Arabic Handwritten Manuscripts Text Recognition: A Systematic Review

Arwa Alghamdi[1], Dareen Alluhaybi[2], Doaa Almehmadi[3], Khadijah Alameer[4], Sundos Bin Siddeq[5] and Tahani Alsubait[6,*]

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

Summary
Handwritten text recognition is one of the active research areas nowadays. The progress in this field differs in every language. For example, the progress in Arabic handwritten text recognition is still insignificant and needs more attentions and efforts. One of the most important fields in this is Arabic handwritten manuscript text recognition which focuses in extracting text from historical manuscripts. For eons, ancients used manuscripts to write everything. Nowadays, there are millions of manuscripts all around the world. There are two main challenges in dealing with these manuscripts. The first one is that they are at the risk of damage since they are written in primitive materials, the second challenge is due to the difference in writing styles, hence most people are unable to read these manuscripts easily. Therefore, we discuss in this study different papers that are related to this important research field.

Keywords:
Arabic handwritten text; Manuscripts; Text recognition; Convolutional Recurrent Neural Network.

1. Introduction

Optical Character Recognition (OCR) is an important subfield of Computer Vision (CV). In this field, we recognize text and encode it into a digital format [1]. We can classify text recognition into two main categories, namely: online recognition and offline recognition [2]. The difference between these two categories is that in online recognition the text is being recognized in real time, while in offline text recognition the text is being recognized from scanned images, whether it is a typed or handwritten text. The text can be recognized either as words, or as separated characters [2]. OCR systems go by different phases, usually the first phase is preprocessing which includes different steps. For example, binarization, noise removal, and skew correction.

There are many types of text recognition problems such as printed text and handwritten text. Handwritten text recognition is one of the active research areas because of its importance and difficulty, as handwritten styles vary depending on different factors, for example, fonts type and unique writers' styles. In order to solve text recognition problems, different methods can be used depending on the type of the problem and the dataset used.

2. Methodology

2.1 Searching for Relevant Articles

In order to gather relevant articles for this research, four scientific databases were used for searching: the ACM digital library, IEEE Xplore digital library, ScienceDirect, and Google Scholar. These databases were chosen because of the quality of their publications and the abundance of their content in the fields of technology and computer science.

The search for the papers was held in two stages. The following keywords were used in the first stage for the first experiment: "manuscript", "Arabic", "text recognition", "OCR", "machine learning", "deep learning", "segmentation", "classification", "extract", "Quran", "CNN", "handwritten", and "historical." While the following keywords were used for the second experiment: "CRNN", "hand-written", "manuscripts", "YOLO", "YOLOv5", "object detection", "text recognition".

2.2. Inclusion and Exclusion Criteria

After completing the search for articles related to the research topic, some of them were excluded based on some criteria. The inclusion criteria that have been identified are the relevance of title of the article, its abstract, and the language of article (English).

2.3. Data Extraction

The basic items that were relied upon to extract data from scientific papers are shown in Table 1.

Manuscript received November 5, 2022
Manuscript revised November 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.11.47

2.4. Data Analysis

After extracting the data, it was analyzed using some criteria based on the field. For example, if the study was to be added under “Text Recognition” section of related articles, the most focus was on the parts of the type of method, the dataset, and the accuracy in order to compare the different papers. On the other hand, if the study was to be added under one of the preprocessing sections, most of the focus was on the theoretical concepts and mathematical equations.

Table 1: Data extraction

Item	***Description***
Title	Title of the paper
Author(s)	Author(s) name
Date	Publishing year
Country	Country of authors
Method(s)	Method(s) used in the paper
Dataset(s)	Dataset(s) used in the paper
Accuracy	Accuracy of the result in the paper
Field	Field of the paper

3. Results

The number of scientific articles that were collected from the previously mentioned databases reached one hundred fifteen, and eight books and theses. Thirty-one scientific papers were excluded. Thus, the final number of scientific studies is eighty-five. sixty-four were collected for the first experiment, and nineteen references were collected for the second experiment. The exclusion was based on the previously mentioned criteria.

Fig. 1 shows the countries with the number of papers that were published from each one of them. It is noticed that China is the most popular country for scientific papers related to text recognition.

