From Script to Digital: Arabic Handwriting Recognition for Historical Manuscripts

Hamza Ahmed Abushahla*
Department of Computer Science and Engineering
American University of Sharjah
Sharjah, United Arab Emirates
b00090279@aus.edu

Ariel Justine Navarro Panopio*
Department of Computer Science and Engineering
American University of Sharjah
Sharjah, United Arab Emirates
b00088568@aus.edu

Layth Al-Khairulla*
Department of Computer Science and Engineering
American University of Sharjah
Sharjah, United Arab Emirates
b00087225@aus.edu

Alex Aklson†
Department of Computer Science and Engineering
American University of Sharjah
Sharjah, United Arab Emirates
aaklson@aus.edu

***Abstract*—Handwritten Text Recognition (HTR) for historical Arabic manuscripts is essential for preserving cultural heritage and enabling digital accessibility. However, the cursive nature of Arabic script, its positional letter shapes, and diacritical marks pose unique challenges. Existing HTR systems often fail to generalize to the complexities of historical texts, further hindered by a scarcity of suitable datasets. In this study, we leverage the Muharaf dataset and develop a deep learning-based approach to address these challenges. By employing a convolutional neural network (CNN) architecture, we create an effective HTR system tailored to historical Arabic manuscripts. This work advances the field of Arabic HTR and supports broader efforts to digitize and preserve historical manuscripts for future research.**

***Keywords*—Arabic Handwriting Recognition, Historical Manuscripts, Muharaf Dataset, Arabic Handwriting, Optical Character Recognition, Deep Learning**

I. Introduction

The Arabic language, with its rich history and cultural significance, has played a pivotal role in preserving the intellectual heritage of the Arab world. Historical Arabic manuscripts encompass a wide range of subjects, including literature, science, religion, and personal correspondence. These manuscripts serve as a window into the past, offering invaluable insights into the evolution of thought and culture [1]. Their preservation and digitization are essential for safeguarding this heritage and ensuring accessibility for scholars, historians, and the general public.

Highly accurate Handwritten Text Recognition (HTR) systems for Arabic manuscripts are essential for converting these historical texts into searchable and analyzable formats. Such systems would enable the creation of digital archives, support linguistic and historical research, and facilitate knowledge dissemination. However, existing HTR systems, particularly those designed for modern or non-historical Arabic texts, fail to effectively process these documents [2]. Traditional systems that rely on handcrafted features often struggle to generalize across the diverse styles, varying quality, and structural complexities of historical manuscripts [3]. Moreover, features extracted manually are highly sensitive to noise and degradation commonly found in older documents, limiting their scalability and accuracy.

Handwriting varies significantly between individuals, particularly in cursive scripts like Arabic. The size and style of handwritten characters often differ due to the writer's personal techniques and preferences, creating additional challenges for recognition systems. Hence, Handwriting recognition is considered one of the most challenging tasks in computer vision [4].

Over the past decade, significant advancements in deep learning have revolutionized the field of HTR. Convolutional Neural Networks (CNNs) have become the cornerstone of HTR systems, excelling in feature extraction and image recognition tasks [2], [4], [5]. Their ability to capture spatial hierarchies makes them particularly effective for processing cursive scripts like Arabic. More recently, Transformers and attention-based models have introduced new paradigms [6]–[8], enabling HTR systems to focus on specific regions of an image and capture long-range dependencies. These advances have significantly improved accuracy and generalizability, particularly for complex and diverse handwriting styles.

Despite these advancements, Arabic HTR presents unique challenges. The Arabic script is cursive, with letters that change shape based on their position within a word. The presence of diacritical marks, known as "harakat," adds another layer of complexity. Furthermore, the scarcity of large,

*These authors contributed equally to this work.
†Author to whom correspondence should be addressed.

publicly available datasets for historical Arabic handwriting exacerbates the difficulty of the task [9].

The Muharaf dataset [9], published in June 2024, addresses some of these challenges by providing a comprehensive resource for Arabic handwriting analysis. This dataset consists of 1,644 scanned pages and 36,311 lines of handwritten Arabic text, covering a period from the early 19th to the early 21st century. Fully annotated and transcribed at the text-line level, it encompasses diverse writing styles and offers an invaluable foundation for developing and evaluating HTR systems tailored to historical Arabic manuscripts.

