Arabic Handwriting off-Line Recognition Using ConvLSTM-CTC

Takwa Ben Aïcha Gader[a], Issam Chibani and Afef Kacem Echi[b]
National Superior School of Engineering, University of Tunis, LR: LATICE, Tunisia

Keywords: Recognition of Arabic Handwritten, Segmentation, Features Extraction, Recurrent Neural Network, ConvLSTM.

Abstract: This work is released in the field of automatic document recognition, specifically offline Arabic handwritten recognition. Arabic writing is cursive and recognized as quite complex compared to handwritten Latin script: dependence on context, difficulties with segmentation, a large number of words, variations in the style of the writing, inter- and intra-word overlap, etc. Few works exist concerning recognizing Arabic manuscripts without constraint, which motivates us to move towards this type of document based on an approach based on deep learning. It is one of the machine-learning approaches reputed to be effective for classification problems. It is about conceiving and implementing an end-to-end system: a convolutional long-short-term memory (ConvLSTM). It consists of a recurrent neural network for spatiotemporal prediction with convolutional structures that allow feature extraction. A connectionist temporal classification output layer processes the returned result. Our model is trained and tested using the IFN/ENIT database. We were able to achieve a recognition rate of 99.01%.

1 INTRODUCTION

The recognition of handwritten words is one of the active and hot research problems in the field of optical character recognition facing increasingly tricky challenges: writing not constrained by the shape of the support and the writer, use of an increasingly large vocabulary, multi-font recognition, multi-writer, multi-script, multi-language.

Although Arabic is spoken by more than 250 million people worldwide, there is no industrial system for automatic recognition of handwritten Arabic writing. The fields of application are, however, numerous: automation of the sorting of postal mail, processing of cheques, invoices, forms, and automatic indexing of old manuscripts. Much work has been done in recent years, but this topic remains an active area of research. The appearance of public databases of a substantial size, such as the IFN/ENIT database, and the organization of competitions have made comparisons between systems possible. They have enabled rapid progress in the field in recent years.

In this work, we are particularly interested in the offline recognition of Arabic handwritten words. This research topic raises a real challenge because the Arabic script is a cursive script and is quite complex compared to the handwritten Latin script: context dependence, segmentation problems, a large number of words, and variations in writing styles, writing, inter- and intra-word overlap. This work aims to perform an OCR of a handwritten Arabic script based on the standard IFN/ENIT database of Tunisian city names. We focus on the segmentation phases, the extraction of primitives, and the power of new deep-learning technologies for word recognition. Deep learning is a machine learning method that is efficient for classification problems, particularly for recognizing images, speech, automatic translation, identification scripting, named entity recognition, and protein sequence classification. We thus aim to recognize handwritten Arabic words by proposing an end-to-end system, more precisely, a convolutional long-short-term memory (ConvLSTM), followed by a connectionist temporal classification output layer.

This paper is structured into four sections: In the next section, we present the state of art in recognizing handwritten Arabic scripts and the work carried out in this field. The third section is devoted to the acquisition and pre-processing of the text image, allowing the preparation of images for the OCR system. The fourth section describes the proposed Arabic handwriting recognition system. The last section

[a] https://orcid.org/0000-0002-3786-3649
[b] https://orcid.org/0000-0001-9219-5228

Gader, T., Chibani, I. and Echi, A.
Arabic Handwriting off-Line Recognition Using ConvLSTM-CTC.
DOI: 10.5220/0011794700003411
In *Proceedings of the 12th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2023)*, pages 529-533
ISBN: 978-989-758-626-2; ISSN: 2184-4313

discusses the obtained results. We end with a general conclusion and some prospects.

2 RELATED WORKS

Due to recent advancements in deep learning technology, numerous deep learning-based solutions were put up for the challenge of Arabic text recognition. In 2008, Graves (Graves and Schmidhuber, 2008) initially published the first deep learning-based method for AHTR from document images. The Multi-Dimensional Long Short Term Memory (MDLSTM) network and Connectionist Temporal Loss (CTC) were employed. The suggested model's accuracy was equal to 91.4% on the IFN/ENIT dataset.

