https://doi.org/10.17576/apjitm-2020-0901-06
Asia-Pacific Journal of Information Technology and Multimedia
Jurnal Teknologi Maklumat dan Multimedia Asia-Pasifik
Vol. 9 No. 1, June 2020: 69 - 81

Received: 12 March 2020
Accepted: 8 May 2020
Published: 15 June 2020

# A REVIEW OF ARABIC TEXT RECOGNITION DATASET

IDRIS SALEH AL-SHEIKH
MASNIZAH MOHD
LIA WARLINA

ABSTRACT

Building a robust Optical Character Recognition (OCR) system for languages, such as Arabic with cursive scripts, has always been challenging. These challenges increase if the text contains diacritics of different sizes for characters and words. Apart from the complexity of the used font, these challenges must be addressed in recognizing the text of the Holy Quran. To solve these challenges, the OCR system would have to undergo different phases. Each problem would have to be addressed using different approaches, thus, researchers are studying these challenges and proposing various solutions. This has motivate this study to review Arabic OCR dataset because the dataset plays a major role in determining the nature of the OCR systems. State-of-the-art approaches in segmentation and recognition are discovered with the implementation of Recurrent Neural Networks (Long Short-Term Memory-LSTM and Gated Recurrent Unit-GRU) with the use of the Connectionist Temporal Classification (CTC). This also includes deep learning model and implementation of GRU in the Arabic domain. This paper has contribute in profiling the Arabic text recognition dataset thus determining the nature of OCR system developed and has identified research direction in building Arabic text recognition dataset.

Keywords: text recognition, dataset, Quranic, OCR, Arabic

## INTRODUCTION

Conventionally, the input of an Optical Character Recognition (OCR) system is a page-image. The page is usually segmented into paragraphs, the paragraphs are segmented into text-lines, the text-lines into words, the words into sub-words, and finally the sub-words into individual characters for the system to be able to convert this image into the equivalent text. A character recognizer would recognize the segmented characters one by one. This method is called the segmentation-based OCR. The second, more recent method is the holistic OCR. In the holistic method, the recognition is performed at a word or a text-line level. This method overcomes the issues of character segmentation. There are two types of OCR methods which are segmentation-based OCR and segmentation-free method.

In the segmentation-based OCR method, the first main step is to detect discrete components, appropriate for the last recognition of the OCR step. Such a method depends heavily on the extraction of the individual characters from the text. It has continued to be the state-of-the-art method for a period of time before a segmentation-free method outperformed the segmentation method. An old version of Tesseract (Smith 2007), which is a popular open-source OCR system, is a good example of a segmentation-based OCR system. A segmentation-based OCR can be classified into two classes, which are template matching and over-segmentation. Template matching is based on connected components, where the characters are extracted and matched upon the possible templates. The recognition is achieved based on some similarity measures. The use of template-matching based methods is quite limited as it greatly suffers from font variations, image noise, and touching characters. On the over-segmentation method, instead of finding a precise segmentation spot between two characters, an approximate

segmentation spot is located and as a result, over-segmentation occurs and is corrected at the next step. The over-segmentation method is very handy with cursive scripts such as the Arabic text where a correct character segmentation is very hard to perform. Ahmed et al. (2007a, 2007b), Jabril et al. (2011, 2016a, 2016b, 2016c) and Atallah & Khairuddin (2009) introduced rules for reconfirming the potential segmentation points of Arabic words using peaks and vertexes points of Voronoi diagrams on the baseline based on peaks detection. Three steps were developed in word and sub-word segmentation approach where a peaks detection function is adopted to model the maximum and minimum peaks. A stroke operator is utilized to extract of potential segmentation points; then a determine baseline process is developed to estimate the parameters depend on the mostly minimum peaks and determine nearest vertexes point to minimum peak on the baseline to confirm the minimum peak as segmentation point. In addition, Jabril et al. (2017) too have applied a novel method to detect correctly of location segmentation points by detect of peaks with neural networks for Arabic word. This method employs baseline and peaks identification; where using two steps to segmenting text. Where peaks identification function is applied which at the sub-word segment level to frame the minimum and maximum peaks, and baseline detection has provide high accuracy.

The difference between segmentation-free and segmentation-based methods is the level of segmentation. Usually, in segmentation-free methods, the text-line is segmented into words or parts of a word. Then, the recognition engine tries to recognize the whole word or parts of the word. In the segmentation-free or sometimes called the holistic approaches, discriminating features are extracted from the word or parts of the word. Then, the HMM or ANN classifier are trained on the extracted features to recognize the word or the sub-word. A recent approach in the segmentation-free domain is to use RNN with CTC. This method is called Sequence Learning, where the classifier has to map the input sequence to the output sequence. In this study we focused on RNN and the CTC.

## RECURRENT NEURAL NETWORKS

The human mind usually analyzes information according to past experience and depending on the current context, traditional neural networks are not context-dependent. Meanwhile Recurrent Neural networks are designed to take advantage of the context information. RNN are connecting units which allow information to be passed from one to another as shown in Figure 1.

