Proposing a Comprehensive Dataset for Arabic Script OCR in the context of Digital Libraries and Religious Archives (Extended Abstract)

Riccardo Amerigo Vigliermo[1,2], Giovanni Sullutrone[1], Sonia Bergamaschi[1] and Luca Sala[1]

[1]*Università di Modena e Reggio Emilia (UNIMORE)*
[2]*Fondazione per le Scienze Religiose (FSCIRE)*

Abstract

Optical Character Recognition (OCR) technology is integral to digitizing and accessing historical documents within digital libraries. However, OCR systems often struggle to accurately recognize and classify complex document structures, especially historical texts with diverse layouts and languages. This preliminary study addresses this challenge by proposing the building of a comprehensive and community accessible dataset of Arabic title pages using advanced Vision Language Models (VLMs) and OCR tools. In this context, by extracting the first pages of each document at high resolution, we focused on accurately classifying frontispieces and distinguishing them from the main text to improve metadata quality and document retrieval in digital libraries. The Qwen-2vl-72B model was utilized to classify each page as either a 'frontispiece'or 'non-frontispiece'using a specially designed prompt. The identified frontispieces will be processed using Google Vision AI to automatically extract a Ground Truth to be evaluated by linguistic experts before completing the dataset. Further steps will envisage training of an open source solution such as Kraken OCR to also evaluate the effectiveness of the dataset. The innovative approach introduced here not only addresses the current lack of comprehensive datasets but also advances the effectiveness and precision of digital library initiatives such as the Digital Maktaba project.

Keywords

Arabic OCR, Datasets, Title pages, Digital Libraries, Religious Archives

1. Introduction

As digital mediums become central to information consumption, the demand for accessible and searchable digital texts has grown exponentially. Optical Character Recognition (OCR) technology has played a pivotal role in this transformation by converting various document formats into editable and searchable data. This technology is crucial for digital archiving, information retrieval, and data analysis. OCR technology still faces notable challenges, particularly with complex and historical documents. These challenges are amplified when processing Arabic script, where unique linguistic, typographic and calligraphic characteristics affect OCR accuracy. One of the primary barriers in OCR research and application is the lack of comprehensive, high-quality datasets designed for library usage. Moreover, in comparison to other languages and scripts, existing datasets often lack the breadth and specificity required to address these complex features, hindering the development and benchmarking of advanced OCR algorithms. This limitation is especially critical for elements like frontispieces (i.e., title pages) which contain unique artistic and typographical components that demand specialized OCR handling. Fig. 1 shows an example of title pages. In developing a cataloging tool, which is the main aim of the

IRCDL 2025: 21st Conference on Information and Research Sciences Connecting to Digital and Library Science, February 20-21, 2025, Udine, Italy

vigliermo@fscire.it (R. A. Vigliermo); giovanni.sullutrone@unimore.it (G. Sullutrone); sonia.bergamaschi@unimore.it (S. Bergamaschi); luca.sala@unimore.it (L. Sala)
0000-0003-4643-6128 (R. A. Vigliermo); 0009-0006-5556-1827 (G. Sullutrone); 0000-0001-8087-6587 (S. Bergamaschi); 0000-0002-4833-8882 (L. Sala)

