğ, entered manually with a mouse in the black square.

Figure 3 : Recognition application screen

Conclusion

Handwritten Character Recognition has a major impact on Amazigh language promotion. However, it remains a fresh research area, which always needs an improvement in accuracy and efficiency.
In this paper, we have applied the most popular deep learning technique, which is the Convolutional Neural Network. Thus, we have firstly elaborated a CNN system, which operates mainly in two steps. And then we have applied for Latin BERBER-MNIST dataset created in previous works. The obtained experimental results have shown that the proposed system provides good performance for the Latin Amazigh character recognition. In future work, we aim to deal with Amazigh handwritten word recognition by developing more reliable deep learning approaches.

Bibliography

[1] Pradeep, J., Srinivasan, E., & Himavathi, S. (2012). Neural network based recognition system integrating feature extraction and classification for english handwritten. International journal of Engineering, 25 (2), 99-106.

[2] Younis, K. S. (2017). Arabic handwritten character recognition based on deep convolutional neural networks. Jordanian Journal of Computers and Information Technology (JJCIT), 3(3), 186-200.

[3] Liu, C. L., Yin, F., Wang, D. H., & Wang, Q. F. (2013). Online and offline handwritten Chinese character recognition: benchmarking on new databases. Pattern Recognition, 46(1), 155-162.

[4] Pal, U., Sharma, N., Wakabayashi, T., & Kimura, F. (2007, September). Handwritten numeral recognition of six popular Indian scripts. In Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) (Vol. 2, pp. 749-753). IEEE.

[5] BERBER-MNIST dataset, https://www.kaggle.com/muqran /the-berbermnist-dataset/version/7. (2021).

[6] Benaddy, M., El Meslouhi, O., Es-saady, Y., \& Kardouchi, M. (2019). Handwritten Tifinagh Characters Recognition Using Deep Convolutional Neural Networks. Sensing and Imaging, 20(1), 9.

[7] Aharrane, N., Dahmouni, A., Ensah, K. E. M., \& Satori, K. (2017, May). End-to-end system for printed Amazigh script recognition in document images. In 2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP) (pp. 1-6). IEEE.

[8] Gajoui, K. E., Allah, F. A., \& Oumsis, M. (2015). Diacritical Language OCR based on neural network: Case of Amazigh language. Procedia computer science, 73, 298-305.

[9] Aharrane, N., Dahmouni, A., El Moutaouakil, K., & Satori, K. (2017). A robust statis-tical set of features for Amazigh handwritten characters. Pattern Recognition and Image Analysis, 27(1), 41-52.

[10] Amrouch, M., Rachidi, A., El Yassa, M., \& Mammass, D. (2009, April). Printed amazigh character recognition by a hybrid approach based on Hidden Markov Models and the Hough transform. In 2009 International Conference on Multimedia Computing and Systems (pp. 356-360). IEEE.

[11] Elouahabi, S., Atounti, M., \& Bellouki, M. (2016, March). Amazigh isolated-word speech recognition system using hidden Markov model toolkit (HTK). In 2016 International Conference on Information Technology for Organizations Development (IT4OD) (pp. 1-7). IEEE.

[12] Deng, L. (2012). The mnist database of handwritten digit images for machine learning research [best of the web]. IEEE Signal Processing Magazine, 29(6), 141-142.

[13] Saady, Y. E., Rachidi, A., Yassa, M., & Mammass, D. (2011). Amhcd: A database for amazigh handwritten character recognition research. Int. J. Comput. Appl, 27(4), 44-48.