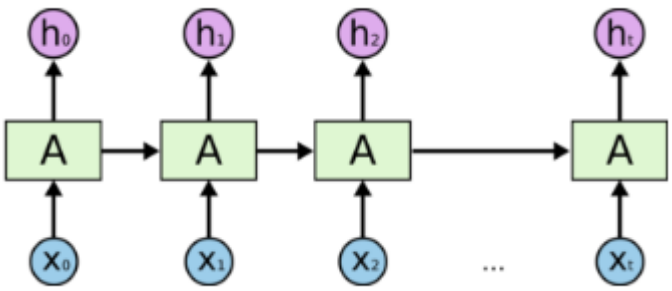



FIGURE 1. Recurrent Neural Network

The RNN is capable of remembering the context of a sequence due to the feedback connections between the hidden layers. However, in practice, it is not capable of remembering very long context due to the Exploding Gradient Problem and Vanishing Gradient Problem, when the training uses gradient descent-based learning, the error signal is propagated back to update the internal weight connections. The values of the first-order gradient values either grow

exponentially, the reason for it to be called the Exploding Gradient; or they vanish to zero exponentially, which is called the Vanishing Gradient, which makes the RNN learning very slow and unusable. In order to solve the Exploding Gradient Problem and Vanishing Gradient problem (Hochreiter & Schmidhuber 1997) used memory cells to replace the activation units at the hidden layer, which is called the Long Short-Term Memory (LSTM) as shown in Figure 2.

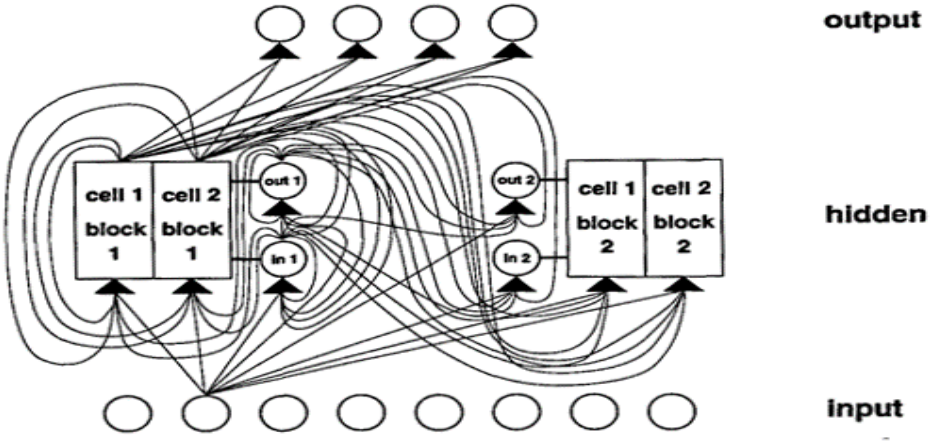



FIGURE 2. Example of a net with eight input units, four output units, and two memory cell blocks of size 2 (Hochreiter & Schmidhuber 1997)

Another RNN introduced by Cho et al. (2014) called Gated recurrent unit (GRU) also aims to solve the vanishing gradient problem, GRU is basically an LSTM without an output gate, which therefore fully writes the contents from its memory cell to the larger net at each time step as shown in Figure 3.

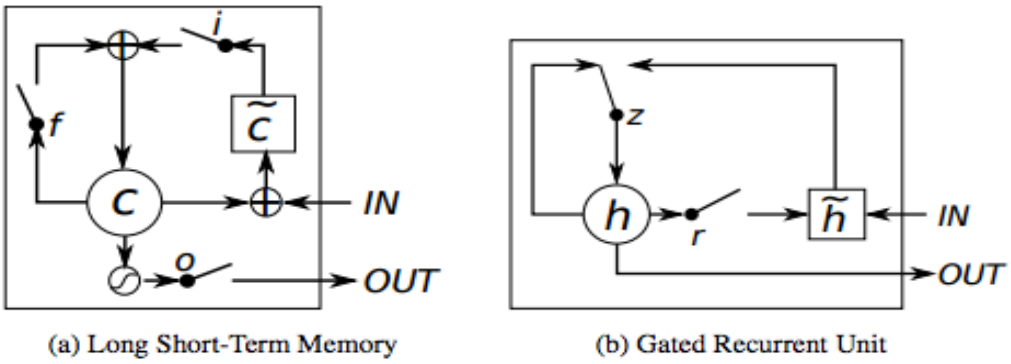



FIGURE 3. Illustrate of (a) LSTM (b) GRU

LSTM has three gates *i*, *f* and *o* are the input, forget, and output gates, respectively, *c* and *c*~ are the memory cell and the new memory cell content. GRU has two gates *r* and *z* are the rest and update gates, and *h* and *h*~ are the activation or the candidate activation. In order to map the input sequence to the output sequence, a specialized algorithm called CTC has been used which is comparable to HMM's forward backward methods.

CONNECTIONIST TEMPORAL CLASSIFICATION

Connectionist Temporal Classification (CTC) was introduced by Graves et al. (2006) to Labelling Unsegmented Sequence Data with Recurrent Neural Networks, usually sequence classification will required pre-segmented training data, and post-processing to transform their outputs into label sequences, the CTC solve those two problems it will map the sequence of input to the sequence of the output using the CTC loss function and CTC decoder transforms the NN output into the final text.

## ARABIC OPTICAL CHARACTER RECOGNITION DATASET

The OCR system requires a dataset for training and for the system to learn how to recognize the text within the image, and then convert that image into digital text. Due to the lack of a standard benchmark, most of the studies in this field were conducted using private datasets without a fair comparison. Hence, although most work would showcase high accuracy results, they may not be up to scale for a large set of problems. Therefore, an extensive list of publicly available datasets is offered in this subsection.

IFN/ENIT (Pechwitz et al. 2002) is an Arabic handwritten word dataset that contains 26,459 handwritten Tunisian town names, which were written by 411 different writers. This dataset is available to the public for research purposes.

The Arabic Handwritten Database (AHDB) (Al-Ma'adeed et al. 2002) contains the most popular Arabic words, numerals, and entities used in cheques, and written by 100 different writers.