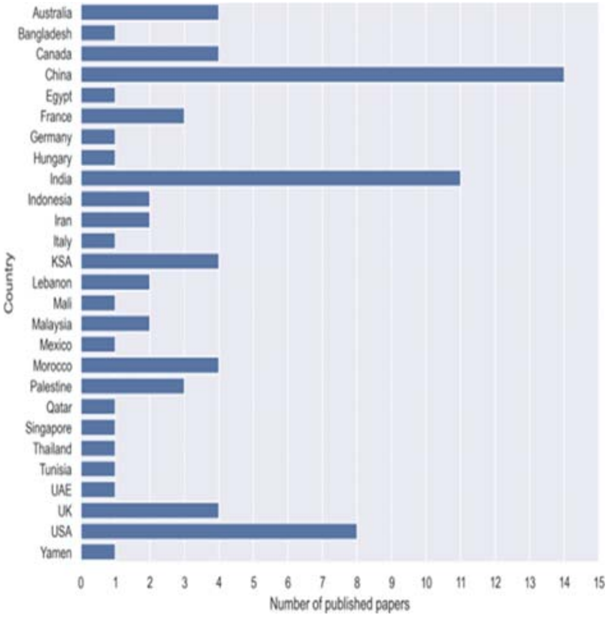

Figure 1: Geographic distribution of the collected studies

Fig. 2 shows that the number of published papers increases by years, which indicates that with the advancement of time, the interest in this field increases. Moreover, two keywords were used to search in ScienceDirect database which are “OCR” and “manuscript”, Fig. 3 shows the number of publications for each year from 1997 to 2020, the number of publications clearly increases with time which also proves how much the interest of this field is increasing.

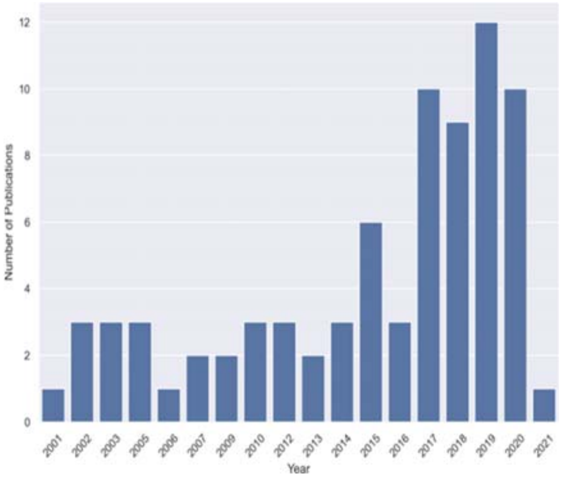

Figure 2: Publication year of the collected studies

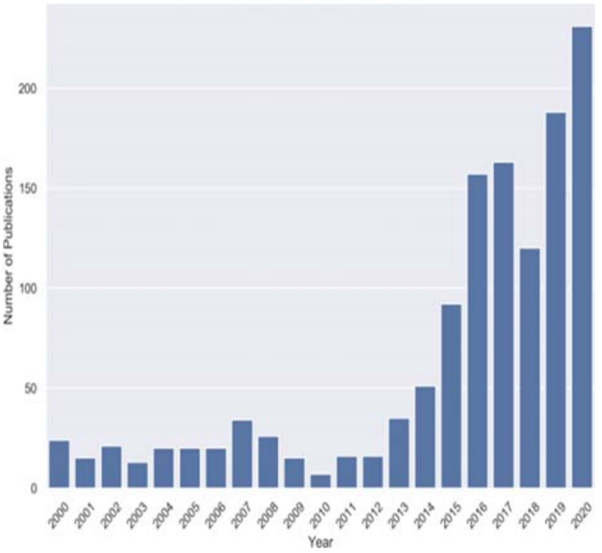

Figure 3: Research interest using "OCR" and "manuscript" keywords

4. Challenges and Limitations

There are many historical manuscripts that have not been studied and have not been discovered yet. Some of them are preserved for fear of losing them. Recognition of handwritten manuscripts will help to read millions of manuscripts all around the world. It also facilitates the access to them and its conversion into a publishable and an editable form.

Manuscripts can also be scanned and archived, algorithms that process manuscript images have also been developed, thus they give better results and high accuracy.

