Arabic Text Recognition with Harakat Using Deep Learning

Ashwag Maghraby[1†] and Esraa Samkari[2††],
aomaghraby@uqu.edu.sa s44380084@st.uqu.edu.sa

Umm Al-Qura University, College of Computer and Information Systems ,Makkah, Saudi Arabia

Summary
Because of the significant role that harakat plays in Arabic text, this paper used deep learning to extract Arabic text with its harakat from an image. Convolutional neural networks and recurrent neural network algorithms were applied to the dataset, which contained 110 images, each representing one word. The results showed the ability to extract some letters with harakat.

Keywords:
Arabic, deep learning, harakat, neural network, text recognition.

1. Introduction

With the progression of technology, text tools are now being utilized for different purposes, such as to guide users via devices that read text, translate text into other languages, and recognize car license plates to record violations. Optical character recognition (OCR) is one of the tools that extracts text (data) automatically from images, such as newspapers, magazines, and documents [1], [2],[3]. The text on an image can be handwritten or digital, but each has its own challenges. Originally, OCR relied on a procedure called word/character segmentation to detect text in an image, which required developed diversity algorithms that depend on the writing style of the selected language characters [4], [5].

One of the languages found most in source documents worldwide is Arabic, which is used as a formal language in 25 countries and in writing by more than 250 million people. As shown in Table. 1, the Arabic alphabet has 28 letters, which take different forms depending on their position (i.e., beginning, middle, and end), and they are written from right to left. It also uses some diacritics, such as harakat. Harakat are the short vowel marks that consists of four forms: fatha, damma, kasra, and sukun, as shown in Table. 2.The fatḥa is a small diagonal line placed above a letter (ً), and represents a short /a/. The kasra is a small diagonal line below a letter (ٍ), and represents a short /i/. The ḍamma is a small curl-like diacritic placed above a letter (ٌ), and represents a short /u/. The sukun is a circle-shaped diacritic placed above a letter (ْ), and indicates that the consonant to which it is attached is not followed by a vowel [6],[7].

Automatic text extraction of Arabic presents some challenges due to the natural method of writing letters, as well as existing diacritics [4],[8],[9].

Table 1: Arabic letter shapes

letter	Shape				letter	Shape			
	isolated	begin	middle	end		isolated	begin	middle	end
Alef	ا	✖	✖	ـا	Dad	ض	ضـ	ـضـ	ـض
Ba	ب	بـ	ـبـ	ـب	Ta	ط	طـ	ـطـ	ـط
Ta	ت	تـ	ـتـ	ـت	Za	ظ	ظـ	ـظـ	ـظ
Tha	ث	ثـ	ـثـ	ـث	Aeen	ع	عـ	ـعـ	ـع
Jeem	ج	جـ	ـجـ	ـج	Keen	غ	غـ	ـغـ	ـغ
Haa	ح	حـ	ـحـ	ـح	Faa	ف	فـ	ـفـ	ـف
Kah	خ	خـ	ـخـ	ـخ	Qaf	ق	قـ	ـقـ	ـق
Dal	د	✖	✖	ـد	Kaf	ك	كـ	ـكـ	ـك
Dhal	ذ	✖	✖	ـذ	Lam	ل	لـ	ـلـ	ـل
Ra	ر	✖	✖	ـر	Meem	م	مـ	ـمـ	ـم
Zai	ز	✖	✖	ـز	Non	ن	نـ	ـنـ	ـن
Sen	س	سـ	ـسـ	ـس	Haa	ه	هـ	ـهـ	ـه
Shen	ش	شـ	ـشـ	ـش	Waw	و	✖	✖	ـو
Sad	ص	صـ	ـصـ	ـص	Yaa	ي	يـ	ـيـ	ـي

