NAACL Student Research Workshop (SRW) 2025

69-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, İstanbul,Türkiye

1.Introduction

- Calligraphy is the art of beautiful, expressive writing
- Found in historical buildings, and Islamic manuscripts
- Conveys Islamic thought and cultural history
- Traditional OCR struggles with overlapping and stylized text

وَرَبَّكَ فَكَبِّرْ

Fig. 1 An example of Arabic calligraphy.
By Aydın Kızılyar and Berna Karabulut

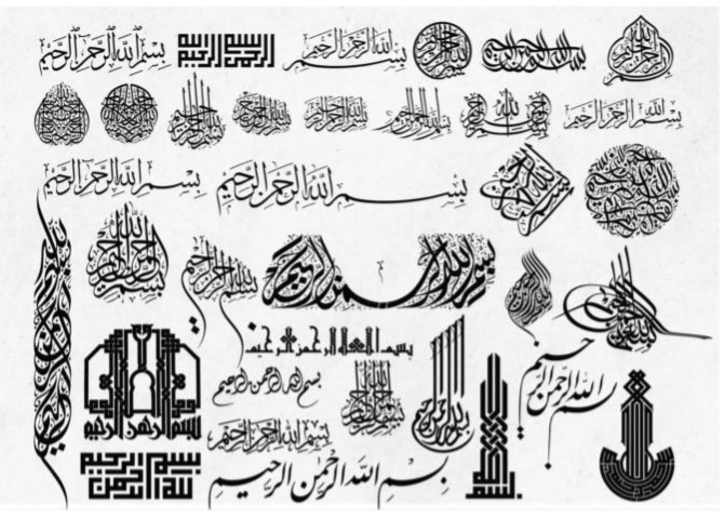

Fig 2. The phrase بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ
(In the name of Allah, the Most Gracious, the Most Merciful) in different styles and with different letter combinations.

- As shown in the figure, recognizing the same sentence in different calligraphic styles is challenging, with non-standard layouts making even the start of the sentence hard to identify.
- In the literature, only one work addresses text extraction, but it suffers from limited data, leading to unsatisfactory results.

2.Research Goals

- Due to the limited amount of available data and the lack of extensive research in this area, our main research question is: **What are the optimal methods for accurately extracting and reconstructing text from Arabic calligraphy images, considering the unique artistic and structural challenges?**
- To address this question, the proposed research is outlined in the flowchart below.