The Arabic Cheque Database (Al-Ohali et al. 2003) is a handwritten cheque for legal and courtesy amount recognition database, which contains 29,498 sub-words and 15,175 digits in the form of Indo-Arabic numerals, and 2,499 legal and courtesy amount words extracted from 3,000 checks.

The Handwritten Arabic Character Database (Asiri & Khorsheed 2005) contains 15,800 isolated handwritten Arabic character images, written by approximately 500 Saudi Arabian secondary school students of both genders. The hand-written pages were scanned at 300 dpi, and each character image was saved as 7×7 grey-scaled image. However, this dataset is unavailable to the public.

The Handwritten Arabic Digit Database (Awaidah & Mahmoud 2009) contains 21,120 scanned samples of digits written by 44 different writers. Each writer wrote the digit from 0 to 9 for 48 times in an Indian format. The images were saved with a resolution of 300 pixels, which were then converted to the binary format. To segment the scanned pages into lines, the Horizontal Histogram was used. Then, a Vertical Histogram was used to segment each line into digits. This dataset is available online for researchers.

The Database for Handwritten Arabic Characters (HACDB) (Lawgali et al. 2013) was developed to cover all shapes of the Arabic characters, including overlapping characters. This dataset contains 6,600 characters written by 50 writers ranging between 14 to 50 years old. This database is available publicly for research purposes.

The UPTI database (Sabbour & Shafait 2013) contains images which are synthetically generated using the Nastaleeq font for the Urdu Printed Text. This database consists of 10,063 images of the Urdu text lines, which consists of both ligature and line versions. This dataset is suitable for training deep learning models, and page segmentation. To segment the images into lines and words, the Baseline Estimation was used by calculating the maximum horizontal projection, then using connected components. This dataset is available publicly for research purposes.

Mohd Sanusi Azmi (2013) introduced a novel feature from combinations of triangle geometry for digital Jawi paleography. A dataset of 69.400 images of Arabic calligraphy characters was built consisting of the handwriting of ten calligraphy experts.

KHATT (Mahmoud et al. 2014) is an open Arabic offline handwritten text database. It has 2,000 unique paragraph images with 9,000 line images. Written by 1,000 different writers, who came from different countries with different qualifications, age, gender, and left or right-handedness. The images are stored in different resolutions of 200, 300, and 600 dpi. The dataset can be used for different research purposes other than handwriting recognition, such as line segmentation, noise removal techniques, binarization, and writer identification. This dataset is divided into 70%, 15%, and 15% for training, validation, and testing, respectively. This dataset is available publicly for research purposes.

The KAFD dataset (Luqman et al. 2014) is an Arabic font database at page-level and text-line level. It consists of 40 fonts with 10 sizes in three resolutions at 100 dpi, 200 dpi, and 300 dpi. KAFD dataset contains 2,576,024 line-images. This dataset is available publicly for research purposes.

The ALIF Dataset (Yousfi et al. 2015a) is a dataset for Arabic embedded text recognition in videos frames. It consists of 6,532 cropped text line images from 8 popular Arabic News channels. This dataset is divided into the ALIF Train of 4,152 text images, the ALIF Test1 that is composed of 900 text images, ALIF Test2 that is composed of 1,299 text images, and ALIF Test3 that is composed of 1,022 text images for benchmark purpose. This dataset can be obtained upon request.

The ACTIV Dataset (Zayene et al. 2015) is a public dataset, which was extracted from 80 videos (more than 850,000 frames) collected from 4 different Arabic news channels. It consists of 4,824 text lines with 21,520 words. This dataset is publicly available for research purposes.

SmartATID (Chabchoub et al. 2016) or the Smartphone Arabic Text Images Database contains both printed and hand-written images captured by mobile devices. The printed version contains 16,472 document images, while the hand-written version contains 9,088 document images. Both sets were captured using two types of mobile phones, namely, Samsung Galaxy S6 edge and iPhone 6S plus. Different parameters were used, such as camera version, light conditions, and position. This dataset is available publicly for research purposes.

Alaa et al. (2017) propose a database for degraded Arabic historical manuscripts dating to the Islamic and ancient Arabic eras. The documents in the database exhibit different types of degradation such as smears, uneven illumination, contrast variation, blur, deteriorated paper, bleed-through, faded ink or faint characters, and thin or weak text.

The Printed PAW Dataset (Bataineh 2017) introduces a database for printed sub-words or Part of Arabic Word (PAWs). The proposed database consists of 415,280 images with 83,056 unique PAWs, which can construct approximately 550,000 different words. This database will be available to the researchers upon request.

The ACTIV 2.0 Dataset (Zayene et al. 2018a) is a public dataset that was extracted from 189 video clips, and produces 4,063 key-frames for detection and 10,415 cropped text images for recognition. This dataset is distributed with open-source tools for annotation and evaluation.

The Quran Text Image Dataset (QTID) (Badry et al. 2018) is the first Arabic dataset that includes Arabic marks (diacritics). It consists of 309,720-word images with a dimension of 192×64. It is synthetically generated from the Quranic words with font sizes of 22, 24, 26, and 28 pixels.

Jabril et al. (2013) introduces, an database (AHDB/FTR) comprising Arabic Handwritten Text Images, which helps the researches associated with recognition of Arabic handwritten text with open vocabulary, word segmentation and writer identification and can be freely accessed by researchers worldwide. This database consists of four hundred and ninety seven images of Libyan cities, which were hand written by five Arabic scholars.

