

A Deep Learning Approach for Handwritten Arabic Names Recognition

Mohamed Elhafiz Mustafa¹, Murtada Khalafallah Elbashir²

College of Computer and Information Sciences, Jouf University, Sakaka, Saudi Arabia^{1,2}

College of Computer Science and Information Technology, Sudan University of Science and Technology, Khartoum, Sudan¹

Faculty of Mathematical and Computer Sciences, University of Gezira, Wad Madani, Sudan²

Abstract—Optical Character recognition (OCR) has enabled many applications as it has attained high accuracy for all printing documents and also for handwriting of many languages. However, the state-of-the-art accuracy of Arabic handwritten word recognition is far behind. Arabic script is cursive (both printed and handwritten). Therefore, traditionally Arabic recognition systems segment a word to characters first before recognizing its characters. Arabic word segmentation is very difficult because Arabic letters contain many dots. Moreover, Arabic letters are context sensitive and some letters overlapped vertically. A holistic recognizer that recognizes common words directly (without segmentation) seems the plausible model for recognizing Arabic common words. This paper presents the result of training a Conventional Neural Network (CNN), holistically, to recognize Arabic names. Experiments result shows that the proposed CNN is distinct and significantly superior to other recognizers that were used with the same dataset.

Keywords—Deep learning; Arabic names recognition; holistic paradigm

I. INTRODUCTION

Arabic is the official language of more than twenty countries and the mother tongue of more than 300 million people [6]. Arabic is one of the six United Nations official languages. The Arabic script is also used as a medium of writing for other languages like Persian. Moreover, Arabic script is the former script (Ottoman script) of the Turkish language [23]. Unfortunately, research in Arabic language recognition is far behind when compared with languages with the same size and importance. For instance, number of big popular Arabic datasets that are widely used is very few [14]. Moreover, most of these datasets are small. For instance, the IFN/ENIT dataset which is widely used by researchers and also used in ICDAR Arabic recognition competition contains only 26559 word images (names of Tunisian cities) [17], [22].

Arabic script is cursive (both handwritten and machine-printed text). Fig. 1, presents the Arabic alphabet and their different shapes. This makes handwriting recognition very challenging, as we need word segmentation before applying character recognition system. The segmentation methods reported in the literature are far from being robust and very accurate [1], [5]. Segmenting Arabic words to characters is very difficult [12]. There are two reasons for this difficulty. First Arabic letter shape is context sensitive. Some Arabic letters have four shapes according to their position in the word (see Fig. 1). Secondly, in Arabic writing dots are very important and they are not few as fifteen letters out of twenty eight have dots above or below them. Arabic writers do not

place these dots carefully on their proper place and this leads to much confusion. although Arabic writing is horizontal from right to left some letters overlapped vertical. For illustration Fig. 2, contains an image of the Arabic word "عنه" (arguments in English). The image shows three different handwriting. For comparison the Figure also shows some printing for the same word. This is an example for a word that expert Arabic reader recognizes it as one unit without trying to figure out each letter and its dot(s). An analytical automatic recognition system should find the proper segmentation first and then it should find the proper coupling of the dots with its corresponding letter.

In the literature, there are good results for isolated Arabic character recognition (in some papers the recognition is more than 97%) [21], [4], [2], [3]. However, publication on word recognition are few with low recognition accuracy rate [15], [16]. This low rate for word recognition accuracy is mainly due to the error in segmentation [1]. Arabic readers generally tend to recognize common Arabic words and names holistically (i.e. without segmenting them). For new words or non-common ones, the Arabic readers identify the word letters and then recognize the word. This idea becomes appealing for automatic recognition and found good support in last couple of decades [10]. For Arabic words, holistic paradigm is faster and, in some cases, more accurate than analytic paradigm [19], [11]. The disadvantage of the holistic paradigm is its need for huge training data. Moreover, the holistic paradigm does not work alone as the recognition system needs to revert to the analytic paradigm when the word under consideration is not in common words lexicon. bypassing the problems that occur as a result of over segmentation and under segmentation is the main advantage of holistic approach-based classifiers for OCR.

Arabic names used today have much repetition. For instance, in SUST-ARG dataset [13], the twenty most frequent male names represent 52% of the total male names. Application forms normally contain the applicant name. Therefore, for the automatic processing of these forms, the holistic paradigm seems appropriate to quickly recognize common names and to resort to the analytical paradigm to recognize uncommon ones.

