Arabic Optical Character Recognition Challenges

Subjects: Computer Science, Artificial Intelligence

Contributor: Rayyan Najam , Safiullah Faizullah

When it is necessary to store or edit a text written by hand in Arabic, this can only be performed manually, which can take significant time. However, fortunately, optical character recognition can be applied for this particular case. Optical character recognition (OCR) is a technique that is used to read and recognize the text present in an image and then convert it into a textual format. Once the text is extracted and digitized, it can utilize the applications of storing, retrieving, searching, and editing. Here, researchers shed light on the challenges accompanying OCR in general, then narrow it down to the case of Arabic OCR.

Arabic optical character recognition Arabic OCR challenges

1. Introduction

Interchangeably, OCR is referred to as “text recognition”. Text recognition has many benefits and applications, especially in the automation of pipelines in libraries and healthcare institutions, the overcoming of learning difficulties, and transportation by auto-plate recognition, leading to reduced human intervention and effort, a promise AI-based solutions still deliver [1][2][3][4][5][6].

The text to be recognized can be divided into one of two types: handwritten or printed. Handwritten-text recognition (HTR) tasks can themselves be divided into two types: online and offline. Offline handwritten-text recognition is the recognition of text that is not produced by computer devices and is not produced at the time of recognition but, rather, is usually found in still documents, records, and images. On the other hand, online text recognition focuses on text produced by computer-based devices, such as digital pens and tablets, rather than being extracted from documents [7][8][9][10][11].

Arabic, the language of the Quran, is acknowledged as the lingua franca in 25 countries and is considered the sixth most commonly spoken native language around the world. It is used by more than one billion Muslims around the globe and is spoken by more than 274 million people [12]. Its script’s morphological, lexical, and semantic richness make it one of the most challenging languages computationally [13]. Specifically, the Arabic writing system is from right to left and consists of 28 letters, each of which can take multiple forms, depending on its position in the word: alone, in the beginning, in the middle, or at the end. Moreover, some of the characters that can be turned into other characters simply by adding dots. Additionally, some characters have loops, while others have similar skeletons. Moreover, the Arabic writing system is connected, meaning that connections can be formed whenever two letters are combined. In addition, diacritics can be present in Arabic text in multiple forms, and different diacritics can

change the meaning of a word slightly in most cases. The fonts are rich in their diversity as they originate from different eras and geographies. All these characteristics make recognition tasks difficult [14][15][16]. However, this challenge can also be viewed as an opportunity.

Predominantly, the performance of OCR by an OCR system or engine involves the following steps: (1) image acquisition, in which the image is collected; (2) pre-processing, which involves steps to preprocess the previously acquired image, such as binarization and thinning; (3) segmentation, which can be performed either line-wise, word-wise, or character-wise; (4) feature extraction, in which the important visual features are extracted, which can help in the subsequent step; (5) classification, in which the recognition is performed by devising models that utilize the features extracted previously; and (6) post-processing techniques, which deal with the produced text to correct it further [17][18].

2. OCR Challenges [19][20][21]

Many challenges can occur in the OCR in images from which text is extracted, such as the following:

- Scene complexity: The segregation of text from non-text can be difficult in cases in which the rate of noise is high or when there are strokes similar to text, but which are part of the background and are not texts themselves;
- Conditions of uneven lighting: Due to the presence of shadows, pictures taken of any document can have the problem of uneven light distribution, in which part of the image has more light while the other is quite dark. Thus, scanned images are preferred, since they eliminate such unevenness;
- Skewness (rotation): The text has skewed lines and deviates from the distinctive orientation. This can be a consequence of manually taking pictures;
- Blurring and degradation: The grade of the image can be that of low quality. A remedy for this problem is to sharpen the characters;
- Aspect ratios: Different aspect ratios can require different search procedures to detect the text;
- Tilting (perspective distortion): The page itself may not appear in a correct, parallel perspective. The taking of photographs of texts by non-experts without the use of a scanner can account for such non-parallelism, making the task difficult. A simple solution is to use scanners or to try to take photographs for each page while perceiving the perspectives. An example is shown in **Figure 1**;
- Warping: This problem is related to the text itself. The page may be perfectly parallel at the time at which the picture is taken; however, the nature of the text and how it is written are warped, which is not usually the case in writing;

- Fonts: Printed fonts and styles vary from those that are handwritten. The same writer can write the same passage in the same space in a different way each time. On the other hand, printed text is easier because the differences tend to be less frequent from one print to another;
- Multilingual and multi-font environments: Different languages feature different complexities. Some languages use cursive scripts with connected characters, as in the Arabic language. Moreover, recent individual pages can contain multiple languages, while old writings can have multiple fonts within the same page if they are written by different writers.

