Text Extraction and Script Completion in Images of Arabic Script-Based Calligraphy: A Thesis Proposal

Dilara Zeynep Gürer and **Ümit Atlamaz** and **Şaziye Betül Özateş**
Boğaziçi University
{dilara.gurer@std., umit.atlamaz@, saziye.ozates@}bogazici.edu.tr

Abstract

Arabic calligraphy carries rich historical information and meaning. However, the complexity of its artistic elements and the absence of a consistent baseline make text extraction from such works highly challenging. In this paper, we provide an in-depth analysis of the unique obstacles in processing and interpreting these images, including the variability in calligraphic styles, the influence of artistic distortions, and the challenges posed by missing or damaged text elements. We explore potential solutions by leveraging state-of-the-art architectures and deep learning models, including visual language models, to improve text extraction and script completion.

1 Introduction

Arabic calligraphy initially emerged as a way to create a visually appealing script, evolving into a highly respected art form. This development was especially prominent during the Ottoman Empire, where calligraphy adorned buildings, mosque foundations, and a variety of other structures. From the reign of Mehmed II onwards, distinct schools of calligraphy began to emerge, gaining further momentum when Sheikh Hamdullah, the founder of the Ottoman calligraphy school, arrived in Istanbul during the rule of Bayezid II. Under Turkish craftsmen, Arabic calligraphy was refined and achieved its most perfect forms (Derman, 1997). This art has experienced significant development in several countries, including Iran, Egypt, Saudi Arabia, and Morocco. However, Istanbul is particularly notable for its diverse and well-established tradition in calligraphy, largely shaped by Ottoman influence. While most of these calligraphic works are in the Arabic language, calligraphy using Arabic letters also appears in other languages such as Urdu, Persian, and Ottoman Turkish. Arabic serves as the common language in these artworks, as it is central to Islamic texts, with many works featuring Quranic verses, Hadith, or prayers (duas) (Gündüz, 1988).

While the artistic dimension of calligraphy is often the main focus, these works also encode significant linguistic, cultural, and historical information. Calligraphy not only conveys Islamic thoughts and architectural aesthetics but also serves as a valuable record of historical and cultural contexts. However, the complexity of calligraphic styles—where sentences vary dramatically in layout and form—often makes them challenging to read, even for those fluent in Arabic. Traditional optical character recognition (OCR) techniques fall short in interpreting such intricate designs, as they are not suited for the overlapping, stylized, or highly curved forms that define calligraphy. Understanding and analyzing these works is crucial for preserving and studying historical and cultural heritage.

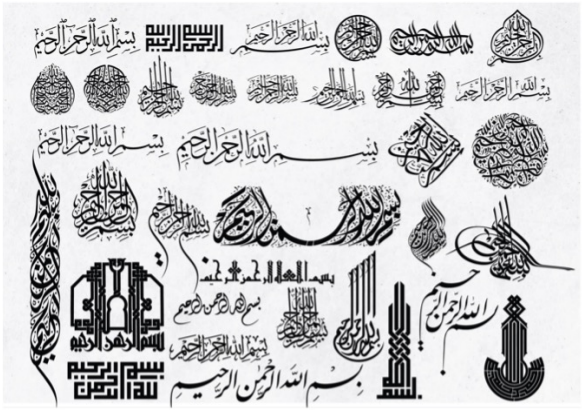

Figure 1: The phrase بِسْمِ ٱللَّهِ ٱلرَّحْمَٰنِ ٱلرَّحِيمِ (In the name of Allah, the Most Gracious, the Most Merciful) in different styles and with different letter combinations.

Figure 1 depicts the phrase بِسْمِ ٱللَّهِ ٱلرَّحْمَٰنِ ٱلرَّحِيمِ (In the name of Allah, the Most Gracious, the Most Merciful) in different styles and with different letter combinations. The phrase consists of five words with twelve distinct letters, but the calligraphic styles vary in

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop), pages 253–259

how these letters are connected, arranged, and stylized through different ligatures and artistic compositions. As can be seen in the figure, same sentence in various calligraphic styles and arrangements is challenging to recognize—even identifying the beginning of the sentence can be difficult due to the non-standard layouts and artistic variations.