In recent years, deep learning has gained great popularity in the pattern recognition filed [9]. It becomes the focus of many researchers since it represents the easiest way to deal with huge data and it automates the feature extraction task. From our literature survey, we notice that although deep learning is the state of the art for OCR, the number of papers which use deep learning for Arabic character recognition is very few [21], [4], [2], [8]. Moreover, All these work address isolated

alone	initial	medial	final	name
ا	ا	ا	ا	alif
ب	ب	ب	ب	bā'
ت	ت	ت	ت	tā'
ث	ث	ث	ث	thā'
ج	ج	ج	ج	jīm
ح	ح	ح	ح	hā'
خ	خ	خ	خ	khā'
د	د	د	د	dāl
ذ	ذ	ذ	ذ	dhāl
ر	ر	ر	ر	rā'
ز	ز	ز	ز	zā'
س	س	س	س	sīn
ش	ش	ش	ش	shīn
ص	ص	ص	ص	ṣād
ض	ض	ض	ض	ḍād
ط	ط	ط	ط	ṭā'
ظ	ظ	ظ	ظ	ẓā'
ع	ع	ع	ع	'ayn
غ	غ	غ	غ	ghayn
ف	ف	ف	ف	fā'
ق	ق	ق	ق	qāf
ك	ك	ك	ك	kāf
ل	ل	ل	ل	lām
م	م	م	م	mīm
ن	ن	ن	ن	nūn
ه	ه	ه	ه	hā'
و	و	و	و	wāw
ي	ي	ي	ي	yā'

Fig. 1. Arabic alphabet and their different shapes

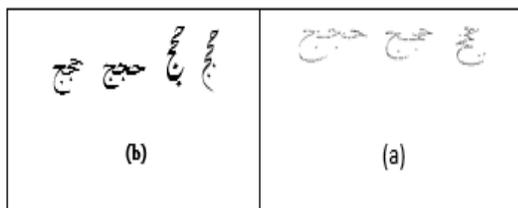


Fig. 2. This figure shows different handwriting and printing for a single Arabic word

Arabic letters and numerals recognition. Therefore, there is a lack of research that uses deep learning to recognize Arabic words holistically (i.e. using holistic paradigm).

The rest of the paper is organized as follows: The next section describes the materials and methods. The material and methods starts by presenting SUST-ARG names dataset

بسم الله الرحمن الرحيم
 جامعة السودان للعلوم والتكنولوجيا
 كلية علوم الحاسوب وتقانة المعلومات
 إستمارة تقديم داخلي لشهادة التخرج

النوع: أنثى ذكر

خاص بالطالب:
 الاسم (رباعي) (حسب الجواز): ميساء عثمان علي محمد نور
 الاسم باللغة الإنجليزية (رباعي): Maisa Osman Ali Mohammed Nur

Fig. 3. Upper part of SUST certificate application form, extracted names are indicated by the red boxes.

followed by describing the preprocessing of the data, and finally describes the proposed model. Section 3, presents and discusses the results. the conclusions and future works are given in Section 4.

II. MATERIAL AND METHODS

A. Dataset: SUST-ARG names

The Dataset used in this paper, SUST-ARG names, is collected by the Arabic Recognition Research Group in the College of Computer Science and Information Technology, Sudan University of Science and Technology. This dataset has been collected in two stages. In the first stage 8028 names written by 2007 students have been extracted from the Certificate application form which used by the registrar office in the college (Fig. 3). From this data the most repeated names have been identified. In the second stage, a new form is designed to collect data for Forty common names (Fig. 4). The right column in Figure 4 contains male names and the left for female names. One thousand forms have been filled by university students. The forms were preprocessed and a dataset that contains forty thousand names is constructed. This dataset contains two parts, twenty thousand male names dataset and twenty thousand female names. Only male names are used in the experiments of this paper. Table I, contains male names samples.

B. Data Preprocessing

The data collection forms were digitized using a scanner and then the names are extracted from the scanned paper. Most of the names are not in the center of their boxes as depicted in Table I. The initial experiments show that raw data is not suitable as input to the proposed CNN. The extracted images are then preprocessed by eliminating the surrounding white spaces and down scaling the image to 28 x 56 pixels. After that all the images are converted to black Background and white foreground.