The difficulty is concentrated in the ability to scan documents properly because they may be in poor quality since papers are usually easily damaged. In addition, in historical manuscripts, the passage of time may lead to the damage of the manuscripts. Furthermore, among the difficulties is the difference in fonts that may cause problems in recognizing texts. Regarding Arabic manuscripts, the most obvious characteristics of the Arabic language is that Arabic scripts are inherently cursive which leads to difficulties in segmentation the words.

5. Comparison and Discussion

Arabic text recognition has become an interesting field for many researchers in the last few years. Moreover, it is a challenging task due to the complexity of Arabic writing. The texts recognition methods are different from one research to another, but the steps that most methods used are preprocessing, segmentation, and classification. There are some research papers that are concerned with Arabic manuscripts text recognition, others are less related but still focus in the same research field. In [3, 4, 5], using SVM classifier yields the best results. Noticing that using HOG and LBP as features extraction algorithms in [3] increased the accuracy up to 97.05%. On the other hand, it is shown that using CNN, DCNN and DNN, we can accomplish features extraction and classification.

In [6], using DNN and DCNN, the resulted accuracy was 98%, and the DCNN model was more efficient than the DNN. Also, in [7, 8, 9], the authors used CNN which resulted a better accuracy in all experiments, where the highest accuracy was 98.47% and the lowest one was 74.29%. Also, it is important to mention that the highest accuracy when using K-NN classifier in [4, 5, 10] was 93.79% and the lowest one was 65.79%.

Looking at another point of view, increasing the data and decreasing number of classes improves the accuracy, for example, in [11] only ten texts written in the same way by one writer were used, thus, the resulted accuracy is 100%, this is also clear in the different experiments that were done in [7]. Some other methods were also used such as HMM in [12, 13, 14], in [14] the achieved accuracy is 72.10%, it also showed encouraging results in [12], however, in [13], the accuracy was not mentioned, which shows that this filed still needs much effort and attention from researchers all around the world. In [15, 16] the dataset size was large, which led to high accuracy results. Moreover, some papers have improved their accuracy by adding improvements to the model. In [17] it used man-machine recognition which increases the accuracy, and in [18] the text was manually inserted to the system in a line-by-line basis.

On the other hand, in [19] the accuracy increases when lexicons were used, despite that, using lexicons has a drawback, which is that if the word does not exist in the lexicon, the model will not be able to recognize it. The highest accuracy of previous papers was in [17] with a value of 99.0% when using the man-machine recognition. And the lowest accuracy was in [15] with a value of 78.2% in IIIT5K dataset when the lexicon was not used.

To wrap up, in this section we made a systematic review of the different collected references in this research. This includes making different statistics, analysis and plots. In addition, we discussed different related works in the specific field of Arabic handwritten manuscripts text recognition, and some other papers that focus on the Arabic handwritten text in general. In these papers, different approaches were used, such as LBP, GF and HOG for features extraction [4, 3], and SVM and K-NN for classification [3, 20]. Also, Deep Learning approaches may be used for both features extraction and classification at the same time, such as DCNN and CNN [7, 8, 9]. Results vary depending on different factors, for example, the complexity of the dataset, the used approaches, and the amount of the data used for training.

6. Conclusion and Future Work

In this study, we made a systematic review of different papers related to the field of Arabic handwritten manuscripts text recognition. We presented the methodology of extracting, filtering, and analyzing the papers. Moreover, we made a comparison and discussion of exiting work in the field. We noticed that the progress in this field is still slow and need more efforts and attentions from the researches all around the world.

As a future work, we aim to increase the number of collected papers and add more factors in the analysis phase, as well as discussing them in more details.

Acknowledgments

We would like to thank the manuscripts department at King Abdullah Library at Umm Al-Qura University for providing their expertise.

References

[1] A. Shinde and D. Chougule, “Text pre-processing and text segmentation for OCR,” International Journal of Computer Science Engineering and Technology, vol. 2, no. 1, pp. 810–812, 2012.

[2] A. Priya, S. Mishra, S. Raj, S. Mandal, and S. Datta, “Online and offline character recognition: A survey,” in 2016 International Conference on Communication and Signal Processing (ICCSP), pp. 0967–0970, IEEE, 2016.

[3] A. Zafar and A. Iqbal, “Application of soft computing techniques in machine reading of Quranic Kufic manuscripts,” Journal of King Saud University- Computer and Information Sciences, 2020.