In this thesis proposal, we outline a specialized methodology designed to analyze and interpret documents and images featuring Arabic script-based calligraphy with high accuracy. This approach aims to bridge the gap between visual artistry and textual extraction, enabling both aesthetic appreciation and understanding of these culturally significant texts.

Thesis contributions will be as follows:

- A system for efficiently labeling datasets with limited annotated samples (scarce datasets), employing semi-supervised learning and transfer learning techniques. The system is designed to generalize well from small datasets and can be adapted to scale up when larger datasets become available.
- A rich dataset of real-world calligraphy images, annotated at the letter and word levels, encompassing diverse calligraphic styles to support tasks like style recognition, text extraction, and sentence reconstruction.
- An optimized pipeline for noise removal and artifact handling tailored for historical and architectural calligraphic content. This pipeline will incorporate deep learning techniques to enhance textual clarity, eliminate ornamental noise, and reconstruct missing portions of letters for improved accuracy.
- Implementation of advanced recognition techniques, leveraging architectures like Visual Question Answering (VQA) and Large Language Models (LLMs), designed to handle the artistic and structural intricacies of Arabic calligraphy while extracting textual content with contextual awareness.

2 Background

This research explores Arabic script-based calligraphy analysis, emphasizing its intricate and artistic nature. To develop a comprehensive approach, we reviewed related studies in Arabic script analysis, calligraphic works, and non-Arabic calligraphy.

Arabic character recognition is challenging due to its cursive and context-sensitive script. Traditional methods like HOG with SVM achieved over 99% recognition on a dataset of 43,000 handwritten words (Jebril et al., 2018), while Random Forests, KNN, and MLP attained 100% accuracy on a dataset of 600 images across 28 classes (Boufenar et al., 2018). Though designed for handwriting, these approaches could aid in labeling calligraphy datasets.

Deep learning methods have further improved recognition. CNNs with LReLU achieved 99% accuracy by mitigating overfitting (Nayef et al., 2022), and foundation models like Qalam have advanced Arabic OCR and handwriting recognition capabilities (Bhatia et al., 2024). Qalam, combining a SwinV2 encoder and a RoBERTa decoder, excels in Arabic script recognition. Trained on over 4.5 million manuscript images and 60,000 synthetic pairs, it achieves a WER of 0.80% in handwriting recognition and 1.18% in OCR tasks. Its support for diacritics and high-resolution inputs addresses limitations of many OCR systems. However, while useful for initial insights, these methods are less suited to the artistic variability and baseline inconsistencies in Arabic calligraphy images.

Arabic calligraphy has been the focus of several research efforts, each contributing to the field with unique datasets, methodologies, and findings. The Calliar dataset (Alyafeai et al., 2022) is a comprehensive resource, featuring 2,500 sentences and over 40,000 strokes. It covers multiple levels—stroke, character, word, and sentence—and includes styles like Diwani, Thuluth, and Farisi, enabling tasks such as style classification, character recognition, and calligraphy generation.

The Salamah dataset (AlSalamah and King, 2018) contains 3,467 images across 32 categories of Arabic calligraphic letters, representing diverse styles. Kaoudja et al. (2021) developed feature descriptors tailored to specific calligraphy styles, achieving superior performance on a dataset of 1,685 images across nine styles compared to existing methods, including deep learning approaches. A complementary study (Gürer and Gökbay, 2024) analyzed two datasets for classification tasks, achieving F1 scores of 90% for style classification and 79% for letter classification using transfer learning techniques.

Efforts in content recognition remain limited despite valuable datasets and classification studies. A study (Alsalamah, 2020) on a dataset of 388

images achieved 75% accuracy in mapping calligraphic images to their corresponding quotations, highlighting the challenges in recognizing content from artistic calligraphy and underscoring the need for more advanced methodologies.

Generative Adversarial Networks (GANs) have been used to synthesize calligraphic styles like Nastaliq, blending traditional and contemporary elements (Sobhan et al., 2024). Similar methods in non-Arabic calligraphy, such as Chinese and Japanese scripts, have applied CNNs and transformers for style recognition and glyph generation (Wen and Sigüenza, 2020; Zhang et al., 2024; Aguilar, 2024; Wong et al., 2024; Kuwata et al., 2024).