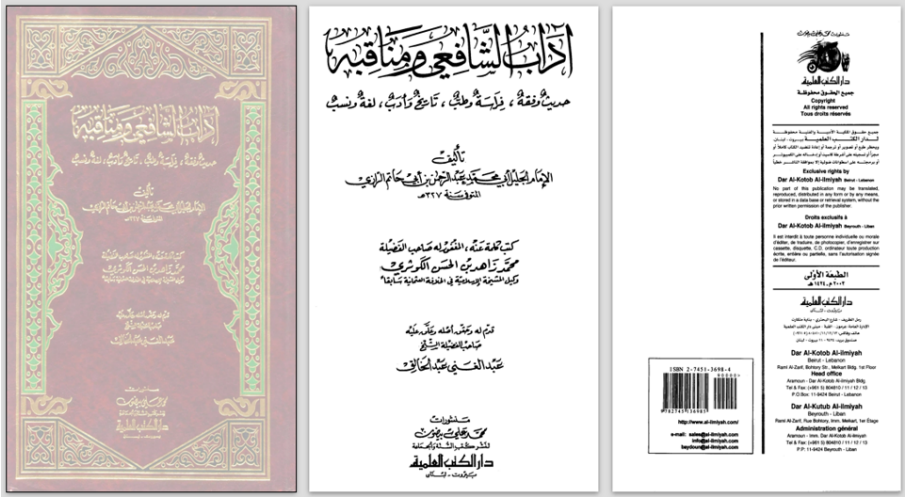

Figure 1: Example of a Frontispiece Pages Group (FPG) as intended in this study

Digital Maktaba project [1], we present here a possible pipeline for the creation of a title pages dataset to effectively train an Open Source OCR model such as Kraken [1], through the escriptorium VRE [2], to extract cataloging metadata from Arabic printed frontispieces. The presented work also considers recent advances in Vision Language Models [3, 4] that could contribute significantly to data extraction from images by integrating visual, as well as textual understanding. By leveraging VLMs and a closed source OCR such as Google Vision AI this study addresses these challenges by developing an extensive, community-accessible dataset of title pages in the Arabic script. We focus specifically on accurately classifying frontispieces and distinguishing them from the main text within the initial pages of each document. Additionally, we aim to generate enriched metadata for improved organization and retrieval in digital libraries. Currently, we are in the process of constructing a dataset that captures the diverse typographic and structural challenges of Arabic texts. Our immediate goal is to finalize this dataset and utilize the best available OCR tools to automatically extract text, or portions of it, from these documents. This extracted text will then be meticulously reviewed and corrected by experts to create a gold-standard reference, ensuring accuracy for future OCR advancements. In the sections that follow, we provide background on OCR technology, elaborate on the unique challenges posed by Arabic script, outline our objectives and methodology, and discuss the preliminary workflow and the anticipated outcomes and implications of our study.

2. Background and Related Works

2.1. Arabic Optical Character Recognition in Digital Libraries: the Title page as FPG and VrD

OCR technology has played a key role in library and archive digitization efforts across the globe. By converting printed text into machine-readable formats, OCR facilitates the storage, retrieval, and analysis of vast document collections. Traditional OCR systems rely on pattern recognition and machine learning techniques to interpret character shapes and word patterns in scanned images [5]. While effective for documents with conventional fonts and layouts, these systems struggle with documents that deviate from these standards. The Arabic script poses several specific challenges due to its unique linguistic and typographic features. Arabic is written in a cursive, right-to-left (RtL) script, complicating

[1]The Digital Maktaba (DM), defined as Work Package (WP) 5 in the ITSERR (Italian Strengthening of the ESFRI RI RESILIENCE) project, functions as the primary source and ultimate goal of the study presented here. This WP is dedicated to crafting a digital library that can analyze and extract information from multi-lingual documents, particularly from Arabic scripts (Arabic, Persian and Azerbaijani), offering a state-of-the-art cataloging methodology designed specifically for religious studies libraries that need to manage multi-lingual and multi-alphabetic cultural resources

character segmentation. Some Arabic graphemes have diacritical dots (above or below the baseline) that can shift in different calligraphic-typographic styles leading to recognition errors. The script primarily represents consonants, with vowels indicated by optional diacritical marks (that are often not presented). Arabic characters change shape depending on their position within a word (initial, medial, final, or isolated). Some graphemes do not connect (bind) with others creating words with multiple disconnected components. Lastly, in some calligraphic styles the characters may overlap, touch, or appear in slanted orientations. In the present preliminary investigation, the title pages could be conceived as part of a series of pages that we refer to as Frontispiece Pages Group (FPG). This definition is motivated by the context of frontispiece OCR analysis and character extraction for the development of a system able to support the librarian's work. FPGs are groups of pages where most of the metadata useful for cataloging is present. It should be also considered as such since we often have the recurrence of the same information in different scripts on different pages and with different layouts (e.g., title is represented in the title page, as well as in other pages, sometimes in other fonts). The title page in many cases is a black-on-white re-proposition of the cover page where the mere binarization of the text does not solve other graphical issues (decorations, vocalization, etc.). Subsequent pages of the FPG usually report useful data (even in a fragmented manner across several pages) in more 'normalized'scripts allowing for easier text extraction and cataloging. In some cases, the information is placed inside special boxes on one of the pages following the title page. Moreover, the title page in the FPG, from a Document Analysis perspective, could be considered in many cases the same as a Visually Rich Documents (VrD) [6, 7, 8, 9], especially considering that scanned PDFs from the physical realm could also bear noise elements such as library stamps or marks and several other issues related to the state of the paper as support.

