A Review of Arabic Optical Character Recognition Techniques & Performance

Yazan M. Alwaqfi, Mumtazimah Mohamad

Faculty of Informatics & Computing, Universiti Sultan Zainal Abidin, Besut, Terengganu, Malaysia

ABSTRACT
The Artificial intelligence one of the most perspective and interesting research area that takes attention from many researchers. Several studies focused the interest in Optical Character Recognition, which is computer software designed for converting images with text into machine processed text. These OCR systems available for several languages including Arabic, Arabic OCR has been developed and improved over decades, which ultimately causes to a huge number quantity of approaches with robust results with high, in some cases about 99%, while using deep learning in Arabic OCR can results up to 100% accuracy with short time and less resources to process the image. The characteristics of Arabic text cause more errors than in English text in OCR. The aim of this paper is to analyze the related works and issues in Arabic language OCRs. The analysis results show that existing OCRs within implementation with other application exhibit defects, or at least just their subsets, such as low-resolution inputs and video-based inputs. Accordingly, it is necessary to review existing approaches that have reliable results. In addition, the review of deep learning for Arabic OCR systems and researches is very important and useful. This paper presents a literature review on the existing systems Arabic text recognition, consists of a typical mechanism, lists the differences, advantages, and disadvantages that help in adopting or expanding these systems in accordance with modern requirements.

Keywords: *Optical Arabic Recognition, OCR, Deep Learning, Machine learning.*

I. INTRODUCTION

Arabic language is one among the most important languages of the globe that's counted a mother language in various countries and millions individuals who Read, speak, and/besides write with it. In addition, there are many similarities between alphabet and alternative Aramaic alphabet based mostly languages, like Persian (Farsi) and Kurdish (Sorani) that could be a sensible reason for concentration to boost the Arabic OCR and build it helpful for similar languages. The studies of recognition for the Arabic character were started in 1975 by Nazif, as compared to other language like Latin's [1] . since 2000s the Arabic character recognition as in computer vision and image processing has won big attention from researchers [2-4], which led to made make big interest in this field with the rapid development of deep learning and image processing algorithms for the different natural languages, like Arabic, Persian, Kurdish, and Urdu languages, to enjoy the benefits which offered by deep learning. OCR involves computer systems which designed to transfer images of typewritten (printed or handwritten) text into machine-editable text to transfer pictures or documents of characters into a standard encoding scheme that representing them in ASCII format [5], where OCR systems are classifying into two categorizes describing the behavior of processing between the input and outputs processes (offline and online), where the offline type receives an images from camera or scan devise as the input file, while the online type receives the image in a real time. The most researchers worked on offline category, which divides into two ways (printed and hand writing).

II. RELATED WORKS

The main idea of Optical character recognition (OCR) is to extract text (outputs) where contained in an image (input documents) under the standard encoding scheme representing them in ASCII format [citation]. The Figure 1 shows the general structural of OCR systems as inputs and outputs.

The input of OCR system maybe, probably, can get a camera or a digital scan device, where the system will deal with image document, to get finally a useful text with the minimums number of errors, and with high accuracy of results. The accuracy is completely the most challenge in OCR systems, where sometimes considers as problems if the accuracy is appears low comparing with other languages like English language.

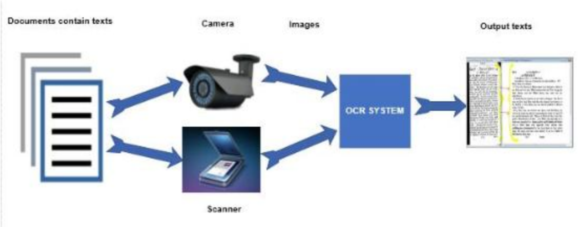

Figure 1: The OCR General Structural [6]

OCR in Arabic language is very difficult process, because the letters in Arabic language are connected without spaces in the same word, and their shape is changed in dependence of characters position in the words[4]. There are several factors effect positive or negative with the accuracy for the OCR results, like cursive written-based languages [7], low resolution scanning image [8], and the noise of image [9].

