From Script to Digital: Arabic Handwriting Recognition on the Muharaf Dataset

Hamza Ahmed Abushahla B00090279@AUS.EDU
Department of Computer Science and Engineering
American University of Sharjah
Sharjah, United Arab Emirates

Ariel J. N. Panopio B00088568@AUS.EDU
Department of Computer Science and Engineering
American University of Sharjah
Sharjah, United Arab Emirates

Layth Al-Khairulla B00087225@AUS.EDU
Department of Computer Science and Engineering
American University of Sharjah
Sharjah, United Arab Emirates

Abstract

Arabic handwritten text recognition (HTR) presents unique challenges, especially when dealing with historical texts. The diverse writing styles and the complex features of Arabic script contribute to these difficulties. Moreover, Arabic handwriting datasets are generally smaller than English counterparts, limiting the effectiveness of training models that are robust and generalizable. This study leverages the Muharaf dataset, which consists of over 1,600 meticulously transcribed images of historical Arabic handwritten pages, to develop a model aimed at enhancing HTR capabilities for archival Arabic. We explore techniques to improve model performance despite dataset constraints, proposing an approach that balances accuracy and computational efficiency for practical deployment. Our evaluation demonstrates the viability of our methods in advancing Arabic HTR for historical documents, with potential applications in cultural preservation and archival digitization.

Keywords: Arabic Handwritten Text Recognition, Historical Manuscripts, Muharaf Dataset, Arabic Script, Machine Learning

1. Introduction

Arabic handwritten text recognition (HTR) has become a crucial tool in cultural preservation, enabling the digitization and analysis of historical manuscripts. However, Arabic HTR faces unique challenges, particularly when dealing with historical texts, due to the intricacies of Arabic script and variations in writing styles across different eras. These difficulties are compounded by the limited availability of large and diverse Arabic handwriting datasets, which restricts the ability to train highly accurate and generalizable models. In this work, we utilize the Muharaf dataset, a significant resource for Arabic HTR that provides a rich collection of historical Arabic handwritten pages, each carefully transcribed by experts.

Our main contributions can be summarized as follows:

- Conduct a comprehensive evaluation of several state-of-the-art …
- something good
- another something good

The remainder of this paper is structured as follows: Section 2 provides necessary background on Arabic script and historical handwriting challenges. Section 3 reviews related work in Arabic HTR. Section 4 details our methodology, including model design and training strategies. Section 5 presents our evaluation and results, and finally, Section 6 concludes with insights and future directions for Arabic HTR research.

2. Background

Arabic HTR is influenced by several factors intrinsic to the Arabic language and script. Arabic letters change form based on their position within a word, and certain letters connect differently depending on the context. Historical Arabic manuscripts further complicate this due to variations in calligraphic styles, which evolved over centuries. These features make Arabic HTR challenging, especially for systems that were initially designed for Latin-based scripts. Additionally, unlike English HTR, which benefits from abundant large-scale datasets, Arabic HTR is constrained by the scarcity of datasets that are sufficiently large and diverse. The Muharaf dataset helps address this gap by providing historical Arabic handwriting samples, which capture a range of stylistic variations and linguistic nuances. This background context underscores the need for specialized models and training techniques to overcome the limitations posed by smaller datasets and script complexity in Arabic HTR.

2.1. Subsection 1

You should use Times Roman style fonts. Please be very careful not to use nonstandard or unusual fonts in the paper. Including such fonts will cause problems for many printers. Headers and Footers should be in 9pt type. The title of the paper should be in 14pt bold type. The abstract title should be in 11pt bold type, and the abstract itself should be in 10pt type. First headings should be in 12 point bold type and second headings should be in 11 point bold type. The text and body of the paper should be in 11 point type.

3. Related Works

Previous research on Arabic HTR has focused on adapting Latin-based HTR models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), to recognize Arabic script. However, due to the distinct structural features of Arabic handwriting, these models often struggle with accuracy when applied to historical texts. Several recent works have explored alternative approaches, including CNN-BiLSTM architectures, to better capture the sequential nature of Arabic script. Additionally, researchers have introduced pre-processing techniques tailored to Arabic text, such as stroke-width normalization and diacritic removal, to enhance model performance. While the Muharaf dataset represents a significant advancement by offering a specialized dataset for Arabic HTR, much work remains in refining model architectures and training strategies to handle Arabic's script-specific complexities, especially in the context of historical manuscripts.