TABLE 1. Datasets on Arabic text recognition

| Dataset | Type of content | Reference | Type | Availability |
|---|---|---|---|---|
| IFN/ENIT | Word | Pechwitz et al. (2002) | Handwritten | Public |
| AHDB | Word and number | Al-Ma'adeed et al. (2002) | Handwritten | Private |
| Arabic Cheque | Sub-words and digits | Al-Ohali et al. (2003) | Handwritten | Private |
| Handwritten Arabic Digit | Digits | Awaidah & Mahmoud (2009) | Handwritten | Public |
| Handwritten Arabic Character | Characters | Asiri & Khorsheed (2005) | Handwritten | Private |
| HACDB | Characters | Lawgali et al. (2013) | Handwritten | Public |
| UPTI | Text lines | Sabbour & Shafait (2013) | Printed | Public |
| Digital Jawi paleography | Images | Mohd Sanusi Azmi (2013) | Jawi paleography | Public |
| KHATT | Text lines | Mahmoud et al. (2014) | Handwritten | Public |
| ALIF | Text line | Yousfi et al. (2015a) | Embedded text | Upon request |
| ACTIV | Text line | Zayene et al. (2015) | Embedded text | Public |
| SmartATID | Page | Chabchoub et al. (2016) | Printed and Hand-written | Public |
| Degraded Arabic historical manuscripts | Image document | Alaa et al. (2017) | Arabic Handwritten Word | Public |
| Printed PAW | Sub-words | Bataineh (2017) | Printed | Upon request |
| ACTIV 2 | Words | Zayene et al. (2018a) | Embedded text | Public |
| QTID | Words | Badry et al. (2018) | Synthetically | Private |
| KAFD | Page and line | Luqman et al. (2014) | Printed | Public |
| AHDB/FTR | Arabic Handwritten Text Images | Jabril et al. (2013) | Handwritten | Public |

Table 1 has clearly shown that more complex tasks, such as recognizing diacritical image texts (for example, the Quranic text) at the word or line level, have not received much attention. Only the QTID dataset deals with the Quranic text, yet, this dataset is not available to researchers. In addition, it is synthetically generated on the word level. This study proposes a dataset based on a printed version of the Holy Quran, on a page and line level. It is also easier to achieve the word level from the line level by applying the vertical histogram projection. Furthermore, this dataset is meant to be publicly available. To the best of our knowledge, there are no publicly available datasets for a diacritical line dataset or Quranic image dataset for text recognition purposes.

## ARABIC TEXT RECOGNITION WITH DEEP LEARNING

Heryanto et al. (2018) proposed Deep Learning approach and using Convolutional Network as learning features to optimize the data representation through end-to-end training of the parameters from raw input data to target class. A multi-classifier implicitly segments the sub-word into sequences of characters where the classifiers consists of one sub-word length classifier and seven character classifiers. This approach is superior to state-of-the-art methods of Jawi handwriting recognition.

Yousfi (2016) presented an Arabic video text recognition system based on the deep learning approach. The proposed model used the input image without any pre-processing or

segmentation. Multi-scaled window-based scanning scheme and deep neural models were applied to extract feature vectors from the input image. The Deep Belief Networks and Multi-Layer Perceptron were used as deep auto-encoders and one with the convolutional neural network. Next, the feature vectors were send to the BLSTM network to learn the sequence labeling, followed by CTC output layers with softmax activation function. A subset of the ALIF dataset was used in this work to train the model with 7,000 text images, to validate the system with 673 text images, and for testing 900 text images. The author compared two approaches to extract features (learned features vs. hand-crafted ones). It was reported that convolutional neural network outperformed the hand-crafted approaches. To show the strength of this model, a comparative study was performed, with 'ABBYY Fine Reader 12'4. This system outperformed the commercial software by almost 11 points in terms of CRR.

Graves (2012) won at the ICDAR 2009 on the Arabic offline handwriting recognition competition. This work was based on the MDLSTM recurrent neural networks. Raw pixel data is used as input and CTC as output. The dataset used for training and test is the IFN/ENIT.

Rashid et al (2013) described a low resolution, multi-font, and open vocabulary system for printed Arabic text. The system is based on MDLSTM and recurrent neural network architecture with CTC layer. The proposed method was trained and evaluated using the APTI database. They reported a result of 99% word recognition rate.

A study by Morillot et al. (2013) was presented by University of Balamand (Lebanon) and Telecom ParisTech (France) for the OpenHaRT 2013 competition. They implemented a system based on BLSTM for the text-line recognition task. The recognition rate of 52% was obtained using a single BLSTM recognizer trained on only 11% of the available NIST/OpenHaRT data (145,000 text-lines).

Chherawala et al. (2013) compared handcraft features and automatic features using the IFN/ENIT dataset. The features used were the concavity features (CCV) for Arabic word image, the distribution features by Rath and Manmatha (R-M) for handwritten word spotting in historical manuscript, and the (M-B) by Marti and Bunke for handwritten text recognition, with HMM, SIFT, Local Gradient Histogram (LGH) features, and automatically learned features by the MDLSTM. The results showed that although the MDLSTM is capable of learning features, the handcraft features had achieved better results.

Pham et al. (2014) reported that a dropout on the first layer can reduce the CER and WER by 10% to 20%, and if the dropout is applied to MDLSTM, the error can be reduced by 30% to 40%. The system was evaluated with three datasets in three languages: RIMES dataset for the French language with character accuracy of 91.1%; IAM dataset for the English language with character accuracy of 85.6%; and OpenHaRT dataset for Arabic handwritten recognition with character accuracy of 90.1%.