Since 1950's, the OCR systems is considering as a popular research. Several works has been done for various scripts in case of English language. Nevertheless, in case of Arabic scripts the research is still limited.

There are several problems that facing Arabic recognitions, these problems can be summarized in the following:

1. Arabic text written from right to left ,

2. Arabic language contains 28 characters; each character has more than two shapes .

3. In Arabic language there are not any upper or lower case characters,

4. The position of the character in the word or in the phrase forms the shape of it (ج ,جـ ,ـجـ ,ـج) ,(م, ,مـ ,ـمـ ,ـم).

5. In Arabic characters the dots play a significant role .

6. Some Arabic characters have the same shape but the differences between them are the position and the number of dots , this can occur either above or below the characters .

7. Short marks such as a "hamza", can be placed above or below five particular characters or can appear as isolated characters .

8. Some Arabic characters have a loop, such as (ص ,ف ,و).

9. There are six characters (ا ,د ,ذ ,ر ,ز ,و) that do not have not shape sat the beginning and in the middle of the word . Therefore, these characters do not connect to a subsequent character in a word and this causes a separation of the word into parts; Spaces separate words and short spaces separate sub words.

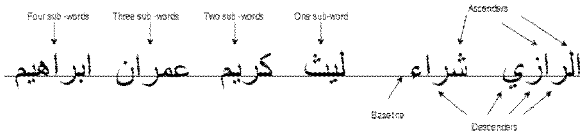

Figure 2: Words with Different Sub-words.

Arabic texts might have short marks. These marks, as shown in Figure 2, can be placed above or below characters. These marks are written as strokes and can affect recognition owing to their location either above or below the characters, like dots. Figure 3 shows marks and dots on an Arabic text. Therefore, these marks need to be recognized and a distinction made between them and the dots

Fathah	Kasrah	Dammah	Maddah	Sukun	Tanwin Fathah	Tanwin Kasrah	Tanwin Dammah	Shaddah	Hamza
ـَ	ـِ	ـُ	~	ـْ	ـً	ـٍ	ـٌ	ـّ	ء

Figure 3: Marks in Arabic Writing

The Arabic OCR

The works in OCR have been started for following segmentation-free approaches. Like, [6] proposed a segmentation free technique for Arabic OCR by performing morphological operations on text images and comparing them with existing symbol models. Thoroughly, using Fourier Transform coefficients from normalized polar images, [10] proposed especially a method to recognize multi-font cursive Arabic words. Recognition was achieved by performing template-matching using Euclidean distance as the loss metric.

Text recognition can generally be divided into two categories; Online and Offline [11, 12]. In online recognition, the characters/glyph recognized while the user writes the text - usually on a digitized pen tracer with a special stylus pen [13, 14]. Whereas; the offline recognition deals directly with the recognition of text from scanned copies of printed or handwritten texts. A comprehensive survey includingly with its focus as the differences between online and offline text recognition was done by [15].

Printed texts generally have the same font styles and sizes across prints, while handwritten texts can have varied font styles and sizes for the same writer as well as among various writers. For languages like Arabic and Urdu, where the script has a complex cursive nature and shows ligature, character level segmentation seems often an arduous task. Hence, segmentation-free techniques has gained a quite the popular appeal. Like, [16] segment words from the input script and likewise; then compute the discrete-cosine transform (DCT) features on a normalized input image. These DCT features are then used to train a neural network which performs word-classification. Another segmentation-free approach was suggested by [2], who describes an 1D HMM offline handwriting recognition system that employing an analytic approach of extracting baseline dependent features from binarized input images.

Deep learning as a type of machine learning approach in which a model learns to satisfy tasks classification immediately from text, sound, or images. Usually Deep learning executed and developed using a neural network architecture. The term "deep" refers to the number of layers in the network—the more layers, the deeper the network. conventional neural networks support only 2 or 3 layers, whereas deep learning networks may have hundreds.