Table 2: Letter “ba” with four forms of harakat

No harakat	Harakat			
	Fatha	Damma	Kasra	Sukun
ب	بَ	بُ	بِ	بْ

One of the consequences of these challenges is lower extraction accuracy [1]. Recently, many researchers applied deep learning to increase the efficiency and accuracy of Arabic text extraction from images that contain text with harakat [1], [2], [10]. They used the convolutional neural network (CNN) algorithm [11], as well as the recurrent neural network (RNN) algorithm [12] to solve both the exploding gradient and the vanishing gradient problems,

Manuscript received January 5, 2023
Manuscript revised January 20, 2023
https://doi.org/10.22937/IJCSNS.2023.23.1.6

which occur when the internal weight values in the hidden layers are either greater than one or fade to zero [1], [10].
CNN is a multi-layered feed-forward unsupervised neural network designed to extract image features without concerns about feature selection problems. RNNs are a type of artificial neural network in which the output from the previous step is used as input in the current phase [13]. CNN can learn to categorize a sentence, a paragraph or an image, while RNNs are excellent at predicting what follows next in a sequence.[1], [10].

In [1], the dataset was based on the Holy Quran (Mushaf Al-Madinah), every line of which was divided into three datasets: normal (containing signs), clean (no marks or dots), and hybrid. The accuracy was tested, and the results show that the character recognition rate (CRR) was 99% and the word recognition rate (WRR) was 95%.
In [2], the objective was to train a model with five diverse types of datasets, each containing digital Arabic text in different font types. The result shows good CRR in most testing scenarios, above 97%, but the WRR accuracy must be enhanced, as it was above 85%.
The researchers in [10] considered the different characteristics of the Arabic text, such as font size, background color, image noise, lighting, and orientation, of the Arabic text. The images in the dataset were from video and natural scenes. Because of the variety of image properties, the CRR and WRR were low, at 75.05% and 39.43%, respectively.

All existing applications and researchers focused on working with images that contain text with harakat, and they tried to measure the accuracy of extracting text without harakat. Harakat is particularly important in the Arabic language, as it signifies both vowels and consonants. In fact, a single character in a word can have four to five different pronunciations depending on the associated harakat mark [14],[15],[16][17]. Therefore, it is essential that Arabic diacritical marks exist. Thus, this paper applied a neural network to extract Arabic words with harakat from an image using both the CNN and RNN algorithms.

The remainder of this paper is organized as follows: Section 2 describes the methodology of this work. In Section 3, the results are discussed. Finally, this paper concludes with a summary and future work in Section 4.

2. Method

The methodology section is divided into three stages. The first stage explains how the dataset was prepared, whereas the second stage shows the implementation of the model. The training images in the model are presented in the third stage.

2.1 Dataset Preparation

This research built its dataset from scratch due to the unavailability of data with the required characteristics. Therefore, data for 110 images was created, with some constraints on the content of the photos, as follows:

- Each image has a single word and associated harakat.
- The dimensions of the image are width = 100 px and height = 70 px.
- Images do not have noise.
- The images are saved with the ".PNG" extension.
- The background color of the images was white for uniformity.
- The text in the images share the same padding.
- All the text in all images share the same orientation.
- All the text in all images differ. However, if the word shape is the same, the harakat differs.
- The text has the same color (black).
- The text font is Arial.
- The text size is 28.

Fig.1 illustrates all constraints of the image. In addition, this dataset can be accessed through the internet [18].
Images were created manually. The PowerPoint program was used to write the text, while the Excel program checked the dedicated words. In addition, the Snipping tool saved the text as images, and image resizing to fixed dimensions was done using an online site (named iloveimg [19]).
After all images were created, they were collected in one folder. In addition, each image's name was used as a label to compare with the prediction words. Therefore, the technique or method of using the table for comparison was not needed.

According to this work, the above dataset was created to identify harakat. Nevertheless, it can be used to test the efficiency of pre-processing (harakat removal); adding specific effects, such as noise; changing the background, e.g., to examine different accuracies; or evaluate the segmentation process.