2.2. Peculiar challenges posed by Arabic script FPGs

Frontispieces present unique challenges for OCR extraction that are not typically encountered with internal pages. Frontispieces often exhibit a high degree of variability and complexity. Several factors contribute to the increased difficulty:

Variety in Layouts and Designs. Frontispieces may include ornate designs, decorative elements, and unconventional layouts that intertwine text and images. This visual richness can confuse OCR systems, which are primarily trained on text-centric pages.

Diverse Backgrounds and Noise. The presence of backgrounds with various colors, textures, or deteriorated conditions adds noise to the images. Such backgrounds can interfere with text recognition by obscuring characters or creating false positives.

Non-Standard Fonts and Scripts. Frontispieces often feature artistic or custom typefaces, including calligraphic styles like *Naskh*, *Nasta'līq*, or *Kūfī*. These fonts have unique graphical peculiarities that are not always well-represented in standard OCR training datasets.

Multiscript Content. They may contain text in multiple scripts, such as Arabic and Latin sometimes within the same page. Moreover, the use of numerals (for dates) and alphabetic script is by itself a challenge since Arabic-indic numerals have a LtR orientation while Arabic script is RtL.

Presence of Vocalization and Diacritics. The inclusion or omission of vowels and diacritical marks can vary, affecting character recognition. Decorations or artistic elements might be mistaken for diacritics, leading to misinterpretation of the text.

These challenges are compounded by external variables such as overall image quality, character resolution, different levels of support degradation and the presence of colored fonts or backgrounds. Previous works [10, 11, 12, 13] highlighted that these issues requires not only advanced OCR algorithms but also carefully curated datasets.

2.3. Related Works: Vision-Language Models in OCR

VLMs models have applications across a range of fields, including image captioning, visual question answering, and object recognition, where both visual and textual data are combined [14]. Leveraging large-scale neural networks and extensive datasets, VLMs are designed to interpret complex images

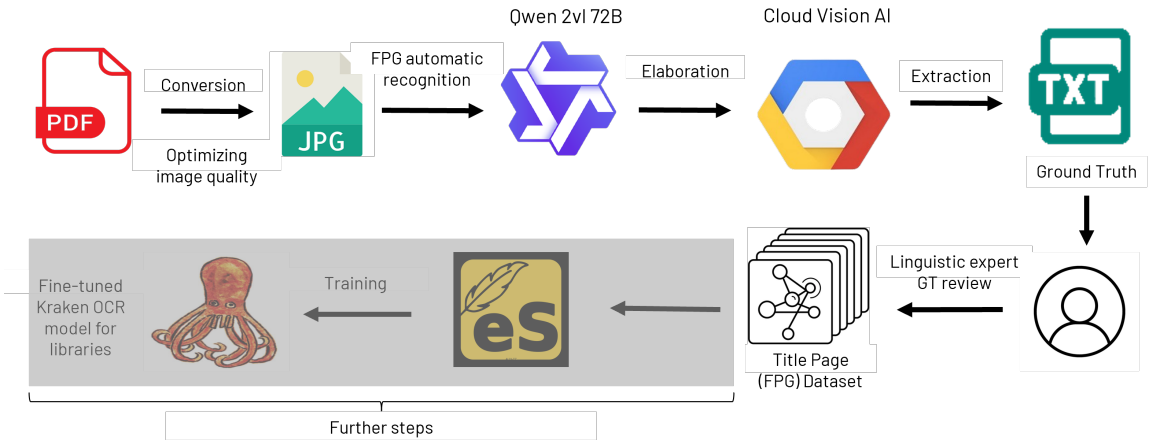

Figure 2: Example of the proposed pipeline. In grey some further step of training and fine tuning a Kraken OCR model for librarians use