The Easy access to huge marked data sets, computer power increasing, and already pre-built models by experts makes deep learning intelligent approach for image processing and OCR.

Based on deep learning Ashiquzzaman, A. proposed a novel algorithm [17] in advance to recognize Arabic hand written numbers using corresponding activation function and normalization layer, the proposed approach results significantly improved accuracy compared to the existing Arabic numeral recognition methods. The proposed model gives 97.4 percent accuracy.

Elleuch and others [18] highlighted the effectiveness of deep learning techniques in Arabic OCR for hand written script using two architectures: Deep Belief Networks (DBN) and Convolutional Deep Belief Networks (CDBN), they applied these architectures accordingly in low and high levels measuring in textual images. The experiments have shown promising results in compression to other Arabic OCR techniques

III. TECHNIQUES FOR ARABIC OCR

A. Basic OCR technique for Arabic

a) Rule-based and word context techniques

Usually, the text image segmented into images of lines. Depending on the used feature extraction and classification techniques, a character-based segmentation phase may or may not be necessary. Since Arabic text is cursive, some techniques require the segmentation of Arabic text before the feature extraction phase. During segmentation, the Arabic text image is segmented into lines. Furthermore, the line images could be segmented into words/sub-words and then to characters or even sub- characters based on the used technique. If the image under consideration contains tables and figures, then their text is extracted for recognition.

OCR system is working based on different stages (Preprocessing, Segmentation, Feature Extraction, Classification, and Post-processing) which are working in homogeneous manner, where each stage depends on previous stage [19].

b) Optical Character Recognition Stages

OCR system is working based on different stages (Figure 4) which are working in homogeneous manner, where each stage depends on previous stage [19]:

1. Pre-processing: the first stage in OCR system, where working clarify and improve the original image quality (scan image), and to reduce original image noise [20].
2. Segmentation: this stage is working to find and extract all symbols images from the original images in pre-processing stage. Segmentation algorithms are working to segment the image into lines or words, where line segmentation is working to separate text lines, while word segmentation is extracting the spacing between the words [20].
3. Feature Extraction: this stage is identifying useful information from symbols images, it's one of the fundamental problems of character recognition [21].
4. Classification: comparing the unknown images of symbols (feature extraction) with predefined stored samples in order to identify their type [2], its determines the region of feature space in which an unknown pattern falls [11].
5. Post-processing: this is the final stage in OCR system, and the most important stage [11]. Post-processing stage is working for checking the result text from previous stage, and correct it to make sure it is free from errors [12].

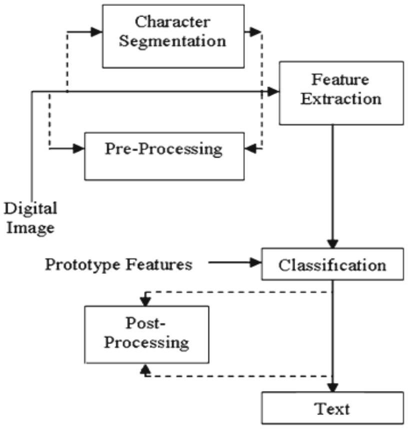

Figure 4: The OCR Stages [1]

B. Modern techniques

a) Holistic Technique

Analytical based mostly approaches in OCR systems will endure a big amount of segmentation errors, particularly once managing cursive languages like the Arabic language with frequent overlapping between characters. holistic based mostly approaches that take into account whole words as single units were introduced as a good approach to avoid such segmentation errors. Still the most challenge for these approaches is their computation complexness, particularly once managing massive vocabulary applications. during this paper, we tend to introduce a computationally economical, holistic Arabic OCR system. A lexicon reduction approach supported agglomeration similar formed words is employed to scale back recognition time. victimization international word level distinct trigonometric function rework (DCT) primarily based options together with native block based options, our planned approach managed to generalize for brand spanking new font sizes that weren't enclosed within the coaching information. analysis results for the approach victimization completely different check sets from fashionable and historical Arabic books area unit promising compared with state of art Arabic OCR systems [20].