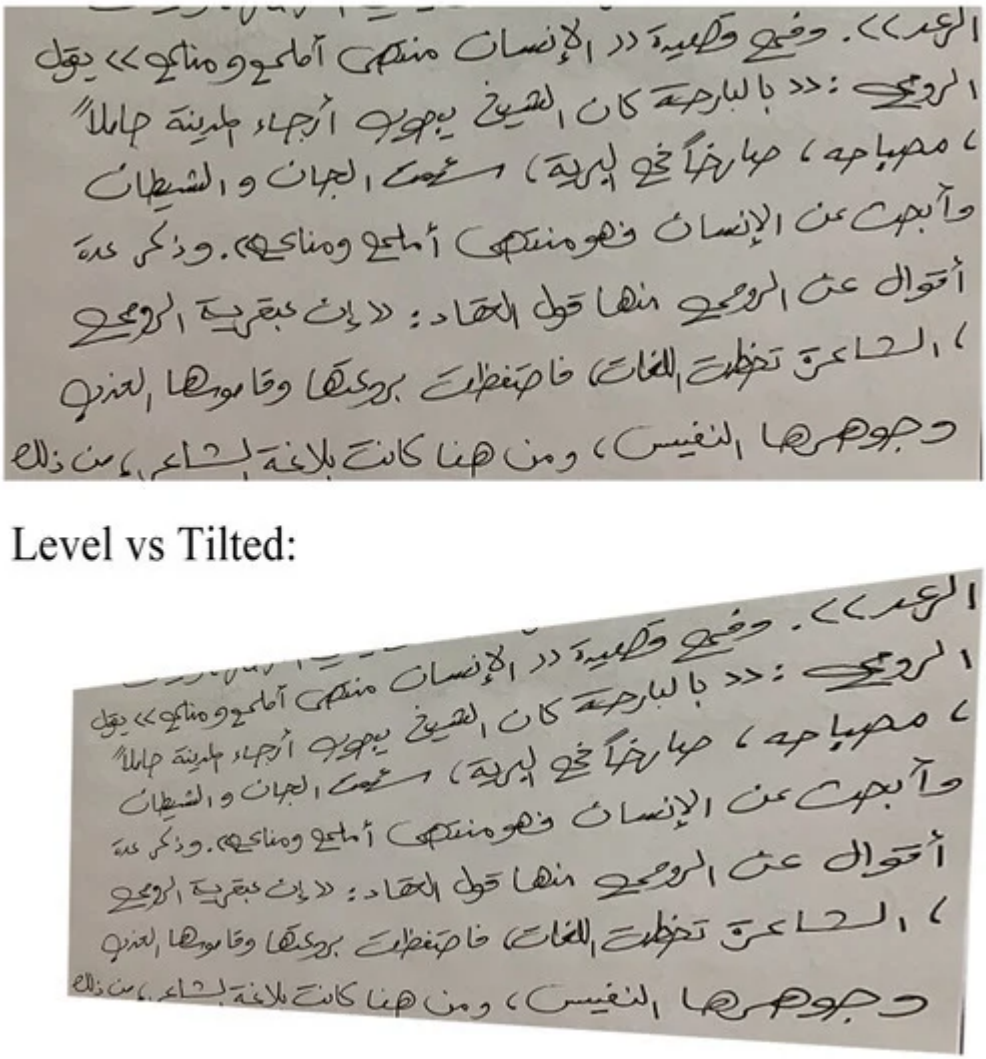

Figure 1. Example of tilting.

3. Arabic OCR Challenges [14][15][16]

- Arabic is written from right to left, while most models are more used to the opposite direction;