In this study, we leverage the Muharaf dataset to develop a convolutional neural network (CNN)-based deep learning model for handwritten text recognition in historical Arabic manuscripts. We focus on preprocessing, training, and evaluating a robust HTR system while addressing the unique challenges posed by the dataset. In general, our contributions can be summarized as follows:

- We develop a CNN-based deep learning approach for Arabic HTR, specifically tailored to historical manuscripts.
- We design a preprocessing pipeline to handle the challenges of historical handwriting, including variability in quality and styles.
- We evaluate the performance of our model using the Muharaf dataset and provide insights into its strengths and limitations.
- We analyze dataset-specific challenges and discuss future directions for improving HTR systems for Arabic manuscripts.

II. Background

Provides necessary context and theoretical framework related to the research topic. This section should explain foundational concepts, terminology, and any relevant theories or models.

A. Handwritten Text Recognition

Optical Character Recognition (OCR) laid the foundation for modern text recognition systems by enabling the automated extraction of text from printed documents. This field gained momentum in the 1990s [3], with early neural network models like LeNet [10], showcasing significant promise in character classification tasks. These systems primarily targeted machine-printed text and laid the groundwork for extending recognition capabilities to handwritten text, evolving into what is now known as HTR.

HTR has progressed significantly since its early focus on isolated character-level recognition [11], which remains prevalent for logographic languages like Japanese [12] and Chinese [13]. For alphabetical languages such as English and Arabic, handwriting recognition expanded to word-level tasks, where single words are transcribed from handwritten images [14], [15]. More advanced techniques have since moved towards line-level HTR, where entire lines of text, including spaces, are transcribed. Line-level HTR can either rely on pre-segmented input or integrate segmentation and recognition into a unified framework. In recent developments, systems also tackle paragraph- and page-level recognition, incorporating layout analysis for handling complex document structures.

Furthermore, recent advancements focus on line-level HTR, where the aim is to transcribe entire text lines, including spaces, which were often omitted in word-level systems. Line-level recognition can be performed on pre-segmented text [23]–[27] or integrated into joint detection and recognition frameworks where both line segmentation and transcription are done at the same [29].

More advanced systems now address paragraph or page-level recognition [28], combining layout analysis techniques such as paragraph and line segmentation [30]–[33].

The sequential nature of handwriting makes its recognition uniquely challenging. Characters within words are connected, influenced by their context, and require a holistic understanding of the sequence for accurate transcription. Cutting-edge HTR algorithms leverage recurrent architectures, such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, which are designed to capture and model these sequential dependencies effectively. These architectures have been instrumental in advancing HTR by processing input as a series of interconnected data points rather than isolated units, enabling robust handling of cursive and complex scripts.

Among recurrent architectures, Multidimensional Long Short-Term Memory (MDLSTM) networks have emerged as a powerful tool for HTR. Unlike standard LSTMs, which operate along a single axis, MDLSTM networks extend recurrence along two axes, making them particularly effective for two-dimensional inputs like line images in HTR. They excel in extracting features from line-level handwritten text, converting 2D data into 1D sequences for transcription. This capability lies at the core of many state-of-the-art line-level HTR systems. However, MDLSTMs are computationally intensive compared to Convolutional Neural Networks (CNNs), which have become the go-to architecture for efficient and scalable feature extraction.

Recent innovations combine CNNs with RNNs or MDLSTMs to capitalize on their complementary strengths. For example, CNNs extract spatial features from input images, which are then processed by recurrent layers to capture the sequential dependencies of text. These hybrid architectures have achieved significant breakthroughs, with deformable convolutions further enhancing CNN-based models by allowing the convolutional kernel to adapt geometrically. Such approaches address the variability in handwriting styles, reducing transcription errors and setting new benchmarks in HTR accuracy.