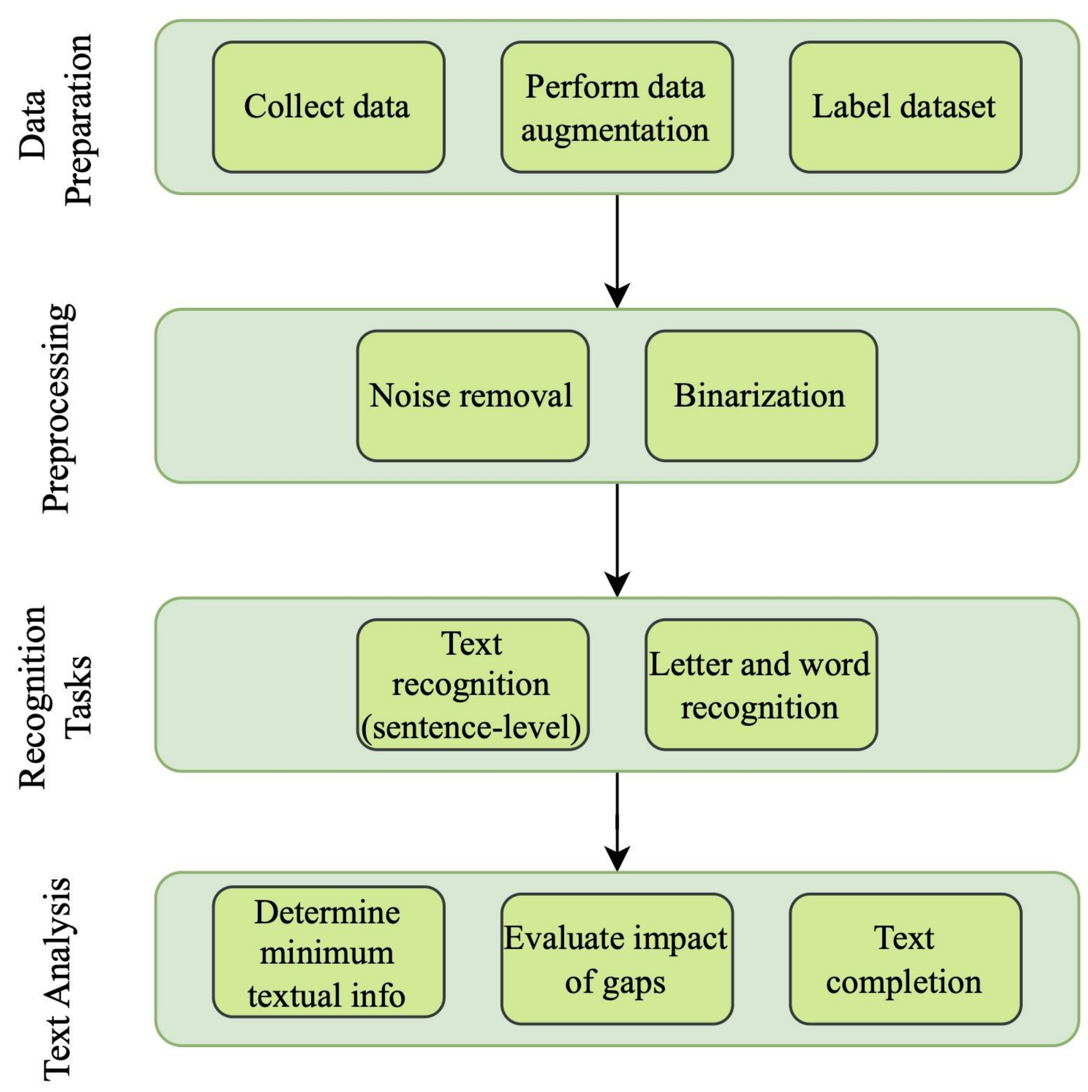

Fig 3. Flowchart of the proposed research

3.Research Questions

To carry out all of these steps in the flowchart , we break down the main problem into several sub-questions that guide our approach.

RQ1 How do we obtain authentic data for investigating the main research question?

- Web & On-site Image Collection
- Arabic & Ottoman Turkish Focus
- Persian & Urdu Expansion
- 136K-Page Archive Utilization

RQ2 How should the collected data be labeled?

- Automated Labeling: Web-sourced data
- Manual Labeling: Physical sources
- Image-Text Dataset
- Online Dataset Training: Guide offline text labeling
- Semi-Supervised Learning: Expand dataset with pseudo-labeling

RQ3 How to enrich the dataset to make it more comprehensive?

- Artistic Variation: Simulate diverse calligraphy styles
- Data Augmentation: Rotation, scaling, and more
- Dataset Enrichment: Structured variations for robustness

RQ4 How can we effectively remove noise from the images?

We will work on removing noise for text extraction.

As shown in Figure 4:

•The first image shows the original artwork.

•The second highlights the region of interest, focusing on the text.

•The last two images show three letters and their digitized form, demonstrating the removal of unwanted noise and decorations while keeping essential diacritical marks for accurate interpretation.