- Arabic consists of a minimum of 28 characters, and each character has a minimum of two and up to four shapes, depending on the position. Fortunately, no upper- or lower-case characters are used in Arabic;
- The position determines the shape of a character. Therefore, a character can have different shapes depending on its position: alone, at the start, in the middle, or at the end. For example, the letter Meem has different shapes according to its position;
- Ijam: Dots can be written above or below characters, and they differentiate between one character and another, such as Baa and Tha;
- Tashkeel: Marks such as Tanween and Harakat can change the context of a piece of writing. Examples applied to the letter Raa are illustrated in **Table 1**;
- Hamza, in various locations, such as above or below Alif, above Waw, after or above Yaa, and in the middle or end of a word;
- Loops, such as Saad and Taa, and Meem and Ain, where the difference between characters is difficult to distinguish due to the similarity of the skeleton;
- Separated characters separate words, such as Raa, Dal, and Waw. **Figure 2** presents an example with different kinds of separation;
- The total number of shapes is at least 81. Some letters share the same shape but differ in terms of dots;
- Without dots, there are at least 52 letter shapes;
- The structural characteristics of the characters, given their high similarity in shape, can be perplexing. Moreover, a slight modification can turn one character into another;
- Arabic is cursively written. Therefore, connections are formed when two characters are connected;
- The presence of other diacritics, such as Shadda, and punctuation, such as full stops can be confused with the dots that are used in Arabic;
- Most models suffer from overfitting, since generalization is not a concern, partly because of the lack of annotated datasets and partly because complex models are trained on small sets or sets that are not balanced;
- The available datasets have the issue of distorted and unclear samples;
- The majority of datasets suffer from imbalances, in which the distribution is not even between the samples, which can affect the model’s performance by overfitting to these specific samples;

- The writing style can vary from one author to another. **Figure 3** shows different types of writing style.

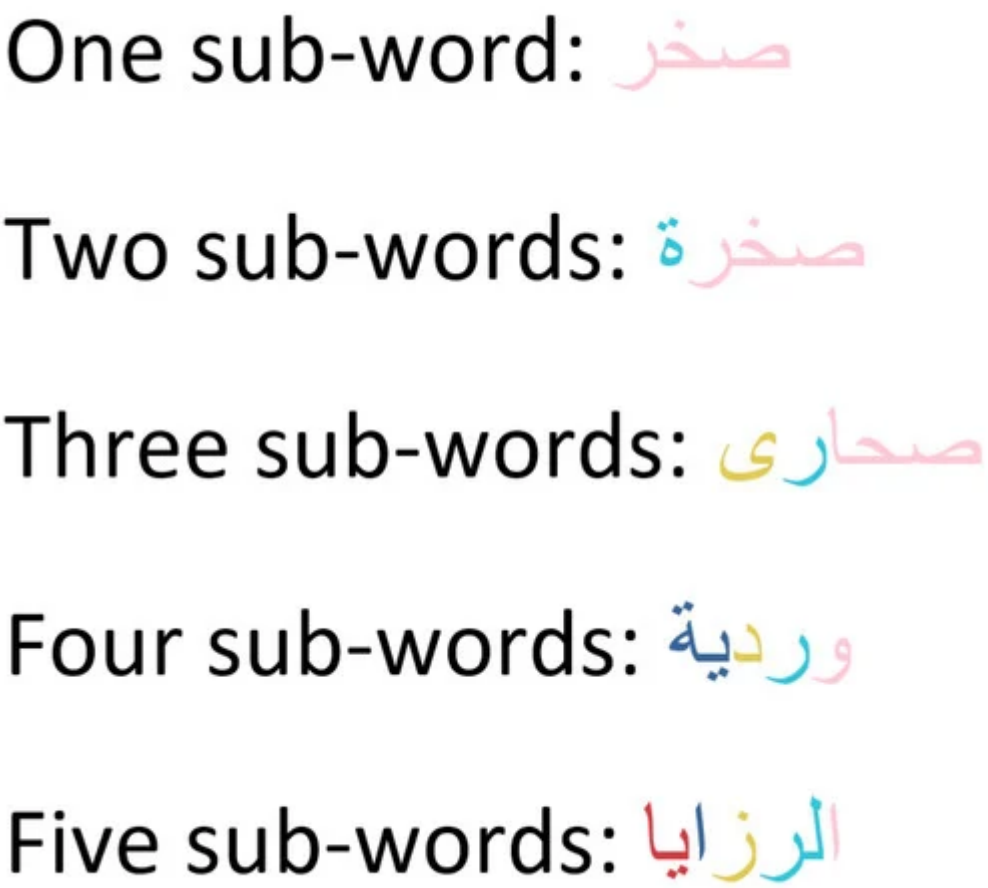

Figure 2. Elucidation of how sub-words are formed in Arabic.