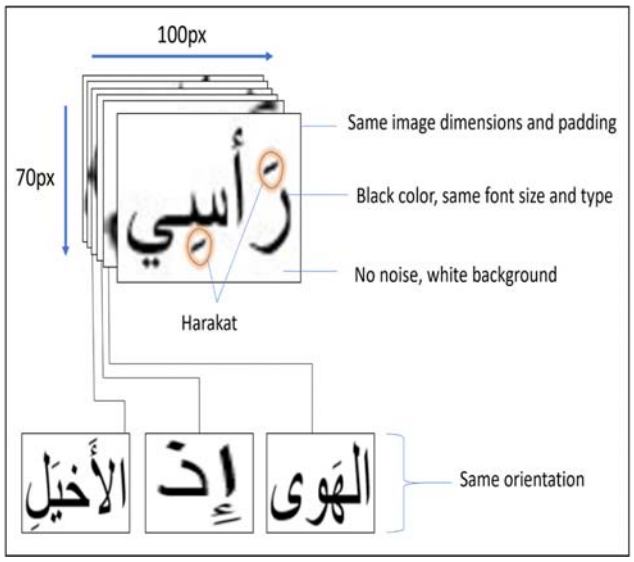

Fig.1: Image constraints

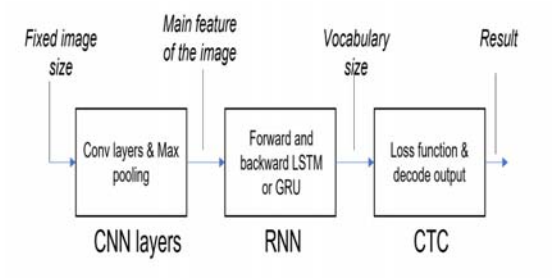

Fig. 2: Overview of model steps

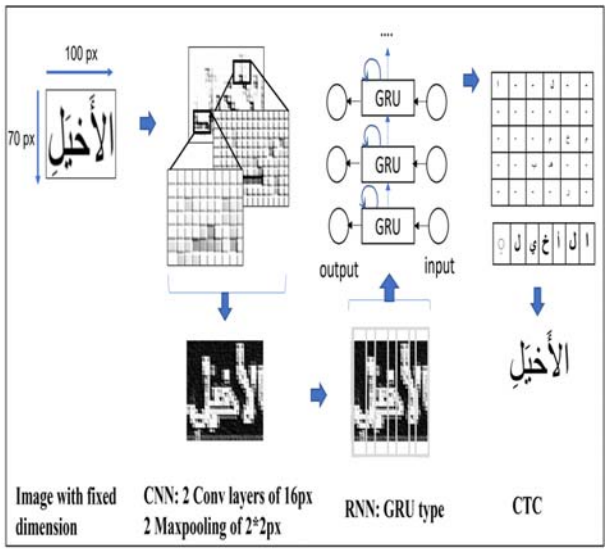

Fig. 3: Research model

2.2 Design the Model

This research followed the same steps used in [1], as shown in Fig. 2. The first step is to determine how many hidden layers should be used in the CNN based on the result achieved in [1], this paper used two convolutional layers with a window of 16 pixels and two max pooling operations with a window of 4 pixels.

Then, in a second step, the Gated Recurrent Unit (GRU)[20], a special type of RNN was selected, as it the work in research [1] showed that using the GRU was better than the Long Short Term Memory (LSTM))[21]. LSTM is a special type of RNN capable of learning long term sequences. GRU uses less training parameter, less memory and executes faster than LSTM whereas LSTM is more accurate on a larger dataset).

In the third step, connectionist temporal classification (CTC) was used to fix the neural weight automatically [1],[22]. Fig. 3 summarizes these steps.