فضلاً قم بملأ الأسماء بالاسماء

سارة	سارة	ابراهيم	ابراهيم
ايمن	ايمن	احمد	احمد
مروة	مروة	ادريس	ادريس
هبة	هبة	ادم	ادم
فاطمة	فاطمة	اسماعيل	اسماعيل
سلمى	سلمى	بابكر	بابكر
منى	منى	بشير	بشير
آلاء	آلاء	جعفر	جعفر
امراء	امراء	حامد	حامد
سحر	سحر	حسن	حسن
سماح	سماح	حسين	حسين
صفاء	صفاء	خالد	خالد
شيماء	شيماء	سعيد	سعيد
وفاء	وفاء	سليمان	سليمان
ايناس	ايناس	صالح	صالح
نقيسة	نقيسة	صديق	صديق
ريان	ريان	طله	طله
زينب	زينب	عباس	عباس
امل	امل	عثمان	عثمان
ابتهال	ابتهال	علي	علي

Fig. 4. Second stage data collection form sample.

TABLE I. MALE NAMES SAMPLES.

#	English name	Arabic name	Samples
1.	Ibrahim	إبراهيم	ابراهيم ابراهيم
2.	Ahmed	احمد	احمد احمد
3.	Idris	إدريس	ادريس ادريس
4.	Adam	آدم	ادم ادم
5.	Ismail	إسماعيل	اسماعيل اسماعيل
6.	Babiker	بابكر	بابكر بابكر
7.	Bashir	بشير	بشير بشير
8.	Gafar	جعفر	جعفر جعفر
9.	Hamid	حامد	حامد حامد
10.	Hassan	حسن	حسن حسن
11.	Hussain	حسين	حسين حسين
12.	Khalid	خالد	خالد خالد
13.	Saeed	سعيد	سعيد سعيد
14.	Soliman	سليمان	سليمان سليمان
15.	Salih	صالح	صالح صالح
16.	Siddiq	صديق	صديق صديق
17.	Taha	طله	طله تله
18.	Abass	عباس	عباس عباس
19.	Osman	عثمان	عثمان عثمان
20.	Ali	علي	علي علي

C. The Proposed CNN Architecture

Convolutional Neural Networks (CNNs) is one of the important classes of deep learning, which is mostly applied in computer vision to analyze and identify visual imagery [20]. They are found to be very efficient in computer vision problem and they have state of the art performance when provided with large amount of training data [18]. Normally, a CNN architecture consists of a sequence of layers that are stacked in specific organization where the output of each layer is used as an input for the layer that comes after it. The basic layers that can be found in any CNN architecture are: convolutional layer, pooling layer, and the fully connected or dense layer. The most important layer in the CNN architecture is the convolutional layer and normally it is the first layer (the input layer). It uses what is known as convolutional filtering to detect the presence of features that can be used to distinguish specific image that is provided as an input by creating features map. These feature map records the exact position of features in the image, where small movements in this position will result in a totally different feature map. Down sampling is one of the approaches to address the problem that can be created by the features map, and it is implemented in the CNN architecture by the means of a max pooling layer, which scales the number of features generated by the convolutional layer and keep the most important features. There are two important functions that can be used in the pooling layer: Max pooling and Average pooling, where the former takes the large value in each patch of features i.e. the large value in a set of pixels and the later calculates the average of each patch. Because the distribution of each layer input changes with the change of its previous layer parameters during the training process, a batch normalization

is needed to normalize layer inputs and hence speed up the learning process [7]. The layer that is used to classify the input images is known as the fully connected layer and normally it is the last layer in the CNN architecture. It produces outputs equivalent to the number of the classes in the classification problem. These outputs are probabilities for the class's labels. Activation functions are essential part of the CNN architecture to accomplish the complex functional mapping between the input and the outputs, which can be used as inputs for the next layer. The proposed CNN architecture model for Arabic names recognition uses four convolutional layers and each layer uses Rectified linear unit (ReLU) activation function, which is very popular activation function. According to the literature, it has very good experimental results. Simply it is defined as shown in equation 1.

$$f(x) = \max(0, x) \tag{1}$$

The function will return zero if the value is negative and the value itself if it is positive, which means that it grows linearly for positive values. Because Max pooling layer can select the brighter pixels from the inputs and it suits the case where the background is dark i.e. the same as the preprocessed images in our dataset, we used it after the second and the fourth convolutional layers. Our proposed model uses two dense layers and it uses softmax activation function after the second Dense layer. The softmax function forces the output of each unit to be between 0 and 1. It divides each output such that the total sum of the outputs is equal to 1 and mathematically can be expressed as shown in equation 2.