Advances in deep learning have driven much of the progress in HTR. Recurrent Neural Networks (RNNs) and their multidimensional variant, MDLSTM (Multidimensional Long Short-Term Memory), are particularly well-suited for sequential data and two-dimensional inputs. MDLSTM networks introduce

recurrence along two axes, making them highly effective for line-level HTR, where 2D data is converted into 1D sequences for character transcription. However, their computational complexity has shifted the focus toward Convolutional Neural Networks (CNNs), which offer a balance of efficiency and accuracy. CNNs extract features from images, often in combination with RNNs, to generate robust text predictions. Recent innovations, such as deformable convolutions, have further enhanced CNN-based models by allowing kernels to adapt geometrically, addressing the variability in handwriting styles. These advancements have established CNNs and hybrid architectures as state-of-the-art approaches in HTR.

The foundation of cutting-edge HTR algorithms lies in recurrent architectures, such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM), which capture the sequential nature of text. Over the last decade, CNNs and, more recently, Transformers have gained prominence for their superior feature extraction and context-awareness capabilities.

B. Characteristics of Arabic Handwriting

Characteristics of Arabic handwriting: cursive nature, positional letter shapes, and diacritics.

C. Challenges in Arabic HTR

Challenges in Arabic HTR and its historical significance. Importance of datasets for advancing HTR research.

III. Related Work

In this section, we review previous research on (), including traditional methods, advancements in deep learning, and the role of Arabic handwriting datasets in advancing the field.

Review of existing HTR systems for Arabic text. Limitations of traditional feature-based approaches in HTR. Advances in deep learning for handwriting recognition. Datasets of Arabic Handwriting: Overview of existing datasets: WAHD, KHATT, and Balamand. Introduction of the Muharaf dataset, its features, and its advantages.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

A.

B.

C. Datasets of Arabic Handwriting

Several datasets have been developed for Arabic handwriting research, (). These datasets differ in focus, size, and availability. Below, we summarize the most relevant ones including WAHD, KHATT, Balamand, and Muharaf.

- **WAHD Dataset [16]:**
  The WAHD dataset is the first dataset explicitly designed for writer analysis tasks in Arabic historical documents. It consists of 353 manuscripts, 333 from the Islamic Heritage Project (IHP) and 20 from the National Library in Jerusalem. Written by 302 writers (23 of them identified), it includes 2,313 pages authored by 11 scribes contributing multiple manuscripts. WAHD is freely available, making it a valuable resource for historical handwriting analysis.
- **KHATT Dataset [17]:**
  The KHATT dataset is a modern (non-historical) Arabic handwriting dataset comprising 4,000 paragraphs written by 1,000 scribes, with six paragraphs contributed by each. It was created mainly for writer identification and was developed jointly by researchers from KFUPM (Saudi Arabia), TU Dortmund (Germany), and TU Braunschweig (Germany). However, its restricted access limits its usability, as it is not publicly available.
- **Balamand Dataset [?]:**
  The Balamand dataset contains 567 historic manuscripts collected from 14 repositories in Lebanon and Syria, including Antiochian Orthodox monasteries and bishoprics, which have been digitized at the University of Balamand. The dataset, spanning the 13th to the 19th century, identifies 256 copyists who produced 329 manuscripts. However, it is not publicly available.

- **Muharaf Dataset [9]:**
  The Muharaf dataset, used in this study, is the largest publicly available Arabic dataset with fully annotated and transcribed historical manuscripts at the text-line level. It includes 1,644 pages (1,216 public and 428 restricted), spanning the early 19th to the early 21st century, with 36,311 text lines (24,495 public). Line-level images, stored in PNG format, were generated using line warping software to create consistent horizontal grids, and extensive metadata is provided in JSON format.
  Although Muharaf was primarily explored for handwriting text recognition (HTR) in previous works [9], [18], our project utilizes it for writer identification, leveraging the existing writer labels (about 25% of the public portion). The dataset contains a rich diversity of handwriting styles, including writings by Arab Levantines living in America, as well as contributors from various other regions. This mix of historical and contemporary writing makes Muharaf a valuable resource for writer identification.

IV. Methodology

A. Data Preparation and Preprocessing

To prepare the data for model training, we utilized the filtered line-by-line images. The following preprocessing steps were applied to ensure consistency and enhance the model's performance...