While these methods offer valuable insights, their direct use for Arabic calligraphy content recognition is limited due to the unique features of Arabic script. However, the strategies in these studies can guide the development of tailored approaches for Arabic calligraphy.

3 Research Goals and Questions

The main goal of this thesis is to analyze Arabic calligraphy images to accurately extract the script contained within them.

Our main research question is:

RQ: What are the optimal methods for accurately extracting and reconstructing text from Arabic calligraphy images, considering the unique artistic and structural challenges?

We examine the challenges arising from the artistic elements and absence of consistent baselines, seeking methods to extract letters, phrases or complete sentences with high accuracy. To visually represent the methodology for tackling these challenges, Figure 2 illustrates the step-by-step process for analyzing and extracting text from Arabic calligraphy images.

We outline three objectives to answer our research question: (i) **Data Collection** - Gathering an extensive dataset of Arabic calligraphy images. (ii) **Text Extraction** - Developing methods to accurately extract textual data from images. (iii) **Script Completion** - Reconstructing incomplete scripts when necessary. In the following sections, we define sub-questions for these objectives.

3.1 Data Collection

A key challenge in this area is the lack of comprehensive datasets. While there is only one publicly available dataset (Alyafeai et al., 2022), obtaining accurate results will require more diverse, rich and versatile dataset. The existing dataset presents two main issues. First, all images are digitalized, meaning they do not represent the real-world images often encountered in historical or architectural contexts. As shown in Figure 3, these digitalized samples lack background noise, such as decorative elements or embellishments, which are often present in real-world calligraphy and can obscure or blend with the text. This results in clean images that do not fully capture the challenges of authentic calligraphy analysis.

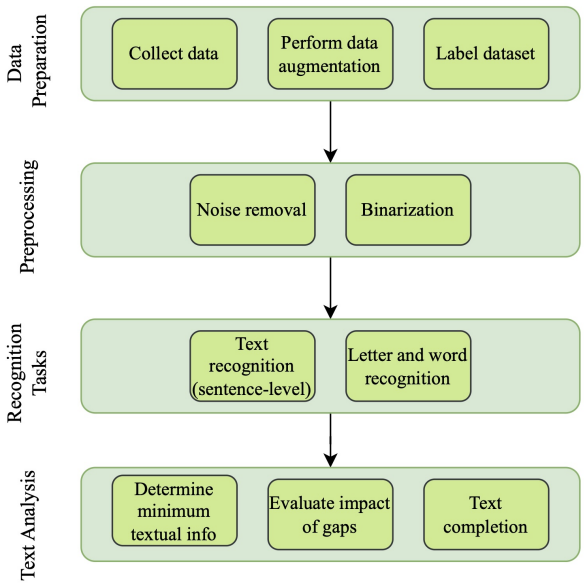

Figure 2: Flowchart of the proposed research.

Figure 3: Examples from the existing dataset (Alyafeai et al., 2022).

Second, our analysis of the dataset revealed that each piece of content appears only a few times, with unique sentences and phrases that lack variations or alternative versions. For effective machine learning, however, the dataset needs multiple versions of each sentence or phrase to better train models in recognizing different stylistic forms of the same text. Although the orientations in the dataset vary, we also need examples with more diverse letter combinations, structural complexity, and re-

alistic variations. Ultimately, a more representative dataset is essential for building an end-to-end architecture that can handle the nuances of Arabic calligraphy in various forms.

RQ1 How do we obtain authentic data for investigating the main research question? Istanbul's rich calligraphic heritage offers a unique advantage for data collection. We will gather images through on-site visits to historical sites, mosques, and architectural landmarks, supplemented by publicly shared photos from tourists and researchers. Since all these places are open to the public, taking photographs does not require special permissions. Additionally, we will meet with the owners of these photos to ensure proper context and information. To ensure diversity, our dataset will capture varied artistic styles and layouts. Image quality may vary due to factors such as different angles and lighting conditions, but this variability will benefit the model's training. The goal is for the model to be able to process and recognize calligraphy accurately, regardless of these variations, when it encounters new images during future data collection or use. Essentially, the model will be robust enough to perform well even if the conditions of the new images differ from the ones it was originally trained on.