As we continue our study of the development of deep-learning-based techniques, we discuss the one presented in (Abandah et al., 2014). It is based on the graphemic division of cursive words. A characteristics vector is retrieved and sent to a BLSTM, which uses graphemes to exploit the transcript sequences.

IA segmentation-free RNN strategy using a four-layer bidirectional Gated Recurrent Unit (GRU) network with a CTC output layer and the dropout technique was described by Chen et al. in 2017 (Chen et al., 2017). The "abcd-e" scenario was used by the authors to assess the system performance on the IFN/ENIT database, and the accuracy rate reached was 86.4 %.

In 2019, a Convolutional Deep Belief Network (CDBN) framework was suggested for handwritten Arabic text recognition in (Elleuch and Kherallah, 2019). The authors employed data augmentation and dropout regularization to improve the model's functionality and prevent over-fitting. The model's accuracy rate was 98.86% when initially tested against the HACDB character database. Further, it was tested on the IFN/ENIT database and attained an accuracy of 92.9

In 2020, the authors of (Ahmad et al., 2020) suggested a deep learning-based strategy for Arabic text recognition. They employed preprocessing, which included de-skewing the skewed text lines and pruning extra white spaces. They also used data augmentation to train the proposed MDLSTM-CTC model using the KHATT database and achieved a character recognition rate of 80.02 %. In the same year, Mohamed Eltay's (Eltay et al., 2020) Exploring approach consists of a CNN for feature extraction followed by a recurrent neural network concatenated to a CTC layer for the learning and transcription of Arabic handwritten words. The model is trained and tested with the INF/ENIT and AHDB databases. Recognition rates of 98.10 % and 93.57 were reached on the IFN/ENIT and AHDB databases, respectively.

As the last work, we cite the recent one presented in (Albattah and Albahli, 2022), where several deep learning and hybrid models were developed. They used deep learning for feature extraction and machine learning for classification to build hybrid models. The transfer-learning model on the MNIST dataset produced the best results among the standalone deep-learning models trained on the two datasets used in the trials, with an accuracy of 99.67%. While accuracy measures for all of the hybrid models using the MNIST dataset were greater than 0.9%, the results for the hybrid models using the Arabic character dataset were inferior.

3 PROPOSED SYSTEM

The proposed deep neural network for resolving the AHTR problem is presented in this section. It consists of three main end-to-end components: a CNN, an RNN, and a CTC (the used architecture is presented in Figure 2(a)). This combination is the most promising alternative because it outperforms all other strategies. CNN performs the extraction of sequence characteristics from the input pictures. Furthermore, information inside this sequence is propagated via the RNN. It produces a matrix of character ratings for each element of the sequence. The suggested model will be trained using the CTC function, which is used to make inferences for the input image. The CTC decodes the output matrix of the RNN to infer the text recognized from the input image. Without the need for character-level segmentation, word-level recognition is made possible by these two associated networks coupled with the CTC.

3.1 Preprocessing

The first fundamental step in an OCR model is preprocessing, which aims to improve the quality of the images in the database by suppressing distortions or enhancing features to get better results. Even though the IFN/ENIT images had already been extracted and binarized, we used a preprocessing step to resize the input images to the 32×128 shape without distortion.