[4] K. Adam, S. Al-Maadeed, and A. Bouridane, “based classification of Arabic scripts style in ancient Arabic manuscripts: Preliminary results,” in 2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR), pp. 95–98, IEEE, 2017.

[5] Z. Kaoudja, M. L. Kherfi, and B. Khaldi, “An efficient multiple-classifier system for Arabic calligraphy style recognition,” in 2019 International Conference on Networking and Advanced Systems (ICNAS), pp. 1–5, IEEE, 2019.

[6] M. Elmansouri, N. E. Makhfi, and B. Aghoutane, “Toward classification of arabic manuscripts words based on the deep convolutional neural networks,” in 2020 International Conference on Intelligent Systems and Computer Vision (ISCV), pp. 1–5, IEEE, 2020.

[7] B. Alrehali, N. Alsaedi, H. Alahmadi, and N. Abid, “Historical Arabic manuscripts text recognition using convolutional neural network,” in 2020 6th Conference on Data Science and Machine Learning Applications (CDMA), pp. 37–42, IEEE, 2020.

[8] R. Alaasam, B. Kurar, M. Kassis, and J. El-Sana, “Experiment study on utilizing convolutional neural networks to recognize historical Arabic hand- written text,” in 2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR), pp. 124–128, IEEE, 2017.

[9] R. Alaasam, B. K. Barakat, and J. El-Sana, “Synthesizing versus augmentation for Arabic word recognition with convolutional neural networks,” in 2018 IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR), pp. 114–118, IEEE, 2018.

[10] H. Boukerma and N. Farah, “Preprocessing algorithms for Arabic handwriting recognition systems,” in 2012 International Conference on Advanced Computer Science Applications and Technologies (ACSAT), pp. 318–323, IEEE, 2012.

[11] L. Dounas, M. I. Azzouzi, and N. Rais, “New algorithm for the transcription of Arabic manuscripts,” in 2012 Colloquium in Information Science and Technology, pp. 86–90, IEEE, 2012.

[12] M. S. Khorsheed, “Recognising handwritten Arabic manuscripts using a single hidden Markov model,” Pattern Recognition Letters, vol. 24, no. 14, pp. 2235–2242, 2003.

[13] N. E. makhfi, “Handwritten Arabic word spotting using speeded up robust features algorithm,” in 2019 5th International Conference on Optimization and Applications (ICOA), pp. 1–6, IEEE, 2019.

[14] E. Chammas, C. Mokbel, and L. Likforman-Sulem, “Arabic handwritten document preprocessing and recognition,” in 2015 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 451–455, IEEE, 2015.

[15] B. Shi, X. Bai, and C. Yao, “An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition,” IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 11, pp. 2298–2304, 2016.

[16] X. Feng, Z. Wang, and T. Liu, “Port container number recognition system based on improved yolo and CRNN algorithm,” in 2020 International Conference on Artificial Intelligence and Electromechanical Automation (AIEA), pp. 72–77, IEEE, 2020.

[17] L. Zhao and K. Jia, “Application of CRNN based OCR in health records system,” in Proceedings of the 3rd International Conference on Multimedia Systems and Signal Processing, pp. 46–50, 2018.

[18] R. Achkar, K. Ghayad, R. Haidar, S. Saleh, and R. Al Hajj, “Medical hand- written prescription recognition using

CRNN," in 2019 International Conference on Computer, Information and Telecommunication Systems (CITS), pp. 1–5, IEEE, 2019.

[19] L. Chen and S. Li, "Improvement research and application of text recognition algorithm based on CRNN," in Proceedings of the 2018 International Conference on Signal Processing and Machine Learning, pp. 166–170, 2018.

[20] M. Elleuch, N. Tagougui, and M. Kherallah, "Arabic handwritten characters recognition using deep belief neural networks," in 2015 IEEE 12th International Multi-Conference on Systems, Signals & Devices (SSD15), pp. 1–5, IEEE, 2015.

[21] Giurgiu, I., Riva, O., Juric, D., Krivulev, I., Alonso, G.: *Calling the Cloud: Enabling Mobile Phones as Interfaces to Cloud Applications*. In: Bacon, J.M., Cooper, B.F. (eds.) Middleware 2009. LNCS, vol. 5896, pp. 83–102. Springer, Heidelberg (2009).