Since most calligraphic works in Istanbul are in Arabic and Ottoman Turkish, our initial focus will be on these languages, which are central to Islamic calligraphy. This approach is particularly significant as Arabic serves as the lingua franca of the Islamic faith. Once a solid foundation is established, we will expand to Persian and Urdu for broader linguistic coverage. Additionally, we will source images from websites with proper permissions.

We have access to a comprehensive 136,000-page textual archive that explores the history, evolution, and artistic styles of Arabic calligraphy, along with its reading techniques, cultural significance, and traditional methodologies. This archive includes scholarly analyses, historical manuscripts, and instructional texts that provide deep insights into the art form. Before training our VQA model, we will fine-tune the language component—not the visual part—of a visual language model using text from these books. This will enhance the model's understanding of calligraphy's artistic and textual nuances.

RQ2 How should the collected data be labeled? The process will begin with the creation of an image-caption dataset, where each image will be paired with a caption describing the text within it. This dataset will be generated using digital tools such as web scraping techniques with BeautifulSoup or Selenium to gather digitized calligraphic images from web sites and manually collecting and labeling images from printed or physical sources. Once the image-caption dataset is established, the next step will involve labeling individual letters and words within the calligraphic images.

For this task, we will leverage an existing small dataset of online handwritten Arabic calligraphic letters and words. Although this small dataset is in the online handwriting format—where temporal stroke data is recorded—the dataset we will collect is in the offline handwriting format, derived from scanned or photographed calligraphic texts. To bridge this gap, we will use the online dataset to inform and guide the labeling process for offline data. By training the model on the online handwriting data first, it will gain a foundational understanding of Arabic calligraphic structures, which will then be applied to label offline handwritten datasets effectively. This alignment between the two formats will allow for a more robust and comprehensive training process.

Using semi-supervised learning techniques, the model will initially be trained on the small, labeled dataset. With pseudo-labeling, it will then generate labels for a larger set of unlabeled offline images. This hybrid approach will enable the model to learn both letter recognition and word recognition from the offline calligraphic images while leveraging the detailed structure of the online handwriting data. This process not only increases the dataset's size and diversity but also will enhance its applicability to real-world offline calligraphic texts.

RQ3 How to enrich the dataset to make it more comprehensive? To simulate the diverse artistic styles and orientations found in Arabic calligraphy, we plan to use data augmentation techniques such as rotation and scaling, allowing us to create variations within a structured dataset.

3.2 Text Extraction

To extract textual data, noise removal is essential as the first step. This leads us to a new sub-question:

RQ4 How can we effectively remove noise from the images? Identifying and removing unwanted elements, such as decorations, background patterns, and ornamental designs, is crucial for this sub-task. However, critical elements like diacritical marks in

the text should not be treated as noise, as they are essential for accurate interpretation.

Template matching can be used to identify and remove consistent unwanted elements, such as decorative patterns, across images. Morphological operations, such as erosion and dilation, help eliminate small artifacts and background patterns while preserving the main calligraphic features. Additionally, Region of Interest (ROI) detection algorithms can focus on the text areas, removing surrounding noise and ensuring the calligraphy is the primary focus.

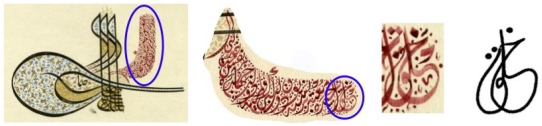

Figure 4: Example of an Arabic calligraphy artwork, from left to right: original artwork, region of interest highlighting the text, removal of unwanted noise and ornamental elements, and the final digitalized form retaining essential diacritical marks for accurate interpretation[1].