- Resizing, Normalization, Augmentation

B. Supervised Learning

1) Model Architecture: We implemented a CNN-based architecture...

2) Model Training and Evaluation: The model was trained using a categorical cross-entropy loss function and the Adam optimizer. Early stopping was employed to prevent overfitting.

V. Results and Discussion

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

VI. Conclusions and Future Work

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

References

[1] M. A. Ayuba, "Information and communication technologies in preserving arabic and islamic manuscripts," *Global Journal Al-Thaqafah*, vol. 3, no. 2, pp. 7–14, 2013.

[2] N. Alrobah and S. Albahli, "Arabic handwritten recognition using deep learning: A survey," *Arabian Journal for Science and Engineering*, vol. 47, no. 8, pp. 9943–9963, 2022.

[3] M. T. Parvez and S. A. Mahmoud, "Offline arabic handwritten text recognition: a survey," *ACM Computing Surveys (CSUR)*, vol. 45, no. 2, pp. 1–35, 2013.

[4] N. Altwaijry and I. Al-Turaiki, "Arabic handwriting recognition system using convolutional neural network," *Neural Computing and Applications*, vol. 33, no. 7, pp. 2249–2261, 2021.

[5] L. Mosbah, I. Moalla, T. M. Hamdani, B. Neji, T. Beyrouthy, and A. M. Alimi, "Adocrnet: A deep learning ocr for arabic documents recognition," *IEEE Access*, 2024.

[6] T. Wang, Y. Zhu, L. Jin, C. Luo, X. Chen, Y. Wu, Q. Wang, and M. Cai, "Decoupled attention network for text recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 12 216–12 224.

[7] M. Li, T. Lv, J. Chen, L. Cui, Y. Lu, D. Florencio, C. Zhang, Z. Li, and F. Wei, "Trocr: Transformer-based optical character recognition with pre-trained models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, 2023, pp. 13 094–13 102.

[8] A. K. Bhunia, S. Ghose, A. Kumar, P. N. Chowdhury, A. Sain, and Y.-Z. Song, "Metahtr: Towards writer-adaptive handwritten text recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15 830–15 839.

[9] M. Saeed, A. Chan, A. Mijar, J. Moukarzel, G. Habchi, C. Younes, A. Elias, C.-W. Wong, and A. Khater, "Muharaf: Manuscripts of handwritten arabic dataset for cursive text recognition," *arXiv preprint arXiv:2406.09630*, 2024.

[10] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[11] N. D. Cilia, C. De Stefano, F. Fontanella, and A. S. di Freca, "A ranking-based feature selection approach for handwritten character recognition," *Pattern Recognition Letters*, vol. 121, pp. 77–86, 2019.

[12] T. Clanuwat, A. Lamb, and A. Kitamoto, "Kuronet: Pre-modern japanese kuzushiji character recognition with deep learning," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2019, pp. 607–614.

[13] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," *Advances in neural information processing systems*, vol. 28, 2015.

[14] A. K. Bhunia, A. Das, A. K. Bhunia, P. S. R. Kishore, and P. P. Roy, "Handwriting recognition in low-resource scripts using adversarial learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4767–4776.

[15] F. P. Such, D. Peri, F. Brockler, H. Paul, and R. Ptucha, "Fully convolutional networks for handwriting recognition," in *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE, 2018, pp. 86–91.

[16] A. Abdelhaleem, A. Droby, A. Asi, M. Kassis, R. Al Asam, and J. Elsanaa, "Wahd: a database for writer identification of arabic historical documents," in *2017 1st International workshop on arabic script analysis and recognition (ASAR)*. IEEE, 2017, pp. 64–68.

[17] S. A. Mahmoud, I. Ahmad, W. G. Al-Khatib, M. Alshayeb, M. T. Parvez, V. Märgner, and G. A. Fink, "Khatt: An open arabic offline handwritten text database," *Pattern Recognition*, vol. 47, no. 3, pp. 1096–1112, 2014.

[18] A. Chan, A. Mijar, M. Saeed, C.-W. Wong, and A. Khater, "Hatformer: Historic handwritten arabic text recognition with transformers," *arXiv preprint arXiv:2410.02179*, 2024.