3.1. Subsection 1

You should use Times Roman style fonts. Please be very careful not to use nonstandard or unusual fonts in the paper. Including such fonts will cause problems for many printers. Headers and Footers should be in 9pt type. The title of the paper should be in 14pt bold type. The abstract title should be in 11pt bold type, and the abstract itself should be in 10pt type. First headings should be in 12 point bold type and second headings should be in 11 point bold type. The text and body of the paper should be in 11 point type.

4. Methodology

Our methodology involves multiple stages, beginning with a detailed pre-processing pipeline tailored to the Muharaf dataset. This includes resizing, normalization, and enhancement techniques that address the specific characteristics of historical Arabic manuscripts, such as faded ink and script inconsistencies. The model architecture used in this study is a convolutional neural network (CNN) combined with a bidirectional long short-term memory (BiLSTM) layer, selected for its ability to capture both spatial and sequential features of the handwritten text. For training, we employed a semi-supervised learning approach, leveraging pseudo-labeling for a subset of the data to overcome limitations in labeled samples. Additionally, we experimented with transfer learning, pre-training our model on existing Arabic HTR datasets before fine-tuning it on Muharaf, which helped enhance performance and robustness. Finally, we selected key evaluation metrics—accuracy, F1 score, precision, and recall—to comprehensively assess the model's effectiveness in recognizing Arabic handwritten text.

5. Evaluation and Results

The evaluation of our model was conducted using a variety of metrics to provide a comprehensive assessment of its performance. We implemented the model on high-performance computing hardware, where training and validation were performed on separate splits of the Muharaf dataset. Our results demonstrate that the model achieves high accuracy and F1 score, indicating effective recognition of Arabic script even with limited training data. To further validate our approach, we compared our model's performance with several baseline models commonly used in HTR for Arabic. In terms of computational efficiency, our model showed promising results, highlighting its potential for deployment in practical applications, such as automated manuscript transcription. Error analysis reveals specific areas where the model struggled, particularly with faded text and less common handwriting styles, suggesting directions for further improvement.

6. Discussion

The results of our study highlight both the strengths and limitations of the proposed model. While our model performed well on standard recognition tasks within the Muharaf dataset, challenges remain when dealing with outliers in writing style or low-contrast images. This underscores the need for additional pre-processing techniques or enhanced data augmentation to improve model robustness. Another notable observation is the model's sensitivity to variations in stroke width, which reflects the diverse stylistic choices in Arabic manuscripts. Addressing these limitations may involve experimenting with more complex architectures, such as attention mechanisms, which could enable the model to focus on specific regions of the text. Overall, this study contributes to Arabic HTR by providing insights into model design and dataset handling that may inform future research and applications in this domain.

7. Conclusion and Future Work

In this paper, we presented a novel approach to Arabic handwritten text recognition using the Muharaf dataset, which comprises a diverse collection of historical Arabic manuscripts. Our findings underscore the feasibility of training effective HTR models even with limited Arabic handwriting data, given the right combination of model architecture and training strategy. Future work may explore expanding the dataset, incorporating additional pre-processing methods, and experimenting with advanced neural network architectures to further improve recognition accuracy. Additionally, we suggest that future studies focus on developing models capable of adapting to a broader range of Arabic historical scripts, ultimately enhancing the accessibility and preservation of Arabic cultural heritage.

References

Include all cited sources, ensuring they cover related studies on Arabic HTR, historical manuscript analysis, the Muharaf dataset, and relevant machine learning techniques.

[1] M. Saeed *et al.*, “Muharaf: Manuscripts of Handwritten Arabic Dataset for Cursive Text Recognition,” Jun. 13, 2024, *arXiv*: arXiv:2406.09630. Accessed: Sep. 25, 2024. [Online]. Available: http://arxiv.org/abs/2406.09630