2.3 Implementation

A similar OCR model was available on Keras' site [23]. However, the used dataset contained images of English characters. In addition, each data had the same label size. Therefore, two changes were applied to Keras' code. The first change was to adapt various image label lengths. Meanwhile, the second change was to modify the code's model to be compatible with the model proposed in this research.

Table 3: Layers and parameters of the model

Layer (type)	Output Shape	Param #	Connected to
image (InputLayer)	[(None, 100, 70, 1)]	0	[]
Conv1 (Conv2D)	(None, 100, 70, 16)	160	['image[0][0]']
pool1 (MaxPooling2D)	(None, 50, 35, 16)	0	['Conv1[0][0]']
Conv2 (Conv2D)	(None, 50, 35, 16)	2320	['pool1[0][0]']
pool2 (MaxPooling2D)	(None, 25, 17, 16)	0	['Conv2[0][0]']
reshape (Reshape)	(None, 25, 272)	0	['pool2[0][0]']
dense1 (Dense)	(None, 25, 32)	8736	['reshape[0][0]']
dropout_1 (Dropout)	(None, 25, 32)	0	['dense1[0][0]']
bidirectional_2 (Bidirectional)	(None, 25, 256)	124416	['dropout_1[0][0]']
bidirectional_3 (Bidirectional)	(None, 25, 128)	123648	['bidirectional_2[0][0]']
label (InputLayer)	[(None, None)]	0	[]
dense2 (Dense)	(None, 25, 41)	5289	['bidirectional_3[0][0]']
ctc_loss (CTCLayer)	(None, 25, 41)	0	['label[0][0]', 'dense2[0][0]']

2.4 Training the Data

The dataset contained 110 images, 80% of which were divided for training, 10% for validation, and 10% for testing. The CRR and WRR were used to compute the extraction accuracy. The CRR counts the number of correctly recognized characters for each word, while the WRR counts the number of correct words. The CRR and WRR Eq.1 and Eq.2 are as follows, where the L notations represent all incorrect letters found in all prediction words:

$$CRR = \frac{AllCharactersWords - L}{AllCharactersWords} \quad (1)$$

$$WRR = \frac{\Sigma CorrectPredictionWords}{AllWords} \quad (2)$$

In this research, the CRR calculated whether the harakat was correctly recognized in addition to the letters, and this changed the formal CRR equation slightly, see Eq.3, where the H notation refers to the incorrect recognition of harakat:

$$CRR = \frac{AllCharactersWords - H - L}{AllCharactersWords} \quad (3)$$

3. Results and Discussion

Despite the small dataset size, this research showed the ability of the NN to recognize Arabic words and extract some letters with harakat. For example, each time the proposed model was fed with 88 images for training and 10 images for validation, the testing dataset results were similar.

Most correct predictions of letters and harakat were of the first and last of the word, as shown in Table. 4. The proper recognition of middle letters might occur because some word images had similar data that was already trained previously in the model (see Table. 5). Table. 6 shows some ambiguity in the text recognition, such as confusion between dots and harakat. The CRR of the 12 images in the test dataset was approximately 32.5%, whereas the WRR was 0%. Table. 7 illustrates how these computations were calculated in detail.

The CRR calculation not only considered the correct letters, but it also calculated the correct harakat, which is why its percentage was around the thirds. Yet, the prediction of characters' words sometimes produces an unknown (UNK) value.

In addition, this research manipulated the parameters of the proposed model to emphasize the CRR and WRR results. The exchange between the GRU and LSTM types in the RNN stage did not change the result significantly. When the epoch value was raised, most prediction words with 700 epochs were far away from the actual words. Furthermore, the loss value was around 0.15. Meanwhile, with 1,000 epochs, the outputs enhanced the loss value to nearly 0.02.

4. Conclusion

Extracting the harakat (the damma, fatha, kasra, and sukun) is as important as extracting the Arabic letters. In this research, a dataset was collected manually, and the CNN followed by RNN algorithms were used to extract Arabic text with harakat from image. The result of the OCR model was promising, as some letters were recognized along with their harakat. In addition, some words with similar letters and different vowels were identified. As future work, this research suggests enhancing the dataset volume by increasing the number of training data images.