After addressing the challenges of noise removal, the next stage in text extraction involves exploring recognition and preprocessing methods that enable accurate analysis of calligraphic images. Consequently, we pose the following questions:

RQ5 Which recognition method is most effective for analyzing the text? Handling the overlap or extreme curvature of letters in Arabic calligraphy is complex and requires advanced segmentation. Arabic calligraphy often merges letters into intricate shapes, which traditional OCR systems may struggle to interpret. We plan to explore different recognition approaches, including character-level, word-level, and sentence-level recognition, to determine which method best captures the artistic variations in the text. To quantitatively evaluate the performance of these recognition methods, we will use several metrics. We will measure the accuracy of text extraction at different linguistic levels using Character Error Rate (CER) and Word Error Rate (WER). Additionally, we will assess the similarity between recognized text and ground truth using Levenshtein Distance.

For baseline comparisons, we will evaluate our approach against existing OCR systems, as well as modern handwriting recognition models such as transformer-based OCR architectures. Furthermore, we will compare our results to human-labeled transcriptions to establish an upper-bound accuracy for the recognition tasks. By incorporating both automated metrics and comparative benchmarks, we aim to identify the most effective recognition method that preserves the accuracy and readability of Arabic calligraphy text.

RQ6 Is preprocessing necessary? As discussed in the noise removal section, it may not be essential to eliminate decorative elements ("noise") entirely. To assess this, we will compare different methods by first testing the original images without noise preprocessing. We will also experiment with state-of-the-art visual question-answering models, using targeted prompts such as "focus on the text in the given image." To assess this, we plan to freeze the visual processing part of a multimodal model, such as BLIP-2 or LLaVA, and focus training on the language components using texts from books on Arabic calligraphy. This will align the model's understanding of the text with the linguistic and contextual nuances of calligraphy. This will allow us to improve the model's understanding of the textual and contextual features of calligraphy. After that, we will incorporate the image-text dataset for further fine-tuning. For this purpose, we will leverage advanced visual-language models such as BLIP-2 and LLaVA, which combine powerful image processing capabilities with language models, enabling them to interpret and understand intricate calligraphic text effectively.

3.3 Script Completion

The extracted data from the previous step may be incomplete due to segmentation failures or unreadable parts of phrases. Additionally, historical documents may be damaged, with some sections missing or too degraded to analyze directly. In such cases, further steps are required to reconstruct the content. We plan to utilize a large language model, trained on Islamic texts such as the Quran and Hadith, to complete sentences or phrases when extracted letters are incomplete.

We will employ several metrics to evaluate the performance of our script completion approach. Reconstruction accuracy will measure how accurately the model completes missing text based on context, compared to ground-truth transcriptions. CER and WER will also quantify the accuracy of the

[1] https://www.ketebe.org/eser/8111?ref=artworks

[2] https://tr.ucoin.net/coin/ottoman_empire-100-para-1808/?tid=83518

Figure 5: Example of images: left, an Ottoman coin with worn or incomplete calligraphic text; right, a wall with partially damaged calligraphic text, illustrating the challenges of dealing with incomplete or unreadable content in historical artifacts[2].

reconstructed text. Contextual accuracy will assess whether the completed sentences align with the linguistic and cultural context of Islamic texts. For baseline comparisons, we will evaluate our model against existing text completion and restoration techniques used for historical documents, including context-based generation models and traditional rule-based methods, to measure the improvement brought by our language model fine-tuned on Islamic texts.

RQ7 What is the minimum required information to understand the content of the images? We plan to analyze the impact of missing letters or words on sentence completion in Arabic scripts by testing different letter sets to reconstruct incomplete phrases or sentences. The goal is to evaluate the accuracy of the reconstructed text by comparing it to the expected phrase. This will involve using a trained model, such as Qalam, which leverages unique features of Arabic script, including its cursive and diacritic-rich structure. These models will be employed to predict and complete missing elements. The model will be assessed based on how well it fills gaps and ensures the reconstructed sentence or phrase is contextually and linguistically accurate. This approach will help improve extraction by reducing reliance on segmentation and allowing more robust handling of incomplete or damaged calligraphic text.

4 Conclusion

This proposal outlines steps to extract textual data from Arabic calligraphy images, addressing challenges from the language's unique features and the art's complexity. It focuses on reconstructing incomplete or damaged calligraphy using Arabic-specific language models to enhance recognition accuracy. The research aims to develop tools for Arabic calligraphy text recognition, benefiting areas such as historical document preservation and cultural heritage digitization. While primarily focused on text, it also acknowledges the importance of calligraphic styles and structures. It lays the groundwork for scalable, precise models capable of handling diverse calligraphic styles.