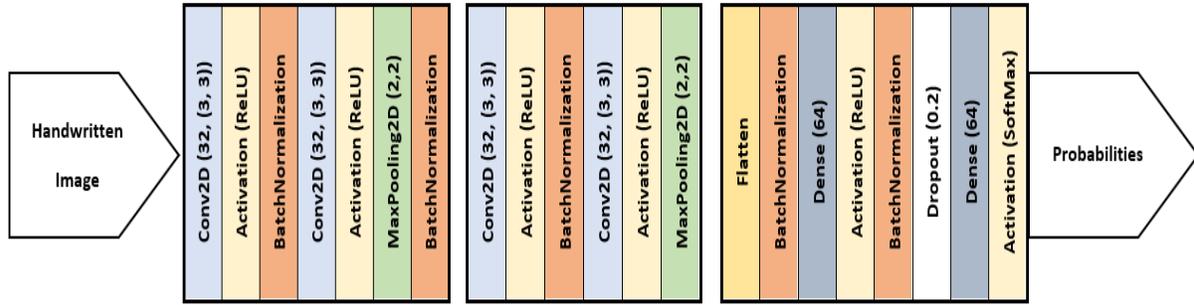


Fig. 5. The architecture of the proposed model

$$\sigma(Z)_j = \frac{e^{z_j}}{\sum_{k=1}^K (e^{z_k})} \quad (2)$$

where z is a vector of the inputs. j is an index for the output units and it goes from 1 to K. The complete architecture of our proposed model is depicted in Fig. 5.

III. RESULTS AND DISCUSSION

We used keras library under TensorFlow environment to construct the proposed CNN model. Ten fold cross validation is used to train and test the proposed model, where the entire dataset is divided into 10 folds. One of these 10 folds is used as a testing set and the remaining folds are used for training the proposed model. This process is repeated 10 times by taking a unique fold in each time. The final results is calculated as the average of the results that are obtained from the repetition. The accuracy of prediction is used as performance measure. The accuracy obtained for the testing names is 99.14%. Table II shows some of the missed classified names.

The literature survey reveals that there are two papers which describe recognition models using the same data set [19], [11]. Both of the papers have used only the male names part of the dataset (same as the experiments of this paper). The classifier of the first paper is HMM and the best classification accuracy is 63%. The classifier of the second paper is probabilistic neural networks, where the classification accuracy is 89% (see Table III).

Despite the fact that the classification accuracy of the proposed model is very high, however, one may consider that the small number of classes (only 20 names) is a major drawback. As we have mentioned holistic classifier does not work alone. The classifier will recognize common names only and the system should use the analytical approach to recognize the new names or uncommon ones. Therefore, there is critical trade-off here, increasing the number of names, speeds up the processing but the cost of training is higher. Moreover, increase the number of classes may decrease the accuracy. This is a question for further research.

IV. CONCLUSION

This paper tests the usage of Holistic paradigm to recognize Arabic names without segmentation. As deep learning is

TABLE II. SAMPLE OF MISSED CLASSIFIED NAMES.

#	English name	Arabic name	Sample	Classified as
1	Ibrahim	إبراهيم		Ismail
2	Ahmed	احمد		Hassan
3	Idris	إدريس		Ibrahim
4	Adam	آدم		Idris
5	Ismail	إسماعيل		Osman
6	Babiker	بابكر		Ahmed
7	Bashir	بشير		Hussain
8	Gafar	جعفر		Saeed
9	Hamid	حامد		Khalid
10	Hassan	حسن		Hussain
11	Hussain	حسين		Hassan
12	Khalid	خالد		Hamid
13	Saeed	سعيد		Bashir
14	Soliman	سليمان		Hussain
15	Salih	صالح		Khalid
16	Siddiq	صديق		Saeed
17	Taha	طه		Khalid
18	Abass	عباس		Osman
19	Osman	عثمان		Soliman
20	Ali	علي		Khalid

TABLE III. THIS TABLE SHOWS THE ACCURACY OF THE PROPOSED CLASSIFIER TOGETHER WITH THE ACCURACY OF TWO CLASSIFIERS TRAINED ON THE SAME DATASET.