b) Neural network and Deep Learning

Artificial Neural Network consists of simple processing elements and a very high degree of interconnection [17]. The weights of the network are learned from training data. The weight are initialized into the initialized input layer, hidden layers and on the final output layer. We have user the cross entropy function to compute the error rate. The extracted information from data will be processed from input layer to output layer gives a character in this task. We have developed this algorithm for learning our artificial neural network [22].

c) Convolutional Neural Network Architecture

(1) CNN Classifier

Being stratified, multi-layer neural networks with a deep supervised learning architecture trained with the back-propagation algorithmic rule, Convolutional Neural Networks are composed of an automatic feature extractor and a trainable classifier. CNN exploited to learn advanced, high dimensional information, and dissent in however convolutional and sub-sampling layers inquired into. The distinction is in their design. Several CNN architectures advised for various issues among that object recognition and handwriting digit/character recognition. The most effective performance on pattern recognition task achieved. Additionally, to ensure some extent of in variance to scale, shift and distortion, CNN combine three main stratified aspects like native receptive fields, weight sharing and special sub-sampling. As shown in Figure 5, cyberspace represents a typical Convolutional Neural spec for written character recognition. It includes a group of many layers. Initially, the input convoluted with a group of filters (C hidden layers) to get the values of the feature map. Next, to diminish the spatiality (S hidden layers) of the special resolution of the feature map, a sub-sampling layer pursues every convolution layer. Convolutional layers alternate sub-sampling layers represent the feature extractor to retrieves discriminating options from the raw pictures. Ultimately, two fully connected layers (FCL) and the output layer pursued these layers. Each layer because the input takes the output of the previous layer.

The image features can automatically extracted in CNN by converting the input structure using the network layers, in which the CNNs are working in dependence with some mathematical operation that called convolution. The convolution defined as some mathematical multiplication operation where each pixel in the image multiplied with each value in the kernel, which results another matrix then summing the products. Convolution operation aim to reproduce another image from the raw image in advance to enhance dissimilar features that extracted from the raw image, this leads to achieve high classification process accuracy [23].

CNN contains different layers type, where the convolution layer (feature extractor) extracts from the raw image the features. At this layer, the features did not identified and where it will be located by CNN in the raw image. Since filter matrix in CNN trying to find the shapes in the image. The second layer is Rectified Linear Unit (ReLU) which is an element wise operation, and considering that the neurons activation function implemented to generate the output after every convolution in the first layer [24]. ReLU aims to convert the bad and unsuitable pixels into zeros in advance to keep the positive pixels only. The Pooling or subsampling is the next layer, the goal in this layer is reduction the dimensionality for each filtered image, and nevertheless it should keep the all the most important selected features in the previous layer. The Pooling layer reproduces images with less number of pixels keeping the same images number from the preceding layer, which leads to fully managing the computational weights. The next layer is the Dropout layer; this layer aims to reduce the retraining.

The random neurons collection in this layer called Dropout, there activation were adjusting to zero. Dropout employed to ensure that the network could summarize the test data, gaining weights that are unresponsive to training samples [25].

Finally, the fully connected layer, which aims to connect every neuron in the previous layer to every neuron in the next layer. The figure below shows the CNN architecture for the OCR problem:

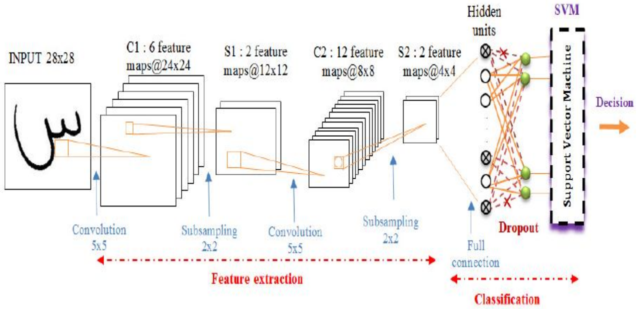

Figure 6: The CNN architecture for the OCR problem

CNN developed to map and processing images as input data to output variables. The CNN have proven so effective that they are a good method of transition to any prediction problem that includes image as input data.

