SAND: A Large-Scale Synthetic Arabic OCR Corpus for Vision-Language Models

Anonymous ACL submission

Abstract

Arabic Optical Character Recognition (OCR) plays a vital role in digitizing Arabic text, yet existing datasets are often limited in scale, diversity, and structured formatting. Most available datasets focus on either printed or handwritten text, lacking the scalability and controlled variation needed for training robust OCR models on book-style documents. To address this gap, we introduce SAND (Large-Scale Synthetic Arabic OCR Dataset), a large-scale, synthetically generated Arabic OCR dataset designed to reflect real-world book formatting. SAND comprises 743,000 document images containing 662.15 million words, spanning five distinct Arabic fonts to enhance typographic diversity and generalization. Unlike traditional datasets, Unlike traditional datasets, SAND offers a scalable and structured resource for training OCR and vision-language models. Its synthetic nature ensures controlled variation in fonts, formatting, and text structures while eliminating common artifacts such as noise, blur, and distortions. This enhances OCR model generalization across diverse Arabic text styles.

1 Introduction

The digitization of Arabic printed materials presents a significant challenge due to the vast number of books and documents that exist in non-text formats, such as scanned PDFs and images. These documents, while rich in historical and scientific value, remain largely inaccessible for text-based search, analysis, and automated processing (Al-Sheikh et al., 2020). Optical Character Recognition (OCR) serves as a crucial technology to bridge this gap by converting scanned Arabic text into machine-readable format, enabling document digitization, automated mail sorting, signature verification, product labeling, and machine vision tasks (Doush et al., 2018). However, OCR systems often encounter errors due to variations in font styles, document quality, and text distortions.

Recent advancements in deep learning, particularly vision-language models, have significantly improved OCR accuracy by leveraging powerful feature extraction and contextual understanding (Najam and Faizullah, 2023). These models require extensive training on well-structured datasets that pair images with their corresponding textual representations. A high-quality dataset is essential for enabling OCR models to learn the complexities of Arabic script, including its cursive nature, contextual letter variations, and diverse typographic styles, ultimately enhancing recognition performance across different document formats (Mosbah et al., 2024).

The Arabic language, spoken by millions and serving as the official language in 26 countries, ranks among the most widely used languages on the internet (UNESCO, 2024). Despite its global significance, the availability of large-scale Arabic OCR datasets remains limited, where most available datasets primarily consist of printed or handwritten text, leaving a gap in synthetically generated large-scale Arabic OCR datasets that can offer controlled, diverse, and scalable training resources for modern OCR models.

To address this gap, we introduce SAND, a large-scale synthetic Arabic OCR dataset designed to reflect real-world book formatting while providing controlled typographic and linguistic diversity. Unlike traditional printed or handwritten datasets, SAND is entirely synthetically generated, ensuring scalability where it can be infinitely expanded by generating additional text samples with diverse structures and noise-free training data. Unlike scanned images, SAND eliminates distortions such as blur, poor lighting, or paper texture artifacts, and lastly, it improves OCR generalization since it enables OCR models to learn from multiple fonts and structured layouts, enhancing real-world performance.

SAND consists of 743,000 document images containing 662.15 million words, covering five diverse Arabic fonts to enhance generalization across different typographic styles. The dataset is structured to simulate real-world book formatting, ensuring that OCR models trained on SAND can handle varying document layouts, font styles, and text complexities.

Our key contributions are as follows:

- **First Large-Scale Synthetic Arabic OCR Dataset:** Unlike existing printed and handwritten datasets, SAND is fully synthetically generated, ensuring controlled variation in text representation, formatting, and document structure.
- **Extensive Typographic and Linguistic Diversity:** With 662.15 million words across five distinct Arabic fonts, SAND provides rich typographic variation to enhance OCR model robustness.
- **We benchmarked baseline OCR models** on structured Arabic book-style text using SAND. Evaluations on Tesseract and EasyOCR provide baseline Character Error Rate (CER) and Word Error Rate (WER), establishing a reference point for future OCR improvements.

To facilitate further research in Arabic OCR and vision-language modeling, we publicly release SAND[1], along with preprocessing scripts[2].

2 Related Work

Arabic OCR research has significantly evolved over the years, focusing on datasets and deep learning-based models to address the script's challenges, such as its cursive nature, contextual letter variations, and diverse typographic styles. Existing datasets primarily fall into two categories: printed and handwritten text.