Table 4: Example of the ability to recognize first and last characters

#	Text of image	Prediction	Similarity
1	فَسِنانُ	أساسانُ	ـا / نُ
2	الجَحفَلِ	المُسدَلِ[UNK]	الـ / لِ
3	شَهِدَ	شِزار	شـ
4	نَكَبتُ	ثُلثُ	نـ / ـتُ
5	الفَتى	العُل	الـ
6	رِقابِ	رأسِ	ر
7	الأَجزَلِ	الأَخيَل	الأ / ل
8	تَحتِ	شَيَبت	ـت

Table 5: Example of the ability to recognize middle characters

#	Text of image	Prediction	Similarity
1	تَسقِني	فأسقِني	سـ / قِـ / نـ / ي
2	عَبس	عَبسَ	عَـ / بـ / س
3	كَريهَة	رَيعَةَ	ر / يـ / ـة
4	غَيرَ	خَيرٌ	يـ / ر

Table 6: Example of the ability to recognize with confusion

#	Text of image	Prediction	Confusion
1	إذ	إذْ	ذ - (Dhal) ذْ - (Dal with sukun)
2	الساقُ	الجَبانُ	قُ - (Qaf with damma) نُ - (Non with damma)

Table 7: Words' images from the test dataset with the CRR and WRR results

#	Text of image	Prediction	Total of characters	Extraction	
				Correct	Incorrect
1	الجَحَلِ	المُسْدِلِ[UNK]	9	4	5
2	كَرِيهَةً	رَبِيعَةً	7	4	3
3	غِيَرَ	خَيْرَ	5	3	2
4	خَوفًا	سَنَاكَ[UNK]	5	1	4
5	أَو	عَرَضَت	3	1	2
6	الخَرِيشَ	الأُوَ	8	1	7
7	بِذَابِهِ	رَرابِرا	9	2	7
8	شَهِدَ	شِزَارِ	6	1	5
9	الهِيذَبَانَ	طَلزِياً[UNK] [UNK]	11	2	9
10	فَبِذَانُ	أَساسَانُ	8	4	4
11	أَلَم	أَحَدِ	5	1	4
12	ذِهرِهِ	رِرامِ[UNK]	7	3	4
	CRR	= (83-56)/83 = 0.325 * 100 = 32.5%			
	WRR	= 0 / 12 = 0 * 100 = 0%			

References

[1]Mohd, M., Qamar, F., Al-Sheik, I., Salah, R.: *Quranic optical text recognition using deep learning models*. In: IEEE Access, vol. 99, pp. 1, 2011.

[2]Fasha, M., Hammo, B., Obeid, N.: *A hybrid deep learning model for Arabic text recognition*. In: International Journal of Advanced Computer Science and Applications (IJACSA), vol.11(8), 2020.

[3]Ranjan, J., Amrita, C., Sk, I.: *Optical character recognition from text imagel*. In: International Journal of Computer Applications Technology and Research, vol.3(4), pp.240-244, 2014.

[4]Anwar, K., Nugroho, H.: *A segmentation scheme of Arabic words with harakat* In: 2015 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT), pp. 111-114, (2015)

[5] Qaroush, A., Awad, A., Modallal, M., Ziq, M.: *Segmentation-based omnifont printed Arabic character recognition without font identification*. In: Journal of King Saud University - Computer and Information Sciences, (2020)

[6]Ryding, K.: *A reference grammar of modern standard Arabic*. In: Cambridge, UK: Cambridge University Press, pp. 25-34, (2005)

[7]Lyovin, A., Kessler, B., Leben, W.: *An introduction to the languages of the world*. In: Oxford, New York: Oxford University Press, pp. 255, (2017)