3.2 Basic Model: CNN+RNN

The hybrid CNN-RNN model has given excellent results in different domains, such as visual description, video recognition of emotions, etc. We performed

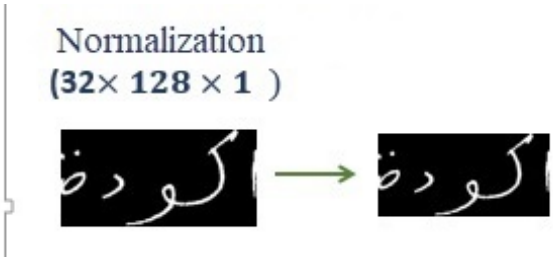

Figure 1: The used reprocessing.

feature extraction using CNN, which acts as an encoder, and then generated a language model and provided a sequence using RNN, which acts as a decoder. Otherwise, the RNN is a language-based model that uses the output features of CNN layers and translates them into natural language sentences. Our end-to-end hybrid NN-RNN model is typically used to map static spatial inputs (images) to sequence outputs (text); the proposed model can perform image classification and text entry leveraging CNN and RNN (see Figure 2(a)). Our model classifies the containing word in the image provided and recognizes it character by character. This allows the model to recognize words even if they are not in the database.

3.3 Methodology: ConvLSTM

The best method for addressing image-related issues is probably a CNN (Convolutional Neural Network). They have demonstrated success in object identification, picture classification, and computer vision. To model sequence-related issues and build predictions on them, LSTMs are used. They are a unique variant of RNN (Recurrent Neural Network) that can recognize enduring dependencies. Generally, LSTMs are frequently employed in NLP (Natural Language Processing) related tasks, including sentence creation, categorization, and machine translation. The CNN-LSTMs models are suggested since conventional LSTMs cannot be applied directly to spatial input sequences. Their inputs either have a spatial structure (two-dimensional (2D) images, one-dimensional (1D) words in sentences, paragraphs, or documents), a temporal structure (the sequence of images in a video or words in a text), or both. Additionally, they can provide output with a temporal structure. As a result, the CNN-LSTM architecture combines LSTM to facilitate sequence prediction with CNN layers for feature extraction on input data. The ConvLSTM model is suggested as a solution to this handwritten text recognition issue.

ConvLSTM is a type of recurrent neural network for spatiotemporal prediction with convolutional structures in input-to-state and state-to-state

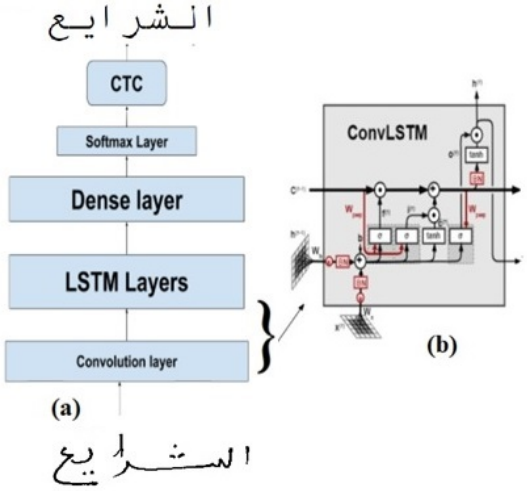

Figure 2: (a): the proposed architecture for text recognition, (b): the ConvLSTM block.

transitions. The ConvLSTM determines the future state of a certain grid cell from the inputs and past states of its local neighbors. This can be easily achieved using an operator convolution in state-to-state and input-to-state transitions (see Figure 2(b)). The key equations of ConvLSTM are shown below:

$$i_t = \sigma(W_{x_i} * X_t + W_{h_i} * H_{t-1} + W_{c_i} \odot C_{t-1} + b_i) \quad (1)$$

$$f_t = \sigma(W_{x_f} * X_t + W_{h_f} * H_{t-1} + W_{c_f} \odot C_{t-1} + b_f) \quad (2)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tanh(W_{x_c} * X_t + W_{h_c} * H_{t-1} + b_c) \quad (3)$$

$$o_t = \sigma(W_{x_o} * X_t + W_{h_o} * H_{t-1} + W_{c_o} \odot C_t + b_o) \quad (4)$$

$$H_t = o_t \odot \tanh(C_t) \quad (5)$$

where, $*$ denotes the operator of convolution and $\odot$ the Hadamard product.