Among the printed OCR datasets, APTI (Mosbah et al., 2024) stands as one of the largest, comprising 45 million synthetically generated Arabic word images. PAW (Bataineh, 2017) takes a different approach by focusing on sub-word recognition, whereas Yarmouk (Doush et al., 2018) consists of printed Arabic text extracted from Wikipedia articles. Classical literature is represented in PATS-A01 (Al-Muhtaseb et al., 2009), which includes 22,000 printed Arabic line images.

For handwritten OCR, MADBase (El-Sawy et al., 2016) and ABAD (El Abed et al., 2009) provide character-based and word-based recognition benchmarks, while OnlineKHATT (Mahmoud et al., 2018) and AlexU-Word (Hussein et al., 2014) offer extensive word-segmented datasets.

Beyond datasets, several deep-learning OCR models have been introduced. ADOCRNet (Mosbah et al., 2024) integrates CNNs and BiLSTM networks for improved Arabic OCR accuracy. Meanwhile, Arabic-Nougat (Rashad, 2024) leverages transformer-based architectures to convert Arabic book pages into structured Markdown text.

A comparative summary of major Arabic OCR datasets is provided in Table 1, highlighting dataset size, type, fonts, and focus.

Despite these advancements, no large-scale synthetic dataset exists for book-style Arabic text recognition. The reliance on printed and handwritten datasets limits OCR models' ability to generalize across structured document layouts. SAND fills this gap by offering a scalable, font-diverse, and book-style synthetic dataset, enabling better OCR model training and evaluation.

3 Methodology

The development of an effective Arabic Optical Character Recognition (OCR) dataset necessitates a structured and comprehensive approach to ensure robustness, diversity, and real-world applicability. This section details our methodology for creating a high-quality dataset that accurately reflects the complexities of Arabic text and common book layouts.

3.1 Dataset Composition

Our dataset is composed of text rendered in five distinct Arabic fonts, each selected to capture variations in typographic representation, stroke complexity, and character spacing. The fonts included are Amiri, Sakkal Calibri, Arial, and Scheherazade New. This selection ensures that OCR models trained on our dataset generalize well across diverse document styles and printed materials.

The dataset spans five distinct Arabic fonts, carefully selected to ensure typographic diversity. Each font contributes approximately 148,541 document images, enabling OCR models to generalize across

[1] https://huggingface.co/datasets/riotu-lab/text2image

[2] https://github.com/riotu-lab/text2image

Dataset Name	Year	Type	Size	Total Words	Fonts Used	Focus
PATS-A01 (Al-Muhtaseb et al., 2009)	2009	Printed	22K	8,248	8 Fonts	Classic Arabic Literature
APTI (Mosbah et al., 2024)	2013	Printed	45M	45M	10 Fonts	Single-word OCR
UPTI (Sabbour and Shafait, 2013)	2013	Printed	10K	12K	Urdu Font	Urdu OCR
Alexuw (Hussein et al., 2014)	2014	Handwritten	25K	10,989	-	Segmented Letter-based Recognition
PAW (Bataineh, 2017)	2017	Printed	400K	550K	5 Fonts	Sub-word OCR
MADBase (El-Sawy et al., 2016)	2017	Handwritten	70K	700K	-	Char-Based Recognition
Yarmouk (Doush et al., 2018)	2018	Printed	9K	436,921	-	Diverse Printed Text
OnlineKHATT (Mahmoud et al., 2018)	2018	Handwritten	10K	22,216	Multiple	Handwritten Arabic
Shotor (Asadi, 2020)	2020	Printed	120K	62,900	Various Farsi Fonts	Farsi OCR
ABAD (El Abed et al., 2009)	2021	Handwritten	15K	14.4M	-	Tunisian Place Names
IDPL-PFOD (Hosseini et al., 2021)	2021	Printed	30K	38,476	Farsi Typefaces	Synthetic Farsi OCR
Arabic-Nougat (Rashad, 2024)	2023	NN Extracted	13.7K	700K	-	Markdown Conversion
SAND (Ours)	**2025**	**Synthetic**	**743K**	**662.15M**	**5 Fonts**	**Large-scale Arabic book OCR**

Table 1: Comparison of SAND with existing Arabic OCR datasets

different text styles and layouts. Our proposed dataset spans multiple domains, as shown in Table 2.