To respect the cultural and historical significance of Arabic calligraphy, the research will involve consultations with domain experts in art history and cultural heritage. This will ensure the accuracy and sensitivity of interpretations, while following best practices in digitization and preserving the integrity of these valuable artifacts.

References

Sergio Torres Aguilar. 2024. Handwritten text recognition for historical documents using visual language models and GANs. Hal-04716654v2.

Seetah Alsalamah. 2020. *Combining Image and Text Processing for the Computational Reading of Arabic Calligraphy*. Ph.D. thesis, The University of Manchester.

Seetah AlSalamah and Ross King. 2018. Towards the machine reading of Arabic calligraphy: A letters dataset and corresponding corpus of text. In *2018 IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR)*, pages 19–23.

Z. Alyafeai, M. S. Al-Shaibani, M. Ghaleb, et al. 2022. Calliar: an online handwritten dataset for Arabic calligraphy. *Neural Computing and Applications*, 34:20701–20713.

Gagan Bhatia, El Moatez Billah Nagoudi, Fakhraddin Alwajih, and Muhammad Abdul-Mageed. 2024. Qalam: A multimodal LLM for Arabic optical character and handwriting recognition. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 210–224, Bangkok, Thailand. Association for Computational Linguistics.

Chaouki Boufenar, Adlen Kerboua, and Mohamed Batouche. 2018. Investigation on deep learning for off-line handwritten Arabic character recognition. *Cognitive Systems Research*, 50:180–195.

M. U. Derman. 1997. Hat. [Accessed: 28 November 2024].

H. Gündüz. 1988. Türk hat sanatında Şeyh Hamdullah ve Ahmed Karahisari ekolleri. Master's thesis, Mimar Sinan Fine Arts University.

Dilara Zeynep Gürer and İnci Zaim Gökbay. 2024. Arabic calligraphy images analysis with transfer learning. *Electrica*, 24(1):201–209.

N. A. Jebril, H. R. Al-Zoubi, and Q. Abu Al-Haija. 2018. Recognition of handwritten Arabic characters using histograms of oriented gradient (hog). *Pattern Recognition and Image Analysis*, 28:321–345.

Zineb Kaoudja, Mohammed Lamine Kherfi, and Belal Khaldi. 2021. A new computational method for Arabic calligraphy style representation and classification. *Applied Sciences*, 11(11).

Wakana Kuwata, Ryota Mibayashi, Masanori Tani, and Hiroaki Ohshima. 2024. Glyph generation for Japanese calligraphy based on encoding both content and style. In *2024 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 207–214.

B. H. Nayef, S. N. H. S. Abdullah, R. Sulaiman, et al. 2022. Optimized leaky relu for handwritten Arabic character recognition using convolution neural networks. *Multimedia Tools and Applications*, 81:2065–2094.

Arshia Sobhan, Philippe Pasquier, and Adam Tindale. 2024. Unveiling new artistic dimensions in calligraphic Arabic script with generative adversarial networks. *Proc. ACM Comput. Graph. Interact. Tech.*, 7(4).

Yuanbo Wen and Juan Alberto Sigüenza. 2020. Chinese calligraphy: Character style recognition based on full-page document. In *Proceedings of the 2019 8th International Conference on Computing and Pattern Recognition*, ICCPR '19, page 390–394, New York, NY, USA. Association for Computing Machinery.

Adam Wong, Joseph So, and Zhi Ting Billy Ng. 2024. Developing a web application for Chinese calligraphy learners using convolutional neural network and scale invariant feature transform. *Computers and Education: Artificial Intelligence*, 6:100200.

Yuqing Zhang, Baoyi He, Yihan Chen, Hangqi Li, Han Yue, Shengyu Zhang, Huaiyong Dou, Junchi Yan, Zemin Liu, Yongquan Zhang, et al. 2024. Philogpt: A philology-oriented large language model for ancient Chinese manuscripts with dunhuang as case study. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2784–2801.