[8]Yousfi, S.: *Embedded Arabic text detection and recognition in videos*. Ph.D. dissertation, University of de Lyon, Lyon, France, (2016)

[9]Elnagar, A., Al-Debsi, R., Einea, O.: *Arabic text classification using deep learning models*. In: Information Processing and Management, vol. 57(1), pp. 102121, (2020)

[10] Jain, M., Mathew, M., Jawahar, C. V.: "*Unconstrained scene text and video text recognition for Arabic script*. In: 2017 1st International Workshop on Arabic Script Analysis and Recognition, (2017)

[11] Tang, Z., Jialing, Y., Zhe, L., Fang, Q.: *Grape disease image classification based on lightweight convolution neural networks and channel-wise attention*. In: Computers and Electronics in Agriculture, vol. 178, pp. 105735, (2010)

[12] Sherstinsky, A.: *Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network.* Physica D Nonlinear Phenomena, vol. 404(8), p.132306, 2020.

[13] Grossberg, S.: *Recurrent neural networks*. Scholarpedia journal, vol.8(2), p.1888, (2013)

[14] Abu-Rabia, S.: *The role of short vowels in reading arabic: a critical literature review*. Journal of Psycholinguistic Research, vol.48(4), (2019)

[15] Abu-Rabia, S.: *The effect of Arabic vowels on the reading comprehension of second- and sixth-grade native Arab children*. Journal of Psycholinguistic Research, vol.28(1), pp.93-101, (1999)

[16] Abu-Rabia, S.: *Reading arabic texts: effects of text type, reader type and vowelisation*. In: Reading and Writing, vol.10, pp.105–119, (1998)

[17] Abu-Rabia, S.: *The role of vowels in reading Semitic scripts: Data from Arabic and Hebrew*. In: Reading and Writing, vol.14(1-2), pp.39-59, (2001)

[18] Samkari, E.: *Arabic Text with Harakat dataset.* (2022) [Online]. Available:https://drive.google.com/drive/folders/1fdCPlDO3L5rTe-HuR26hq Y5ojM9GCwYS?usp=sharing.

[19] iLoveIMG: *The fastest free web app for easy image modification*. (2022) [Online]. Available: https://www.iloveimg.com/

[20] Chung, J. Gulcehre, C., Cho, K. Bengio, Y.: *Empirical evaluation of gated recurrent neural networks on sequence modelling*. arXiv preprint, vol.1412.3555, (2014)

[21] Hochreiter, S., Schmidhuber, J.: *Long short-term memory*. In:Neural Computation, vol.9(8), pp.1735- 1780,(1997)

[22] Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: *Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks.* In: ICML 2006 - Proceedings of the 23rd International Conference on Machine Learning, (2006)

[23] A_K_Nain: *OCR model for reading Captchas.* (2022) [Online]. Available: https://keras.io/examples/vision/captcha_ocr/.

[24] Google Colab: *FrequentlyAsked Questions*. (2022) [Online]. Available: https://research.google.com/colaboratory/faq.html#resource-limits.

Ashwag Maghraby received the B.S. (2002) and M.S. (2007) degrees in computer science from King Abdulaziz University, Jeddah University and the Ph.D. degree in Intelligent software engineering, The Centre for Intelligent Systems and their Applications (CISA)University of Edinburgh , 2013. Since 2013, she has been an Assistant Professor with the Department of Computer Science, Umm Al-Qura University, and Mecca, Saudi Arabia. For the last three years her student researches awarded the best college research project in Umm Al-Qura University. Her research interest includes software and intelligent systems engineer which focuses on using NLP and machine learning to improve health care and enhance people's daily lives**.**

Esraa Samkari[2] received the B.S. degrees in computer science from the University of Umm Al-Qura University, and Mecca, Saudi Arabia. Right now, she studies M.S. in Umm Al-Qura University. Her research interest includes intelligent systems which focuses on using NLP and machine learning.