that may include text, graphics, and other visual elements [4]. Although VLMs are still emerging in document analysis, their potential for handling mixed-media layouts and intricate document structures is promising. Unlike traditional OCR systems, which are focused solely on recognizing text, VLMs take a holistic approach by analyzing both visual and textual components of an image. For instance, the Qwen-2vl-72B [3] model is designed to analyze multimodal data and can perform tasks such as image captioning and visual context interpretation, which could theoretically aid in recognizing text within complex visual contexts. At the time of writing, it is the best overall performing open-source VLM according to benchmark evaluations [15].

2.4. Related Works: Arabic Printed Characters Datasets

In the last two decades Arabic script OCR studies have made significant steps forward. An example of dataset useful for both handwritten and printed Arabic text is ARABASE [16]. The Dataset is a collection of documents images and Part of Arabic Words (PAWs). In 2009, the APTI dataset (Arabic Printed Text Image) was composed of 45,313,600 word images covering 250 million characters representing one of the most extensive [17]. In 2010 PATDB (Printed Arabic Text Data Base) [18] was published as a corpus consisting of 6954 scanned pages images with different dpi resolutions. Few years later the multi-font dataset APTID/MF (Arabic Printed Text Image Dataset/Multi-Font) emerged as another solution for segmentation and automatic font identification researches with 387 pages of documents scanned in grayscale, from which 1845 text-blocks and a large dataset of 27,402 samples were extracted [19]. Similarly, the ALTID dataset was composed by 1,845 text blocks in Arabic and 2,328 in Latin alphabet from 731 greyscale images [20]. Worth to mention is also the KAFD dataset [21], which consists in 40 fonts in 10 sizes and 4 different styles (e.g., bold, italic, etc.), while in 2017 another PAW dataset was created from 83,056 text images representing all the words of the Arabic language in different Arabic fonts: (e.g., *Thuluth*, *Naskh*, etc.) for a combined total of 415,280 images [22]. In 2022 a bilingual dataset with the name of BPTI (Arabic/English) was ideated to address the lack of availability of bilingual text datasets. It consists of 97,812 text images categorized into two groups: Scanned page and digitized line images [23]. Finally, two datasets designed and developed for the recognition of Arabic printed text with examples and text images collected from the Qur'ān: QTID (Quran Text Image Dataset) consisting of 309,720 images with a total of 2,494,428 characters from the Qur'anic text [24]; and the second [25] containing 604 images at page level and 8,927 images at text line level from the Medina Qur'ān (*muṣḥaf al-madīna*). At the best of our knowledge no specific dataset has been developed for the analysis of Arabic printed title pages OCR handling in the context of librarian use and digital libraries.

3. Methodology

The primary focus of this study is to improve the OCR capabilities for digital libraries by developing a comprehensive, community-accessible frontispiece dataset. Our methodology involves assembling an initial set of historical documents from the FSCIRE "La Pira" digital archive, classifying their pages as 'frontispiece' or 'non-frontispiece', and creating an initial OCR draft of the frontispieces using Google Vision AI, which will be evaluated with common metrics such as Carachter Error Recognition (CER) and Word Error Recognitin (WER), analyzed and corrected by linguistic experts. The decision to implement this tool for dataset creation is based on previous tests of open-source solutions such as EasyOCR, Tesseract, and Google Docs [11] where the latter emerged for better performances on the languages in our collection. Despite that, the large number of images to be processed and the accuracy and flexibility needed for our specific goal led us to the selection of Google Vision AI. A visual representation of the pipeline is shown in Fig.2 while further details are provided in the following sections.

Document Selection and Preparation. We collected approximately 140,000 donated documents, selected for their diversity in content, formats and cultural significance. This collection ensures that the dataset spans a range of subjects, periods, languages, layouts, fonts, and visual elements, aligning with our goals of cataloging and preserving large non-latin cultural heritages. The documents were then divided into digitized and non-digitized with the latter selected for the next processing steps.