Classifier	Accuracy
The proposed model	99.14%
Hidden Markov Model [19]	63%
Probabilistic neural network [11]	89%

currently the state of the art for such system, a CNN model is designed, trained and tested using SUST-ARG-names dataset. The goal of the paper is to study the idea of building a system which recognizes common Arabic names very quickly. Although the accuracy is very high, the number of names selected for the experiments of this paper is relatively small (20

names). Comparing the proposed model with previous similar work on the same dataset is the main reason for selecting this small number. According to the statistics on the given dataset, these 20 names represent more than 50% of SUST students' names. It seems clear that increasing the number of names to represent, for instance, more than 90% of the used name is the future work that must be done to support and prove the practicality of this model.

REFERENCES

- [1] Yasser M Alginahi. A survey on arabic character segmentation. *International Journal on Document Analysis and Recognition (IJDAR)*, 16(2):105–126, 2013.
- [2] Chaouki Boufenar and Mohamed Batouche. Investigation on deep learning for off-line handwritten arabic character recognition using theano research platform. In *2017 Intelligent Systems and Computer Vision (ISCV)*, pages 1–6. IEEE, 2017.
- [3] Murtada Khalafallah Elbashir and Mohamed Elhafiz Mustafa. Convolutional neural network model for arabic handwritten characters recognition. *International Journal of Advanced Research in Computer and Communication Engineering*, 7(11), 2018.
- [4] Mohamed Elleuch, Najiba Tagougui, and Monji Kherallah. Arabic handwritten characters recognition using deep belief neural networks. In *2015 IEEE 12th International Multi-Conference on Systems, Signals & Devices (SSD15)*, pages 1–5. IEEE, 2015.
- [5] Hashem Ghaleb, P Nagabhushan, and Umapada Pal. Segmentation of overlapped handwritten arabic sub-words. *International Journal of Computer Applications*, 975:8887, 2015.
- [6] Barbara F Grimes. *Ethnologue: Languages of the world*. 14. Aufl. Dallas, 2000.
- [7] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.
- [8] Ahmed Lawgali. A survey on arabic character recognition. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 8(2):401–426, 2015.
- [9] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.
- [10] Sriganesh Madhvanath and Venu Govindaraju. The role of holistic paradigms in handwritten word recognition. *IEEE transactions on pattern analysis and machine intelligence*, 23(2):149–164, 2001.
- [11] Wafa A. Mohamed and M. E. M. Musa. Recognition of arabic handwritten names using probabilistic neural networks. *Computer Studies Journal*, 1(1), 2009.
- [12] Bodoor Bashir Mohamed E. M. Must and Mohamed N. I. Ismail. Designing an arabic handwritten segmentation system. *International Journal of Computing*.
- [13] Mohamed EM Musa. Towards building standard datasets for arabic recognition. *International Journal of Engineering and Advanced Research Technology (IJEART)*, 2(2), 2016.
- [14] Mohammad Tanvir Parvez and Sabri A Mahmoud. Offline arabic handwritten text recognition: a survey. *ACM Computing Surveys (CSUR)*, 45(2):23, 2013.
- [15] Mohammad Tanvir Parvez and Sabri A Mahmoud. Offline arabic handwritten text recognition: a survey. *ACM Computing Surveys (CSUR)*, 45(2):23, 2013.
- [16] MS Patel and Rohith Kumar. Offline handwritten word recognition: A survey.
- [17] Mario Pechwitz, S Snoussi Maddouri, Volker Märgner, Nouredine Ellouze, Hamid Amiri, et al. Ifn/enit-database of handwritten arabic words. In *Proc. of CIFED*, volume 2, pages 127–136. Citeseer, 2002.
- [18] Waseem Rawat and Zenghui Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, 29(9):2352–2449, 2017.
- [19] Talaat M. Wahbi and Mohamed E. M. Musa. Holistic approach for arabic word recognition. *international Journal of Computer Applications Technology and Research*, 5(3):141–146, 2016.
- [20] Wenzhe Wang, Bin Wu, Sixiong Yang, and Zhixiang Wang. Road damage detection and classification with faster r-cnn. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 5220–5223. IEEE, 2018.
- [21] Khaled S Younis. Arabic handwritten character recognition based on deep convolutional neural networks. *Jordanian Journal of Computers and Information Technology*, 3(3), 2017.
- [22] Oussama Zayene, Jean Hennebert, Rolf Ingold, and Najoua Essoukri BenAmara. Icdar2017 competition on arabic text detection and recognition in multi-resolution video frames. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 1460–1465. IEEE, 2017.
- [23] Erik J Zürcher. *Turkey: A modern history*. Bloomsbury Publishing, 2017.