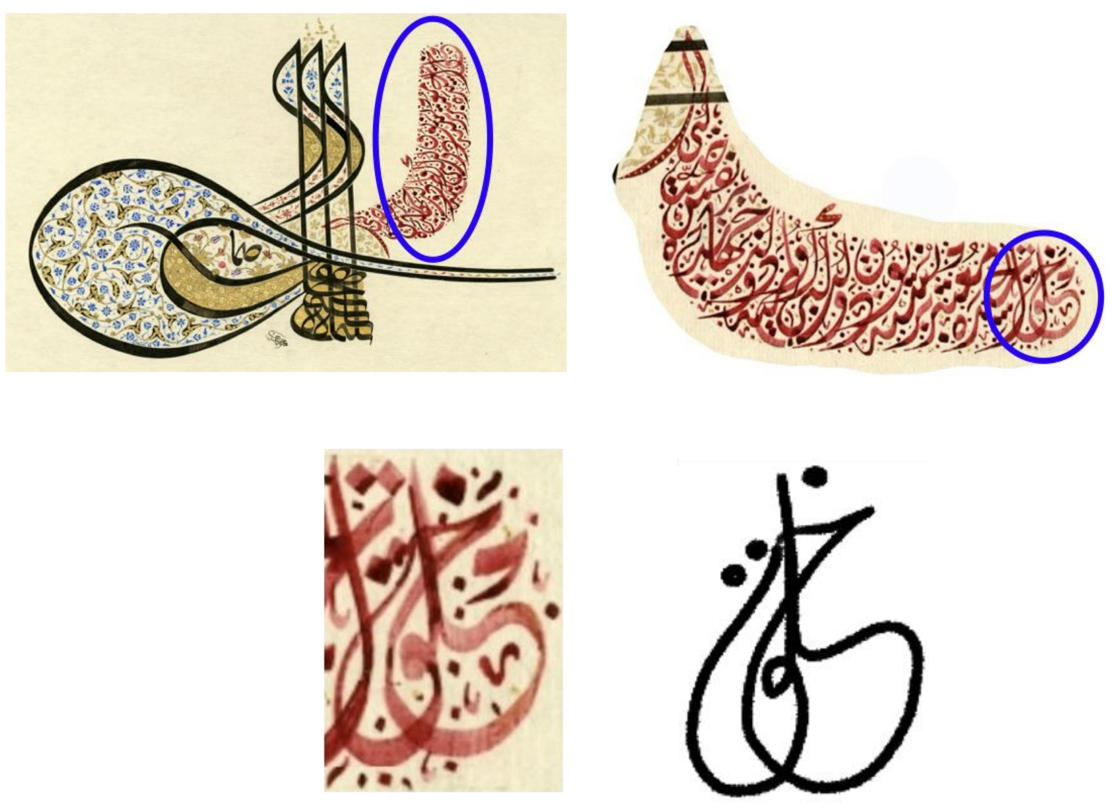

Fig 4. Example of an Arabic calligraphy artwork and steps for removing noises

RQ5 Which recognition method is most effective for analyzing the text?

We'll test character, word, and sentence-level approaches to identify the most effective one.

- Approach: Character, word, and sentence-level recognition
- Metrics: CER, WER, Levenshtein Distance for evaluation
- Comparison: Baseline OCR & transformer-based models

RQ6 Is preprocessing necessary?

We're also exploring whether decorative elements should be removed.

- Decorative Elements: Evaluate need for complete noise removal
- Language Training: Freeze visual components, train language part of VQA (LLaVa)
- Fine-Tuning: Refine models with image-text datasets
- Testing: Test VQA models on original images without noise removal

RQ7 What is the minimum required information to understand the content of the images?

- As a final step, we'll test sentence completion by reconstructing missing letters or words in calligraphic images.
- This supports text extraction from damaged documents, coins, or walls shown in the Fig 5.
- We'll compare results to the original text to measure accuracy—aiming for better recognition without perfect segmentation.

Fig 5. Examples; 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

4.Conclusion

- This research addresses the challenge of extracting text from Arabic calligraphy by combining linguistic insight with artistic sensitivity.
- We propose a reconstruction approach using Arabic-specific language models to improve recognition of incomplete text.
- This lays the groundwork for scalable systems that support cultural heritage preservation.

5.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 offline 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 S¸eyh Hamdullah ve Ahmed Karahisari ekolleri. Master's thesis, Mimar Sinan Fine Arts University.

Dilara Zeynep Gürer and ˙Inci 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.