Resource management.To manage resources effectively and maintain consistency, we extracted the first six pages of each document at high resolution, as these initial pages typically contain frontispieces and other introductory materials. Each page is scaled proportionally to 4096x4096 pixels to reduce the amount of converted tokens for the VLM used in the next step.

Processing Documents. To process these pages, we selected the Qwen-2vl-72B model, one of the best open-source VLMs currently available. Using a specialized prompt, we guided the model in classifying pages as frontispieces or non-frontispieces. After obtaining the subset of frontispieces, we used Google Vision AI to produce an initial text extraction. Given that the analyzed pages are particularly challenging (as stated in Section 2.2), the results of this step will be given to human experts for correction, resulting in a high-quality OCR data set at the end of the pipeline.

4. Conclusion and Future Directions

In this extended abstract, we have presented our planned approach to improve OCR capabilities for Arabic frontispieces in digital libraries. Recognizing the unique challenges of Arabic script and the scarcity of specialized datasets, we aim to develop a comprehensive, high-quality dataset by processing approximately 140,000 historical documents. By employing advanced Vision-Language Models like Qwen-2vl-72B for page classification and OCR tools such as Google Vision AI for initial text extraction, we intend to create a reliable resource for training and benchmarking OCR algorithms. Our future work will focus on finalizing this dataset, refining the OCR pipeline, and collaborating with linguistic experts to ensure accuracy. We believe this effort will significantly contribute to the preservation and accessibility of Arabic texts in digital libraries, supporting advanced cataloging and research initiatives. Moving forward, we plan to publicly release our curated frontispieces dataset, providing a valuable resource for the research community. With this dataset, we plan to train the open source Kraken OCR engine to develop a model specifically tailored for the librarian and cataloger usage, also aiming to improve OCR accuracy for these complex pages and to facilitate better metadata extraction and cataloging. Additionally, by offering the dataset as a benchmark, we hope to support the evaluation and advancement of OCR systems focused on frontispiece recognition.

Acknowledgments

This work was supported by the PNRR project Italian Strengthening of Esfri RI Resilience (ITSERR) funded by the European Union – NextGenerationEU (CUP:B53C22001770006).

References

[1] B. Kiessling, G. Kurin, M. T. Miller, K. Smail, Advances and Limitations in Open Source Arabic-Script OCR: A Case Study, Digital Studies / Le champ numérique 11 (2021). URL: http://arxiv.org/abs/2402.10943. doi:10.16995/dscn.8094, arXiv:2402.10943 [cs].

[2] P. Stokes, B. Kiessling, D. Stökl Ben Ezra, R. Tissot, E. Gargem, The EScriptorium VRE for Manuscript Cultures, Classics@ Journal, Ancient Manuscripts and Virtual Research Environments 18 (2021).

[3] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, J. Zhou, Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. URL: https://api.semanticscholar.org/CorpusID:261101015.

[4] H. Liu, C. Li, Y. Li, Y. J. Lee, Improved baselines with visual instruction tuning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26296–26306.

[5] S. Naz, A. I. Umar, S. H. Shirazi, S. B. Ahmed, M. I. Razzak, I. Siddiqi, Segmentation techniques for recognition of Arabic-like scripts: A comprehensive survey, Education and Information Technologies 21 (2015) 1–20. doi:10.1007/s10639-015-9377-5.

[6] X. Liu, F. Gao, Q. Zhang, H. Zhao, Graph Convolution for Multimodal Information Extraction from Visually Rich Documents, in: A. Loukina, M. Morales, R. Kumar (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 32–39. URL: https://aclanthology.org/N19-2005. doi:10.18653/v1/N19-2005.

[7] H. Wang, Q. Wang, Y. Li, C. Wang, C. Chu, R. Wang, DocTrack: A Visually-Rich Document Dataset Really Aligned with Human Eye Movement for Machine Reading, 2023. URL: http://arxiv.org/abs/2310.14802. doi:10.48550/arXiv.2310.14802, arXiv:2310.14802 [cs].

[8] K.-A. L. Nguyen, Document Understanding with Deep Learning Techniques. Document and Text Processing., Ph.D. thesis, Sorbonne Université, 2024., 2024. URL: https://theses.hal.science/tel-04626992/file/140733_NGUYEN_2024_archivage.pdf.