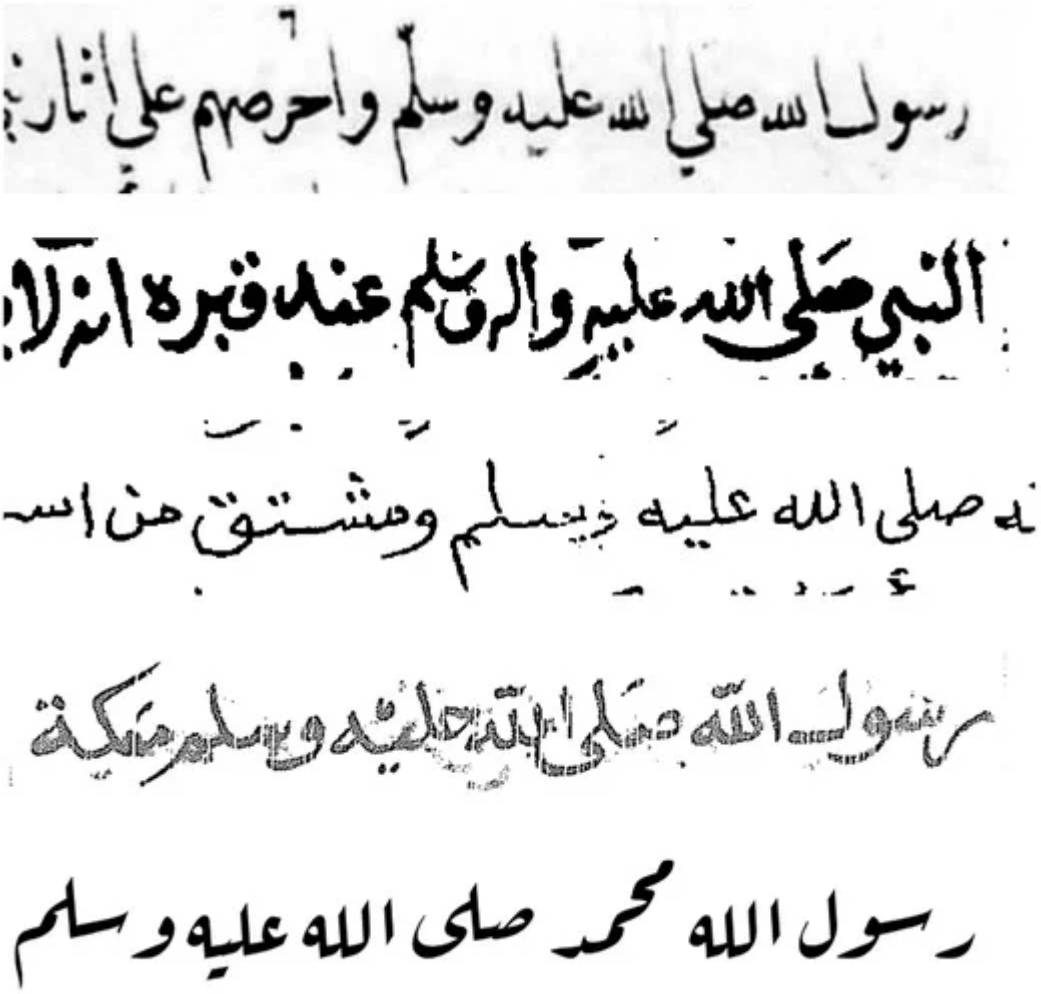

Figure 3. Richness and variation in Arabic fonts.

Table 1. Different types of Tashkeel.

Fatha	Fathatain	Dammah	Dammatain	Kasra	Kasratain	Sukoon	Shaddah
ـَ	ـً	ـُ	ـٌ	ـِ	ـٍ	ـْ	ـّ

References

1. Hsu, T.-C.; Chang, C.; Jen, T.-H. Artificial Intelligence image recognition using self-regulation learning strategies: Effects on vocabulary acquisition, learning anxiety, and learning behaviours of English language learners. Interact. Learn. Environ. 2023, 31, 1–19. [Google Scholar] [CrossRef]
2. Bhattamisra, S.K.; Banerjee, P.; Gupta, P.; Mayuren, J.; Patra, S.; Candasamy, M. Artificial Intelligence in Pharmaceutical and Healthcare Research. Big Data Cogn. Comput. 2023, 7, 10. [Google Scholar] [CrossRef]