The advantage of victimization CNN lies in their ability to develop an enclosed illustration of a two-dimensional image. this enables the model to check the position and scale in various data structures, that is vital when operating with illustrations.

The CNNs can be used for Image data, prediction problems classification and prediction problems regression.

Traditionally the CNN input as two-dimensional, a matrix or a field, and can be modified to one-dimensional, that permitting CNN to develop an internal representation of a one-dimensional sequence.

This benefit permits the CNN to be used additional typically on different kinds of information that incorporates a special relationship. Though not specifically developed for non-image information, CNNs come through progressive results on issues such as document classification employed in sentiment analysis and related problems.

While Habeeb in [8] post-processing stage, which it the last stage in the OCR system. He considered the post-processing stage is working to corrects all possible errors of OCR output text. The proposed approach used hybrid techniques, which it combines of differentiation, alignment, and voting to overcome the identified drawbacks, and then evaluate each one of this technique separately using non-word error rate (NWER), word error rate (WER), and Character error rate (CER) with each technique. The experiment results for all measurement show error rate in WER is 30.35%, CER is 52.42%, and 47.86% for the NWER. However, the implemented hybrid techniques outputs recognition results with high error rate, which is minimizes the accuracy of the processed text image.

In [26] present an approach for improve Arabic character recognition methods, aims to segmentation-free character recognition. For the recognition process in the proposed approach, he selected a sub-image, from the document image (original image) using slide widow with the size of a reference character. The highest probability characters in Arabic writing will begin firstly in the proposed approach, while the characters that produce a recognition error will excluded. The proposed approach grouped all the characters in one cluster and recognized together for each characters that have same dimensions, to make the approach procedure faster. They tested their model using dataset Arabic characters' images; they implement five types of fonts, and nine font sizes. The unknown characters are classified and carried out by extracting their transform features and their structural using the height, width, and the number of pixels above baseline. The experimental results show that time needed for recognition process is low with the small font size characters, because of a few numbers of the lines for the small character's documents.

In [15] considered segmentation stage as an essential stage in OCR systems, and it is error-prone stage, where extensive portion of processing is devoted and a considerable share of recognition errors is attributed. They proposed a new approach using novel segmentation for Arabic printed text with diacritics. For the lost data, the proposed approach used the skeleton methods, which give more advantage for reducing computation, errors, with a clear description for the sub-word. The experimental results show 98.7% for the system accuracy.

IV. RELATED WORKS ON CONVOLUTONAL NEURAL NETWORK TECHNIQUES

In [2] present a model using two classifiers for offline Arabic handwriting recognition (OAHR): Convolutional Neural Network (CNN), and Support Vector Machine (SVM). The proposed approach modifies the classifier of the convolutional neural network (CNN) by the support vector machine (SVM) classifier. Using the raw image, the proposed model found that both classifiers are automatically extracts the features. They evaluate the recognition of Arabic characters on the handwritten; they used the HACDB and IFN/ENIT databases. The experimental results show the efficiency of SVM based-CNN model with dropout performs than the CNN based-SVM model without dropout and the standard CNN classifier. The performance of their model compared with character recognition accuracies gained from state-of-the-art Arabic Optical Character Recognition, producing favorable results.