Hamdani et al. (2014) used the Hidden Markov Models (HMM) for sequence modeling and the BLSTM for feature extractions was used to train the HMM. The Minimum Phone Error (MPE) discriminative training was used to enhance the training. They used the OpenHaRT dataset, and implemented the n-gram language model, which was pre-smoothed using the Modified Kneser-Ney method.

Yousefi et al. (2015) performed a similar experiment as Chherawala et al. (2013). However, in this experiment, they showed that LSTM, which was faster to learn and converge compared to MDLSTM, had also achieved better results in the same IFN/ENIT dataset, with the same handcraft features, namely, CCV, RM, MB, LGH. The LSTM had automatically extracted features from the row images, and this result was obtained by applying a normalization scheme to the input to reduce the translation to a horizontal axis. They had also showed that the LSTM with automatic features had obtained a better result compared to the handcraft features.

Ahmad et al. (2017) proposed a system based on MDLSTM for Arabic character recognition, with CTC layer as the output. A preprocessing technique was introduced, which would remove extra white spaces and de-skews the text-lines for precise height normalization. This system was able to improve the recognition rate by 29% and the accuracy rate was 75.8% CER on text-lines of KHATT dataset.

Nashwan et al. (2017) proposed a holistic Arabic OCR approach that is computationally efficient. To reduce the word recognition time, they used a lexicon reduction technique by clustering similar shaped words features. This approach consisted of two modules training and a recognition module. To train the extracted holistic features, they extracted hybrid features, with a combination between global word level-based Discrete Cosine Transform and local block-based features. Then, they used clusters based on similar word shapes, and these clusters were subsequently used on the recognition module. After preprocessing the input image to extract the lines and words, the features were extracted for each word image. Then, the model would try to get the best n-clusters that have the minimum Euclidean distance with the test image vector. As a result, a word list from the selected cluster was used to construct a word matrix for possible recognition hypotheses of the whole line. This word matrix was rescored using the language model based on the 4-gram model to achieve the best recognition hypothesis. Different sets were used to test the proposed system; the first set contained 1,152 words, with three different fonts and four font sizes, and achieved 99.3% of WRR. The second set contained 2,730 words of recent computerized book's text and achieved 84.8% of WRR. The third set of old non-computerized books consisted of 2,276 words with not well-known fonts achieved. These results have been compared with Sakhr, ABBYY, and NovoDynamics, which are known commercial Arabic OCR systems, and the results were promising.

Zayene et al. (2018b) presented an Arabic video embedded text recognition system based on deep learning approach, they used MDLSTM network as input layers, so the MDLSTM learn the features from the raw input image, for the output layer they use the CTC with softmax activation function. The suggested method has been trained and evaluated using the AcTiV- R database which is part of AcTiv dataset consists of 10,415 text-lines images, 44,583 words. They report 96.5% as a character recognition rate. Also, they report that their system outperformed the previous work on the ALIF dataset, more particularly those based on the combination of CNN and BLSTM on (Yousfi et al. 2015a, 2015b).

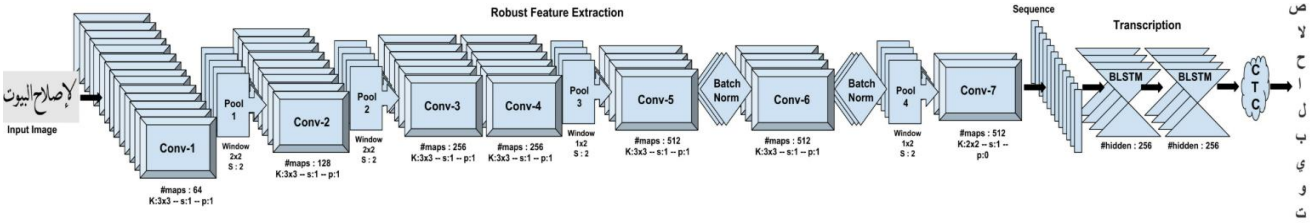
Rahal et al. (2018) proposed the holistic text recognition system, which was based on statistical features. They adopted the Bag of Features (BoF) model, using Sparse Auto-Encoder (SAE) for feature representation and for the recognition process, the Hidden Markov Model (HMM) was used. As a preprocessing step, the Gaussian smoothing was used to reduce the noise normally associated with text images and image re-scaling to obtain a standardized height for all images. This system was evaluated in an experiment with three datasets, namely, KHATT, APTI, and MNIST. The obtained average accuracies of recognition had varied between 99.65% and 99.96% for the mono-font and exceeded 99% for the mixed-font.

Jain (2018) introduced an end-to-end system using a combination of CNN and RNN architecture and showed the superiority of using the hybrid CNN and RNN over a system which additionally depends only on RNN. This method is reported as having outperformed the previous methods on the existing benchmarks. Figure 4 shows the visualization of the hybrid CNN-RNN architecture with a 7-layer-deep convolutional block.

Suvarnam & Ch (2019) use combination of CNN-GRU Model to Recognize Characters of a License Plate number without Segmentation, CNN was used for feature extraction and GRU was used for sequencing without using any segmentation methods the testing precision of the proposed framework is 100% and training accuracy is 90%.