3.4 CTC Function

The Connectionist Temporal Classification CTC (Graves et al., 2006) loss layer provides end-to-end training and free-segmentation transcription. It takes as input the output matrix of the last RNN layer and the ground truth text (GT), then computes the loss value and infers/decodes the matrix to get the text represented in the image. The loss calculation is done by summing up all scores from all possible alignments of the GT text. Alternatively, decoding is done in two steps: first, it takes the most likely character per time step and calculates the best path. Second, it removes duplicate characters and blanks from the path to represent the recognized text. In summary, it converts a prediction into a label sequence. A series of observations serve as its input, while a series of labels, including blank outputs, serves as its output. Our design

uses the ”best path” decoder, with a maximum sequence length of 50 and a class number of 49 (48 different characters included in the IFN/ENIT database + the blank or space character).

4 EXPERIMENTS AND RESULTS

We implemented the proposed offline Arabic handwritten text recognition system using Keras on an HP Z-440 workstation with 16 GB of RAM.

4.1 IFN/ENIT Database

In the IFN/ENIT database (Pechwitz et al., 2002), there are more than 2200 binary images of examples of handwriting forms from 411 writers. Approximately 26,000 images of binary words were isolated from the forms and saved individually for easy access. A ground truth file for each word in the database was compiled. This file contains information about the word, such as the baseline position of the words and information about the individual characters used in the word. Figure 3 represents examples of images extracted from the IFN/ENIT.

Figure 3: Examples of images extracted from the IFN/ENIT.

4.2 Obtained Results

We evaluated the model's training on the IFN/ENIT dataset using accuracy and loss measures. The parameter settings used to train the models are shown in Table 1.

Our model trained over 18 epochs for the dataset and achieved a training accuracy of 94.72% and 96.01% for validation and a training loss of 0.1 and 0.2 during validation. Figure 4 and Figure 5 show the training curves with groups a, b, and c for the training and d for the testing, which reflect a good fit. As can be seen, the learning loss curve lowers until it reaches a stable point, and the loss validation curve does the same, showing a slight gap with the learning loss plot. We conclude that the training dataset has a smaller model loss than the validation dataset. The accuracy

Table 1: Parameter settings.

Image pre-processing	All images are resized to the shape $(32 \times 128 \times 1)$. No other processing.
Training setting	$initial_{epoch} = 0$, $epochs = 25$, $Learning_{rate} = 10^{-6}$, $initial_{weights} = 0.001$, $Batch - size = 30$, Optimizer = $Adam$, Evaluation metrics: Accuracy and Loss,
CTC setting	$max_{sequence\_length} = 50$, Decoder_Type = bestpath and the number of classes $C = 49$

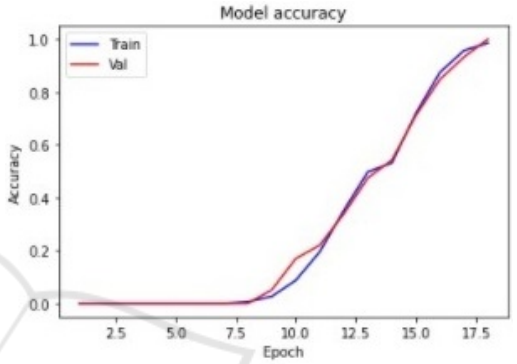

Figure 4: Training and Validation Accuracy curves.

plots for training and validation rise together, stabilizing with a slight gap. The accuracy and loss curves demonstrate a satisfactory match, to sum up. The results show that the model classifies the data with good accuracy, but further improvements are still possible, given that the model has been built from scratch. We also used other database segmentation scenarios and calculated the recognition rate for each one.

Table 2 represents the different results obtained for different scenarios. Interestingly, compared to other state-of-the-art methods evaluated on the IFN/ENIT database, our method outperforms them. This discovery further strengthened our belief that the suggested model is script-independent. Figure 6 presents examples of the model's inferences on Arabic handwritten text images.