In [4] present, a model for recognizing Arabic handwritten characters, and besides comparing it with other deep learning architecture, to take the advantages for managing large dimensions. In supervised model, the proposed model used the convolutional Neural Network (CNN) as a special type of feed-forward multilayer to increase the performance of CNN. Using database that contain 16800 of handwritten Arabic characters, indeed, they trained and tested the convolutional Neural Network (CNN). The experimental results on testing data show that CNN had a 5.1% as an average for misclassification error, and it had a better performance in both of big data and images.

In [9] considered the OCR system is the auto conversion for the handwritten or typewritten (scanned images) into machine-encoded text (ASCII format). The proposed approach depended on Artificial Neural Networks (ANNs) to recognizes Arabic handwritten characters, through implement an off-line OCR system; they used ANN as a system classifier, where trained using Hopfield algorithm. The proposed system worked on image through a preprocessing stage in the beginning, and then followed by a feature extraction stage and a recognition stage. The accuracy for the recognition process is precisely accurate a certain property (extracted features from the image) for each letters that are calculated. The most important factor for achieving a high performance for recognition sounds the feature extraction, where the collection of vectors or features defines the character uniquely by the means of an ANN. According to the experimental results, the proposed system is able to recognize some of Arabic handwritten letters (أ، ب، ث، س، خ، ض، ع، و), with a high accuracy (77.25) for the recognition crystally.

In [3] study, they present an approach of Alphanumeric VGG net, which it is inspired from deep state-of-the-art VGGNet, and it is constructed by thirteen convolutional layers (three fully connected layers, and two max-pooling layers). The proposed approach is fast, reliable and aims to recognize Arabic handwritten alphanumeric characters, it has improves the performance for the classification process, besides reducing the overall complexity of VGGNet. Using two benchmarking databases, they evaluate their approach. The experimental results show that: the accuracy of 97.32% for the HACDB database and 99.66% for the ADBase database have been achieved.

While [20] considered the text preprocessing and segmentation algorithms as a presentative example for: the most important factors for the accuracy in the Optical Character Recognition (OCR) systems. For the line and word recognition, they used the column vector sum and row vector sum (total no. of lines, total no. of words), and finding the maximum valued patterns. Obviously, they present algorithm for correction of skew angle that generated in scanning of the text document, and a novel profile based method for segmentation of printed text, which assertively separates the text in document image into lines, words and characters. The proposed algorithm, which implemented in MATLAB, and tested it with several printed document images. The experimental results show the algorithm has a high accuracy only of printed documents without broken or overlapping characters.

While [16] developed an Arabic scene text recognition especially through using Convolutional Neural Networks (ConvNets), which is considered as a deep learning classifier. They have applied and evaluated their approach on an intrinsic Arabic script and have used five orientations with respect to the single occurrence of a character in order to; deal with maximum variations. Their experimental results showed that the ConvNets could deeply enhance the accuracy of a different and large dataset for a captured Arabic scene text pattern in order to; get better performance. They assumed that: ConvNets could extract more detailed features, which considered robust, but not handling features through layers, which may get a more precised detail of the image.

In [27] presented a model for Optical Character Recognition (OCR) in Telugu language, which includes three parts: a database of Telugu characters, a deep learning-based OCR algorithm and an online client-server application for the developed algorithm. Their model based on Convolutional Neural Networks (CNNs) algorithm reasonably to classify the characters. They have applied their OCR system to real data and the results were being good. It needs to extend the work to include other languages with regardingly the scope of having a common OCR system.

In [5] proposed a region-based Deep Convolutional Neural Network model for document structure learning. The main level of 'inter-domain' transfer learning demanded by getting weights from a pre-trained VGG16 architecture on the ImageNet dataset. Then, the secondary level of 'intra-domain' transfer learning was implemented for a quick training on deep learning models for image segments. In last, the stacked generalization-based assembling applied for merging the forecasting of the base deep neural network models. Their proposed model achieved an accuracy of 92.2% for the known RVL-CDIP document image dataset, meanwhile; exceeded other existing algorithms on benchmarks set. The multi-domain trained DCNN models can enhanced to act as general text detectors and on other application such as Automated Driving Assistance Systems (ADAS).