Jiang et al. (2018) use End-to-End Learning OCR Technologies to solve the CAPTCHA problem , the use two pipelines to solve this arithmetic operation, the deep

convolutional neural network (DCNN) with parallel dense layers and component-connection-based detection, the second use the convolutional recurrent neural network (CRNN) with connectionist temporal classification was adopted, combined with the text region detection technique to recognize more complex pictures with both assignment operations and calculation formulas, which achieved 98.08% accuracy.




| | | | | | |
|---|---|---|---|---|---|
| Graves (2012) | Handwritten Arabic | MDLSTM | Pixels | IFN/ENIT | 93.37 |
| Rashid et al. (2013) | Printed Arabic | MDLSTM | Pixels | APTI | 99.0 RR |
| Morillot et al. (2013) | Handwritten Arabic | BLSTM | Features | OpenHaRT/NIST | 52.0 Word |
| Pham et al. (2014) | Handwritten French English Arabic | MDLSTM | Pixels | RIMES- IAM – OpenHaRT | 91.10<br>85.60<br>90.10 |
| Chherawala et al. (2013) | Handwritten Arabic | MDLSTM | Pixels | IFN/ENIT | 89.0 RR |
| Yousefi et al. (2015) | Handwritten Arabic | BLSTM | Pixels | IFN/ENIT | 87.4 |
| Hamdani et al. (2014) | Handwritten Arabic | BLSTM | Pixels | OpenHaRT (LMs) | 80.1 WRR<br>94.1 CRR |
| Ahmad et al. (2017) | Handwritten Arabic | MDLSTM | Pixels | KHATT | 75.8 CRR |
| Zayene et al. (2018b) | Embedded Arabic | MDLSTM | Pixels | AcTiV | 96.5 CRR |
| Yousfi (2016) | Embedded Arabic | BLSTM | CNN | ALIF | 90.71 CRR |
| Nashwan et al. (2017) | Printed Arabic | Cluster with LBG | Discrete Cosine Transform and local block LM 4gram | 1152 W from newspaper<br>2730 W from book<br>2276 W from an old book | 99.30 WRR<br>84.80 WRR<br>53.0 WRR |
| Rahal et al. (2018) | Handwritten Arabic | HMM | BoF with SAE | KHATT, APTI, MNIST | 99.0 CRR |
| Jain (2018) | Printed Aurdo | BLSTM | CNN | UPTI, IIIT-Urdu | 98.80 - 89.84 CRR |
| Suvarnam & Ch (2019) | license plate | GRU | CNN | Unknown license plate | 100 |
| Jiang et al. (2018) | CAPTCHA-style | BGRU | CNN | Simple Arithmetic Operation | 98.08 |

Table 2 clearly shows that the RNN was able to become the state-of-the-art system in the text recognition domain. MDLSTM or BLSTM can be used to obtain good results, with benefits from using every one of them. The LSTM was faster at learning and converging than MDLSTM. GRU, on the other hand, did not used with the Arabic text recognition yet, some researcher implement GRU on OCR for license plate and CAPTCHA, but many researchers use GRU in the Arabic domain for different task like speech recognition (Zerari et al. 2019),

Arabic Neural Machine Translation (Almahairi et al. 2016), Arabic Named Entity (Gridach & Haddad 2017), and Arabic discretization (Moumen et al. 2018).

It was concluded that complex tasks, such as recognizing diacritical image texts (Quranic text) at word or line level has not received much attention and this could lead for future research directions in this area.

## DISCUSSION

We discovered that complex tasks, such as recognizing diacritical image texts (Quranic text) at the word or line level in Arabic OCR have not received much attention. This can be the future work or research direction in preparing Arabic dataset. Furthermore to the best of our knowledge, there are no publicly available datasets for a diacritical line dataset or Quranic image dataset for text recognition purposes. Only QTID dataset deals with the Quranic text, however it is not available to researchers. In addition, it is synthetically generated at word level.

This work has described Arabic OCR dataset with various types of data such as handwritten text, the printed text and the embedded text. The presented dataset has tremendous potential in fully automated OCR using machine learning and deep learning approaches. We discovered that RNN was able to become the state-of-the-art system in the text recognition domain. Comparison of performance revealed that LSTM was faster at learning and converging compared to MDLSTM. Few researchers have implemented GRU in OCR for license plate and CAPTCHA, but many researchers use GRU in the Arabic domain for different task like speech recognition and in Arabic neural machine translation, Arabic named entity recognition and Arabic discretization. Meanwhile for complex tasks, such as recognizing diacritical image texts (Quranic text) at word or line level has not received much attention.

## SUMMARY

We have highlighted the different approaches of the Arabic OCR system. The discussion is made on the different types of OCR dataset, which reflects the type of OCR system such as the handwritten text, the printed text and the embedded text. The discussion is also made on the different types of segmentation such as the character, sub-words, word, text line and paragraph segmentation. The description of the general techniques used for feature extraction has also been reviewed. Apart from that, the techniques and architecture of recognition such as the MDLSTM, BLSTM, CNN, HMM have also been explained in detail in this study. Finally, the review of the recent related studies in the area of Arabic OCR system has also been discussed.

## ACKNOWLEDGMENTS

This study is supported by the Universiti Kebangsaan Malaysia grant: GGP-2017-022.

## REFERENCES

Ahmad, R., Naz, S., Afzal, M. Z., Rashid, S. F., Liwicki, M. & Dengel, A. 2017. KHATT: A Deep Learning Benchmark on Arabic Script. 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), pp. 10–14. IEEE.

Ahmed M. Zeki, Mohamad S. Zakaria & Choong-Yuen Liong, 2007. Isolation of Dots for Arabic OCR using Voronoi Diagrams, Proceedings of the International Conference on Electrical Engineering and Informatics, Bandung, Indonesia, June 2007. pp.199-202.

Ahmed M. Zeki, Mohamad S. Zakaria & Choong-Yuen Liong, 2007. The Use of Area-Voronoi Diagram in Separating Arabic Text Connected Components, Third International Workshop on Advances Computation for Engineering Applications (ACEA07), Al-Zaytoonah University, Amman, Jordan, 9-11 May 2007. pp. 251-288.