Category	No. of Articles
Culture	13,253
Fatawa & Counsels	8,096
Literature & Language	11,581
Bibliography	26,393
Publications & Competitions	1,123
Shariah	46,665
Social	8,827
Translations	443
Muslim's News	16,725
Total Articles	**133,105**

Table 2: Distribution of articles across different domains

The diversity in topics ensures that the dataset captures a broad spectrum of writing styles, lexical variations, and contextual structures, further strengthening its effectiveness for OCR applications. To mimic real-world book layouts and maintain textual fidelity, the dataset is generated using a structured multi-step pipeline. 1 shows the key stages of the dataset creation process.

As shown in 1, each document adheres to industry-standard book formatting guidelines to ensure realistic text representation. Table 3 details the specifications adopted across different font styles.

The documents conform to standard A4 page dimensions, ensuring compatibility with real-world printed materials. The page layout follows the specifications outlined in Table 4.

By adhering to this systematic methodology, we have created a high-quality, large-scale Arabic OCR dataset that mirrors real-world book layouts. The use of multiple fonts, structured formatting, and a diverse range of topics ensures that this dataset is well-suited for training and evaluating deep learning-based OCR models. Our structured approach enables robust text recognition, improving OCR accuracy across different Arabic fonts and document styles.

Font	Words Per Page	Font Size
Sakkal Majalla	50–300	14 pt
Arial	50–500	12 pt
Calibri	50–500	12 pt
Amiri	50–300	12 pt
Scheherazade	50–250	12 pt

Table 3: Word count and font size specifications for different Arabic fonts.

Specification	Measurement
Left Margin	0.9 inches
Right Margin	0.9 inches
Top Margin	1.0 inch
Bottom Margin	1.0 inch
Gutter Margin	0.2 inches
Page Width	8.27 inches (A4)
Page Height	11.69 inches (A4)

Table 4: Page layout and margin specifications.

4 Evaluating OCR Models on SAND

To assess OCR model performance on structured Arabic book-style text, we evaluate Tesseract and EasyOCR on a subset of 500 images from the Amiri font split. Two key metrics are reported: Character Error Rate (CER) and Word Error Rate (WER). The results, summarized in Table 5, reveal substantial differences in OCR accuracy.

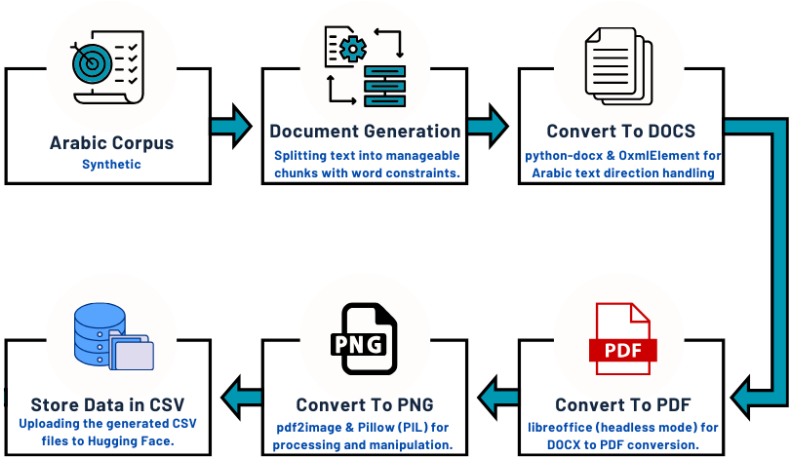

Figure 1: SAND Corpus Generation Pipeline

Figure 2: Example of Amiri Font in SAND

As shown in Table 5, Tesseract achieves the lowest CER (12.68%) and WER (32.23%), demonstrating its effectiveness in handling Arabic book-style text. EasyOCR, despite leveraging deep learning, performs significantly lower, with a CER of 52.02% and WER of 76.53%, suggesting poor generalization to structured Arabic text.

OCR Model	CER ↓	WER ↓
Tesseract	**0.1268**	**0.3223**
EasyOCR	0.5202	0.7653

Table 5: OCR performance on SAND Sample

These results serve as a benchmark for future Arabic OCR research. The high error rates in deep-learning-based models suggest that synthetic Arabic datasets like SAND could be used for fine-tuning OCR architectures, improving their generalization to diverse font styles, text structures, and book layouts.