The CNN Techniques can be useful and fully developed in Arabic OCR, developing the CNN techniques and using the benefits of the good layers implementation can bring perfect results in Arabic OCR systems. One of the most important layers in the CNN that can be modified is the features selection in which the accuracy of the results can be achieved with high values.

V. ANALYSIS AND DISCUSSION

In this document, it is clearly to denote various techniques for Arabic Optical Character Recognition (OCR), as previously mentioned, there is a variation in excising systems and techniques in different aspects. In general, six pre-processing phases are specifically implemented regardless the developed system but especially depending on the characteristics of the input. In addition, the systems that are depend on the phases, have used the pre-processing phase to improve and enhance the outputs (results), The classification algorithm have already been the general processing phase that applied in the systems. Nevertheless, the most evident difference between exactly the existing systems weather in case if the system is segmentation free or it is segmentation based. Privately, in the classification algorithms utilized. The variations usually in the inter-process, which used in the classification algorithms, feature extraction and character segmentation.

So, if the main phases are generally in all recent implemented systems and the variation is in the inter-processing phase, the deep learning can be stratified the needs of enhancing Arabic optical character recognitions that taking in care carefully the specification of Arabic language complexities, Arabic characters and features.

All presented studies and techniques take Arabic OCR in the same processing stages, segmentation and pre-processing stages have a big attention from the most studies, and they are needed phases in text recognition process, to dramatically extract the characters from the image.

However; the featured extraction has less attention, while the performance of an each character recognition system that depends on the features that extracted, although; the feature extraction phase gets the important phase in finding the high accuracy output text results. Thus, choosing appropriate features matrix can provides to more optimal and acceptable results.

VI. CONCLUSION

Finally, as a perfect illustration over the above, Arabic optical character recognition (OCR) is still an opened domain for the researchers; few efforts have been put particularly on the research of Arabic characters, because due to, the complexities of Arabic language comparing with other languages like English. In this paper, we have deeply argued and studied literatures in the field of optical character recognition (OCR) in the Arabic language; we present technique that was used, advantage, furthermore, disadvantage for each study. Regarding future work, authors intend to offer an efficient Optical character recognition (OCR), which makes recognition process more application-aware using into a neural network.

ACKNOWLEDGEMENT

This work is supported by UniSZA Center of Excellence Management and Research Incubator and University Sultan Zainal Abidin, Terengganu, Malaysia.

REFERENCES

[1] Elleuch, M., R. Maalej, and M. Kherallah, "*A new design based-SVM of the CNN classifier architecture with dropout for offline Arabic handwritten recognition*". Procedia Computer Science, 2016. 80: p. 1712-1723.

[2] Mudhsh, M. and R. Almodfer, "*Arabic handwritten alphanumeric character recognition using very deep neural network*". Information, 2017. 8(3): p. 105.

[3] El-Sawy, A., M. Loey, and E. Hazem, "*Arabic handwritten characters recognition using convolutional neural network.*" WSEAS Transactions on Computer Research, 2017. 5: p. 11-19.

[4] Al-Masoudi, A.F.R. and H.S.R. Al-Obeidi, "*Smoothing Techniques Evaluation of N-gram Language Model for Arabic OCR Post-processing*". Journal of Theoretical and Applied Information Technology, 2015. 82(3): p. 432.

[5] Boling, C. and K. Das, "*Semantic Similarity of Documents Using Latent Semantic Analysis*". 2014 NCUR, 2014.

[6] Estes, Z. and S. Simmons. "*Using Latent Semantic Analysis to Estimate Similarity*". in Proceedings of the Annual Meeting of the Cognitive Science Society. 2006.

[7] Hussien, R.S., A.A. Elkhidir, and M.G. Elnourani. "*Optical character recognition of arabic handwritten characters using neural network*". in 2015 International Conference on Computing, Control, Networking, Electronics and Embedded Systems Engineering (ICCNEEE). 2015. IEEE.