5 CONCLUSION AND FUTUR WORK

Handwriting recognition is a dynamic area of study that frequently requires improvement. We proposed a novel approach to carry out Arabic handwriting recognition using deep learning in this work. Our model comprises neuronal layers with convolutional short-term memories (ConvLSTM), a recurrent neu-

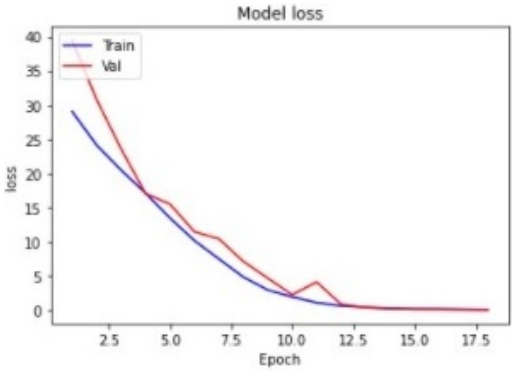

Figure 5: Training and Validation Loss curves.

Figure 6: Examples of recognition.

Table 2: Performance of handwritten Arabic word recognition systems based on IFN/ENIT.

	IFN-ENIT		
	abc_d	abcd_e	abcde_f
Mohamed 2015	-	-	73.5%
Rabie M. 2017	86.73%	-	-
Miled 97	-	-	60%
(Eltay et al., 2020)	98.99%	96.01%	93.57%
Proposed system	99.01%	95.05%	96.57%

ral network used for Spatio-temporal prediction and feature extraction and learning at times, followed by a connectionist temporal classification (CTC) output layer. tries to find an alignment between inputs and outputs. In our case, it allows finding an alignment between the probability vector provided by the RNN and the text of the ground-Truth. The IFN/ENIT database is used to validate the suggested model. We used several training and validation scenarios. We segmented the database into training, validation, and testing sets according to different possibilities, and a maximum accuracy of 99.01% was reached. The model's feature extraction, training, and recognition processes are all intended to be independent of scripts. Without requiring time-consuming, handwritten rules, the model parameters are automatically calculated from the training data. As a result, it can handle languages with cursive handwriting with ease.

REFERENCES

Abandah, G. A., Jamour, F. T., and Qaralleh, E. A. (2014). Recognizing handwritten arabic words using grapheme segmentation and recurrent neural networks. *International Journal on Document Analysis and Recognition (IJDAR)*, 17(3):275–291.

Ahmad, R., Naz, S., Afzal, M. Z., Rashid, S. F., Liwicki, M., and Dengel, A. (2020). A deep learning based arabic script recognition system: benchmark on khat. *Int. Arab J. Inf. Technol.*, 17(3):299–305.

Albattah, W. and Albahli, S. (2022). Intelligent arabic handwriting recognition using different standalone and hybrid cnn architectures. *Applied Sciences*, 12(19):10155.

Chen, L., Yan, R., Peng, L., Furuhata, A., and Ding, X. (2017). Multi-layer recurrent neural network based offline arabic handwriting recognition. In *2017 1st international workshop on Arabic script analysis and recognition (ASAR)*, pages 6–10. IEEE.

Elleuch, M. and Kherallah, M. (2019). Boosting of deep convolutional architectures for arabic handwriting recognition. *International Journal of Multimedia Data Engineering and Management (IJMDEM)*, 10(4):26–45.

Eltay, M., Zidouri, A., and Ahmad, I. (2020). Exploring deep learning approaches to recognize handwritten arabic texts. *IEEE Access*, 8:89882–89898.

Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Graves, A. and Schmidhuber, J. (2008). Offline handwriting recognition with multidimensional recurrent neural networks. *Advances in neural information processing systems*, 21.

Pechwitz, M., Maddouri, S. S., Märgner, V., Ellouze, N., Amiri, H., et al. (2002). Ifn/enit-database of handwritten arabic words. In *Proc. of CIFED*, volume 2, pages 127–136. Citeseer.