5 Limitations

Despite its scale and diversity, SAND has certain limitations. As a synthetically generated dataset, it does not capture real-world imperfections such as scanned document noise, distortions, or handwritten text variability, which are commonly encountered in practical OCR applications. Additionally, SAND focuses on five selected Arabic fonts, and while these fonts cover a broad spectrum of typographic styles, they do not encompass the full range of Arabic typefaces. Future work will extend SAND by integrating scanned Arabic book pages with real-world imperfections, such as noise, distortions, and handwritten annotations, to further enhance OCR model robustness.

6 Conclusion

In this work, we introduced SAND, the first large-scale synthetic Arabic OCR dataset designed to facilitate the training and evaluation of vision-language models for Arabic text recognition. SAND consists of 743,000 document images containing 662.15 million words, covering five diverse Arabic fonts to ensure textual and typographic diversity. By simulating real-world book layouts, SAND provides a structured, scalable, and high-quality dataset that enables better OCR model generalization across varied document styles.

Acknowledgments

The authors thank Prince Sultan University for their support.

References

Husni Al-Muhtaseb, Sabri Mahmoud, and Rami Qahwaji. 2009. Automatic arabic text image optical character recognition method. Patent, US8150160B2.

I Saleh Al-Sheikh, MASNIZAH Mohd, and L Warlina. 2020. A review of arabic text recognition dataset. *Asia-Pacific J. Inf. Technol. Multimedia*, 9(1):69–81.

Amir Abbas Asadi. 2020. Shotor dataset. urlhttps://github.com/amirabbasasadi/Shotor. Accessed on 13/02/2025.

Bilal Bataineh. 2017. A printed paw image database of arabic language for document analysis and recognition. *Journal of ICT Research and Applications*, 11:199–211.

Iyad Abu Doush, Faisal AIKhateeb, and Anwaar Hamdi Gharibeh. 2018. Yarmouk arabic ocr dataset. In *2018 8th International Conference on Computer Science and Information Technology (CSIT)*, pages 150–154.

Haikal El Abed, Volker Märgner, and Adel Alimi. 2009. Icdar 2009 online arabic handwriting recognition competition. In *Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR 2009)*, pages 1388–1392.

Ahmed El-Sawy, Hazem M. El-Bakry, and Mohamed Loey. 2016. Cnn for handwritten arabic digits recognition based on lenet-5. *ArXiv*, abs/1706.06720.

Fatemeh Sadat Hosseini, Shima Kashef, Elham Shabaninia, and Hossein Nezamabadi-pour. 2021. IDPL-PFOD: An image dataset of printed Farsi text for OCR research. In *Proceedings of the Second International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2021) co-located with ICNLSP 2021*, pages 22–31, Trento, Italy. Association for Computational Linguistics.

Mohamed Hussein, Marwan Torki, Ahmed Elsallamy, and Mahmoud Fayyaz. 2014. Alexu-word: A new dataset for isolated-word closed-vocabulary offline arabic handwriting recognition.

Sabri Mahmoud, Hamzah Luqman, Baligh Al-helali, Galal Binmakhashen, and Mohammad Parvez. 2018. Online-khatt: An open-vocabulary database for arabic online-text processing. *The Open Cybernetics & Systemics Journal*, 12:42–59.

Lamia Mosbah, Ikram Moalla, Tarek M Hamdani, Bilel Neji, Taha Beyrouthy, and Adel M Alimi. 2024. Adocrnet: A deep learning ocr for arabic documents recognition. *IEEE Access*.

Rayyan Najam and Safiullah Faizullah. 2023. Analysis of recent deep learning techniques for arabic handwritten-text ocr and post-ocr correction. *Applied Sciences*, 13(13):7568.

Mohamed Rashad. 2024. Arabic-nougat: Fine-tuning vision transformers for arabic ocr and markdown extraction. *Preprint*, arXiv:2411.17835.

Nazly Sabbour and Faisal Shafait. 2013. A segmentation-free approach to arabic and urdu ocr. In *Document recognition and retrieval XX*, volume 8658, pages 215–226. SPIE.

UNESCO. 2024. World Arabic Language Day. UNESCO Official Website. The Arabic language is a pillar of the cultural diversity of humanity. It is one of the most widely spoken languages in the world, used daily by more than 400 million people.