Alaa Sulaiman, Khairuddin Omar, & Mohammad Faidzul Nasrudin, 2018. A Database for Degraded Arabic Historical Manuscripts. The 2017 6th International Conference on Electrical Engineering and Informatics (ICEEI) 25 - 27 November, 2017. pp. 1-6. ISBN: 978-153860475-5 Electronic ISSN: 2155-6830 (Date added to IEEE Xplore: 12 March 2018) DOI: 10.1109/ICEEI.2017.8312375.

Al-Ma'adeed, S., Elliman, D. & Higgins, C. A. 2002. A data base for Arabic handwritten text recognition research. Frontiers in Handwriting Recognition, 2002. Proceedings. Eighth International Workshop on, pp. 485–489. IEEE.

Almahairi, A., Cho, K., Habash, N. & Courville, A. 2016. First result on Arabic neural machine translation. *arXiv preprint arXiv:1606.02680*.

Al-Ohali, Y., Cheriet, M. & Suen, C. 2003. Databases for recognition of handwritten Arabic cheques. Pattern Recognition 36(1): 111–121.

Asiri, A. & Khorsheed, M. S. 2005. Automatic Processing of Handwritten Arabic Forms using Neural Networks. IEC (Prague), pp. 313–317.

Atallah al-Shatnawi & Khairuddin Omar. 2009. Detecting Arabic Handwritten Word Baseline Using Voronoi Diagram. International Conference on Electrical Engineering and Informatics 2009. Vol I. pp: 18-22.

Awaidah, S. M. & Mahmoud, S. A. 2009. A multiple feature/resolution scheme to Arabic (Indian) numerals recognition using hidden Markov models. Signal Processing 89(6): 1176–1184.

Badry, M., Hassan, H., Bayomi, H. & Oakasha, H. 2018. QTID: Quran Text Image Dataset. *International Journal Of Advanced Computer Science And Applications* 9(3): 385–391.

Bataineh, B. 2017. A Printed PAW Image Database of Arabic Language for Document Analysis and Recognition. *Journal of ICT Research and Applications* 11(2): 200–212.

Chabchoub, F., Kessentini, Y., Kanoun, S., Eglin, V. & Lebourgeois, F. 2016. SmartATID: A mobile captured Arabic Text Images Dataset for multi-purpose recognition tasks. *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pp. 120–125. IEEE.

Chherawala Y, Roy PP, & Cheriet M. 2013. Feature design for offline arabic handwriting recognition: Handcrafted vs automated? *Proceedings of the International Conference On Document Analysis and Recognition, ICDAR*. pp. 290-294. DOI: 10.1109/ICDAR.2013.65.

Cho et al. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1724–1734, October 25-29, 2014, Doha, Qatar.

Graves, A. 2012. Offline arabic handwriting recognition with multidimensional recurrent neural networks. *Guide to OCR for Arabic scripts*, pp. 297–313. Springer.

Graves, A., Wayne, G., Reynolds, M. et al. 2016. Hybrid computing using a neural network with dynamic external memory. Nature 538, 471–476. https://doi.org/10.1038/nature20101

Gridach, M. & Haddad, H. 2017. Arabic named entity recognition: A bidirectional GRU-CRF approach. *International Conference on Computational Linguistics and Intelligent Text Processing*, pp. 264–275. Springer.

Hamdani, M., Doetsch, P., Kozielski, M., Mousa, A. E.-D. & Ney, H. 2014. The RWTH large vocabulary Arabic handwriting recognition system. *Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on*, pp. 111–115. IEEE.

Hochreiter, S. and Schmidhuber, J. 1997. LSTM can solve hard long time lag problems. In Advances in Neural Information Processing Systems. MIT Press, Cambridge MA. Presented at NIPS 96.

Jabril Ramdan & Khairuddin Omar. 2011. Comparative Study of Algorithms for Voronoi Diagram Construction on Segmentation of Arabic Hand Writing. Australian Journal of Basic and Applied Sciences, 5(11): 1653-1667. ISSN 1991-8178.

Jabril Ramdan, K. Omar, M. Faidzul. 2016. A New Rule to Reconfirm Potential Segmentation Points with Vertexes Points of VDS. Middle-East J. Sci. Res., 24 (3): 657-662, 2016. ISSN: 1990-9233 DOI: 10.5829/idosi.mejsr.2016.24.03.23185

Jabril Ramdan, Khairuddin Omar and Mohammad Faidzul. 2016. Determine Characters by Mathematical Model for Segmentation Arabic Words by Voronoi Diagrams. Indian Journal of Science and Technology, 9(40): 1-7, October 2016. ISSN (Print): 0974-6846 ISSN (Online): 0974-5645. DOI: 10.17485/ijst/2016/v9i40/84801

Jabril Ramdan, Khairuddin Omar and Mohammad Faidzul. 2016. Segmentation of Arabic Words Using Area Voronoi Diagrams and Neighbours Graph. International Journal of Soft Computing, 11(5): 282-288. ISSN : 1816-9503 (Print) DOI: 10.3923/ijscomp.2016.282.288.

Jabril Ramdan, Khairuddin Omar, and Mohammad Faidzul. 2017. A Novel method to detect segmentation points of Arabic words using peaks and neural network. International Journal on Advanced Science, Engineering and Information Technology, 7(2): 625-631, DOI:10.18517/ijaseit.7.2.1824. ISSN: 2088-5334.