[8] Habeeb, I.Q., "*Hybrid model of post-processing techniques for arabic optical character recognition*". 2016, Universiti Utara Malaysia.

[9] Modi, H. and M. Parikh, "*A review on optical character recognition techniques*". Int J Comput Appl, 2017. 160(6): p. 20-24.

[10] Habeeb, I.Q., et al., "*Two bigrams based language model for auto correction of Arabic OCR errors*". International Journal of Digital Content Technology and its Applications (JDCTA), 2014. 8(1): p. 72-80.

[11] Mohammad, K., et al. "*Printed Arabic optical character segmentation. in Image Processing: Algorithms and Systems XIII*". 2015. International Society for Optics and Photonics.

[12] Ahmed, S.B., et al. "*Deep learning based isolated Arabic scene character recognition*". in 2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR). 2017. IEEE.

[13] Konkimalla, C.P., et al., "*Optical Character Recognition (OCR) for Telugu: Database, Algorithm and Application*". arXiv preprint arXiv:1711.07245, 2017.

[14] Das, A., et al. "*Document Image Classification with Intra-Domain Transfer Learning and Stacked Generalization of Deep Convolutional Neural Networks*". in 2018 24th International Conference on Pattern Recognition (ICPR). 2018. IEEE.

[15] Srivastava, N., et al., "*Dropout: a simple way to prevent neural networks from overfitting*". The Journal of Machine Learning Research, 2014. 15(1): p. 1929-1958.

[16] Mars, A. and G. Antoniadis. "*Handwriting recognition system for Arabic language learning*". in 2015 World Congress on Information Technology and Computer Applications (WCITCA). 2015. IEEE.

[17] Ashiquzzaman, A. and A.K. Tushar. "*Handwritten Arabic numeral recognition using deep learning neural networks.*" in 2017 IEEE International Conference on Imaging, Vision & Pattern Recognition (icIVPR). 2017. IEEE.

[18] Elleuch, M., N. Tagougui, and M. Kherallah. "*Deep learning for feature extraction of Arabic handwritten script*". in International Conference on Computer Analysis of Images and Patterns. 2015. Springer.

[19] Al-shatnawi, A.M. and K. Omar, "*The Thinning Problem in Arabic Text Recognition-A Comprehensive Review*". International Journal of Computer Applications, 2014. 103(3): p. 0975-8887.

[20] Shahin, A.A., "*Printed Arabic text recognition using linear and nonlinear regression*". arXiv preprint arXiv:1702.01444, 2017.

[21] Besbas, W.S., M.R. Sunni, and A.F. Elbokhare, "*Improved Method for Sliding Window Printed Arabic OCR*".

[22] Samra, Y.K.A., "*Tag Recommendation for Short Arabic Text by Using Latent Semantic Analysis of Wikipedia*". Tag Recommendation for Short Arabic Text by Using Latent Semantic Analysis of Wikipedia, 2017

[23] Liu, W., et al. "*ITNLP-AiKF at SemEval-2017 Task 1: Rich features based svr for semantic textual similarity computing*". in Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). 2017.

[24] Mahmoud, A. and M. Zrigui. "*Semantic similarity analysis for paraphrase identification in Arabic texts*". in Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation. 2017.

[25] Mohebbi, M. and A. Talebpour, "*Texts semantic similarity detection based graph approach*". Int. Arab J. Inf. Technol., 2016. 13(2): p. 246-251.

[26] Singh, P. and S. Budhiraja, "*Feature extraction and classification techniques in OCR systems for handwritten Gurmukhi Script–a survey*". International Journal of Engineering Research and Applications (IJERA), 2011. 1(4): p. 1736-1739.

[27] Mars, A. and G. Antoniadis, "*Arabic online handwriting recognition using neural network*". International Journal of Artificial Intelligence and Applications (IJAIA), 2016. 7(5).