3. Ranjan, S.; Sanket, S.; Singh, S.; Tyagi, S.; Kaur, M.; Rakesh, N.; Nand, P. OCR based Automated Number Plate Text Detection and Extraction. In Proceedings of the 9th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 23–25 March 2022; pp. 621–627. [Google Scholar] [CrossRef]
4. Onim, S.H.; Nyeem, H.; Roy, K.; Hasan, M.; Ishmam, A.; Akif, A.H.; Ovi, T.B. BLPnet: A new DNN model and Bengali OCR engine for Automatic Licence Plate Recognition. Array 2022, 15, 100244. [Google Scholar] [CrossRef]
5. Azadbakht, A.; Kheradpisheh, S.R.; Farahani, H. MultiPath ViT OCR: A Lightweight Visual Transformer-based License Plate Optical Character Recognition. In Proceedings of the 12th International Conference on Computer and Knowledge Engineering (ICCKE), Mashhad, Iran, 17–18 November 2022; pp. 92–95. [Google Scholar] [CrossRef]
6. Bi, S.; Wang, C.; Zhang, J.; Huang, W.; Wu, B.; Gong, Y.; Ni, W. A Survey on Artificial Intelligence Aided Internet-of-Things Technologies in Emerging Smart Libraries. Sensors 2022, 22, 2991. [Google Scholar] [CrossRef]
7. Qureshi, F.; Rajput, A.; Mujtaba, G.; Fatima, N. A novel offline handwritten text recognition technique to convert ruled-line text into digital text through deep neural networks. Multimed. Tools Appl. 2022, 81, 18223–18249. [Google Scholar] [CrossRef]
8. Singh, T.P.; Gupta, S.; Garg, M. A Review on Online and Offline Handwritten Gurmukhi Character Recognition. In Proceedings of the 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 13 October 2022; pp. 1–6. [Google Scholar] [CrossRef]
9. Tan, Y.F.; Connie, T.; Goh, M.K.O.; Teoh, A.B.J. A Pipeline Approach to Context-Aware Handwritten Text Recognition. Appl. Sci. 2022, 12, 1870. [Google Scholar] [CrossRef]
10. Ott, F.; Rügamer, D.; Heublein, L.; Bischl, B.; Mutschler, C. Representation Learning for Tablet and Paper Domain Adaptation in Favor of Online Handwriting Recognition. arXiv 2023. [Google Scholar] [CrossRef]
11. Ghosh, T.; Sen, S.; Obaidullah, S.; Santosh, K.; Roy, K.; Pal, U. Advances in online handwritten recognition in the last decades. Comput. Sci. Rev. 2022, 46, 100515. [Google Scholar] [CrossRef]
12. Statista. The Most Spoken Languages Worldwide 2022. Available online: https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide/ (accessed on 6 February 2023).
13. Haddad, B.; Awwad, A.; Hattab, M.; Hattab, A. PRo-Pat: Probabilistic Root–Pattern Bi-gram data language model for Arabic based morphological analysis and distribution. Data Brief 2023, 46, 108875. [Google Scholar] [CrossRef]

14. Al Waqfi, Y.M.; Mohamad, M. A Review of Arabic Optical Character Recognition Techniques & Performance. Int. J. Eng. Trends Technol. 2020, 1, 44–51. [Google Scholar] [CrossRef]
15. Mohd, M.; Qamar, F.; Al-Sheikh, I.; Salah, R. Quranic Optical Text Recognition Using Deep Learning Models. IEEE Access 2021, 9, 38318–38330. [Google Scholar] [CrossRef]
16. Alrobah, N.; Albahli, S. Arabic Handwritten Recognition Using Deep Learning: A Survey. Arab. J. Sci. Eng. 2022, 47, 9943–9963. [Google Scholar] [CrossRef]
17. Moudgil, A.; Singh, S.; Gautam, V. Recent Trends in OCR Systems: A Review. In Machine Learning for Edge Computing; CRC Press: Boca Raton, FL, USA, 2022. [Google Scholar]
18. Avyodri, R.; Lukas, S.; Tjahyadi, H. Optical Character Recognition (OCR) for Text Recognition and its Post-Processing Method: A Literature Review. In Proceedings of the 1st International Conference on Technology Innovation and Its Applications (ICTIIA), Tangerang, Indonesia, 23 September 2022; pp. 1–6. [Google Scholar] [CrossRef]
19. Emon, I.H.; Iqbal, K.N.; Mehedi, H.K.; Mahbub, M.J.A.; Rasel, A.A. A Review of Optical Character Recognition (OCR) Techniques on Bengali Scripts. In Emerging Technologies in Computing; Springer: Cham, Switzerland, 2023; pp. 85–94. [Google Scholar] [CrossRef]
20. Gan, J.; Chen, Y.; Hu, B.; Leng, J.; Wang, W.; Gao, X. Characters as graphs: Interpretable handwritten Chinese character recognition via Pyramid Graph Transformer. Pattern Recognit. 2023, 137, 109317. [Google Scholar] [CrossRef]
21. Singh, S.; Garg, N.K.; Kumar, M. Feature extraction and classification techniques for handwritten Devanagari text recognition: A survey. Multimed. Tools Appl. 2022, 82, 747–775. [Google Scholar] [CrossRef]