Jabril Ramdan, Khairuddin Omar, Mohammad Faidzul, Ali Mady. 2013. Arabic Handwriting Data Base for Text Recognition. The 4th International Conference on Electrical Engineering and Informatics (ICEEI 2013). Procedia Technology 11 (2013) 580 – 584. ISSN: 2212-0173 doi: 10.1016/j.protcy.2013.12.231

Jain, M. 2018. Unconstrained Arabic & Urdu Text Recognition using Deep CNN-RNN Hybrid Networks. International Institute of Information Technology Hyderabad.

Jiang et al. 2018. OCR with a convolutional neural networks integration model in machine vision. Tenth International Conference on Digital Image Processing (ICDIP 2018). pp.834-840.

Lawgali, A., Angelova, M. & Bouridane, A. 2013. HACDB: Handwritten Arabic characters database for automatic character recognition. *Visual Information Processing (EUVIP), 2013 4th European Workshop on*, pp. 255–259. IEEE.

Luqman, H., Mahmoud, S. A. & Awaida, S. 2014. KAFD Arabic font database. *Pattern Recognition* 47(6): 2231–2240.

Mohd Sanusi Azmi. 2013. A Novel Feature from Combinations Of Triangle Geometry For Digital Jawi Paleography. Phd Thesis, Department of Computer Science, Universiti Kebangsaan Malaysia.

Mahmoud, S. A., Ahmad, I., Al-Khatib, W. G., Alshayeb, M., Parvez, M. T., Märgner, V. & Fink, G. A. 2014. KHATT: An open Arabic offline handwritten text database. *Pattern Recognition* 47(3): 1096–1112.

Morillot, O., Oprean, C., Likforman-Sulem, L., Mokbel, C., Chammas, E. & Grosicki, E. 2013. The UOB-telecom Paristech Arabic handwriting recognition and translation systems for the openhart 2013 competition. *12th International Conference on Document Analysis and Recognition (ICDAR), 2013*, pp. NIST. hal-00948985.

Nashwan, F., Rashwan, M. A. A., Al-Barhamtoshy, H. M., Abdou, S. M. & Moussa, A. M. 2017. A Holistic Technique for an Arabic OCR System. *Journal of Imaging* 4(1): 6.

Pechwitz, M., Maddouri, S. S., Märgner, V., Ellouze, N. & Amiri, H. 2002. IFN/ENIT-database of handwritten Arabic words. Proc. of CIFED, Vol. 2, pp. 127–136. Citeseer.

Pham, V., Bluche, T., Kermorvant, C. & Louradour, J. 2014. Dropout improves recurrent neural networks for handwriting recognition. *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pp. 285–290. IEEE.

Rahal, N., Tounsi, M. & Alimi, A. M. 2018. Auto-Encoder-BoF/HMM System for Arabic Text Recognition. *arXiv preprint arXiv:1812.03680*.

Rashid, S. F., Schambach, M.-P., Rottland, J. & von der Nüll, S. 2013. Low resolution arabic recognition with multidimensional recurrent neural networks. *Proceedings of the 4th International Workshop on Multilingual OCR*, pp. 6. ACM.

Sabbour, N. & Shafait, F. 2013. A segmentation-free approach to Arabic and Urdu OCR. *Document Recognition and Retrieval XX*, pp. Vol. 8658, 86580N. International Society for Optics and Photonics.

Smith, R. 2007. An Overview of the Tesseract OCR Engine, *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, Parana, pp. 629-633.

Suvarnam, B. & Ch, V. S. 2019. Combination of CNN-GRU Model to Recognize Characters of a License Plate number without Segmentation. *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)*, pp. 317–322. IEEE.

Yousefi, M. R., Soheili, M. R., Breuel, T. M. & Stricker, D. 2015. A comparison of 1D and 2D LSTM architectures for the recognition of handwritten Arabic. *Document Recognition and Retrieval XXII*, pp. Vol. 9402, 94020H. International Society for Optics and Photonics.

Yousfi, S. 2016. Embedded Arabic text detection and recognition in videos. PhD thesis Université de Lyon.

Yousfi, S., Berrani, S.-A. & Garcia, C. 2015a. ALIF: A dataset for Arabic embedded text recognition in TV broadcast. *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, pp. 1221–1225. IEEE.

Yousfi, S., Berrani, S.-A. & Garcia, C. 2015b. Deep learning and recurrent connectionist-based approaches for Arabic text recognition in videos. *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, pp. 1026–1030. IEEE.

Zayene, O., Hennebert, J., Touj, S. M., Ingold, R. & Amara, N. E. Ben. 2015. A dataset for Arabic text detection, tracking and recognition in news videos-AcTiV. *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 996–1000. IEEE.

Zayene, O., Touj, S. M., Hennebert, J., Ingold, R. & Amara, N. E. Ben. 2018a. Open Datasets and Tools for Arabic Text Detection and Recognition in News Video Frames. *Journal of Imaging* 4(2): 32.

Zayene, O., Touj, S. M., Hennebert, J., Ingold, R. & Amara, N. E. Ben. 2018b. Multi-dimensional long short-term memory networks for artificial Arabic text recognition in news video. *IET Computer Vision*. 12(5): 710-719.

*Idris Saleh Al-Sheikh*
*Masnizah Mohd*
Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia
43600 Bangi, Selangor
MALAYSIA
1alshikh@gmail.com, masnizah.mohd@ukm.edu.my

*Lia Warlina*
Faculty of Engineering and Computer Science
Universitas Komputer Indonesia
Jl. Dipati Ukur 112-116 Bandung 40132
INDONESIA
lia.warlina@email.unikom.ac.id