[9] L. Nguyen, B. Piwowarski, J. Laborde, G. Moyse, Learning Reading Order via Document Layout with Layout2Pos, in: A. Antonacopoulos, A. Hinze, B. Piwowarski, M. Coustaty, G. M. Di Nunzio, F. Gelati, N. Vanderschantz (Eds.), Linking Theory and Practice of Digital Libraries, Springer Nature Switzerland, Cham, 2024, pp. 3–19. doi:10.1007/978-3-031-72437-4_1.

[10] S. Bergamaschi, R. Martoglia, F. Ruozzi, R. A. Vigliermo, S. De Nardis, L. Sala, M. Vanzini, Preserving and conserving culture: First steps towards a knowledge extractor and cataloguer for multilingual and multi-alphabetic heritages, in: Proceedings of the Conference on Information Technology for Social Good, GoodIT '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 301–304. URL: https://doi.org/10.1145/3462203.3475927. doi:10.1145/3462203.3475927.

[11] S. Bergamaschi, S. De Nardis, R. Martoglia, F. Ruozzi, L. Sala, M. Vanzini, R. A. Vigliermo, Novel perspectives for the management of multilingual and multialphabetic heritages through automatic knowledge extraction: The digitalmaktaba approach, Sensors 22 (2022). URL: https://www.mdpi.com/1424-8220/22/11/3995. doi:10.3390/s22113995.

[12] R. Martoglia, L. Sala, M. Vanzini, R. A. Vigliermo, A tool for semiautomatic cataloguing of an islamic digital library: A use case from the digital maktaba project, in: A. Paschke, G. Rehm, C. Neudecker, L. Pintscher (Eds.), Proceedings of the Third Conference on Digital Curation Technologies (Qurator 2022), Berlin, Germany, Sept. 19th-23rd, 2022, volume 3234 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022. URL: https://ceur-ws.org/Vol-3234/paper1.pdf.

[13] R. Martoglia, S. Bergamaschi, F. Ruozzi, M. Vanzini, L. Sala, R. A. Vigliermo, Knowledge extraction, management and long-term preservation of non-Latin cultural heritages - Digital Maktaba project presentation, in: B. Alessia, F. Alex, F. Stefano, M. Stefano, R. Domenico (Eds.), Proceedings of the 19th Conference on Information and Research Science Connecting to Digital and Library Science, volume 3365 of *CEUR Workshop Proceedings*, CEUR, Bari, Italy, 2023, pp. 153–161. URL:

https://ceur-ws.org/Vol-3365/#short11, iSSN: 1613-0073.

[14] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, Z. Muyan, Q. Zhang, X. Zhu, L. Lu, B. Li, P. Luo, T. Lu, Y. Qiao, J. Dai, Intern vl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks, 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 24185–24198. URL: https://api.semanticscholar.org/CorpusID:266521410.

[15] Q. Team, Qwen-2vl: Open multimodal large model, https://qwen2.org/vl/, 2024. Accessed: December 01, 2025.

[16] N. E. B. Amara, O. Mazhoud, N. Bouzrara, N. Ellouze, ARABASE: A Relational Database for Arabic OCR Systems., Int. Arab J. Inf. Technol. 2 (2005) 259–266.

[17] F. Slimane, R. Ingold, S. Kanoun, A. Alimi, J. Hennebert, A New Arabic Printed Text Image Database and Evaluation Protocols, in: 10th International Conference on Document Analysis and Recognition, Barcelona, Spain, 2009, pp. 946–950. doi:10.1109/ICDAR.2009.155.

[18] A. G. Al-Hashim, S. A. Mahmoud, Printed Arabic text database (PATDB) for research and benchmarking, in: Proceedings of the 9th WSEAS international conference on Applications of computer engineering, ACE'10, World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, USA, 2010, pp. 62–68.

[19] F. K. Jaiem, S. Kanoun, M. Khemakhem, H. El Abed, J. Kardoun, Database for Arabic Printed Text Recognition Research, in: A. Petrosino (Ed.), Image Analysis and Processing – ICIAP 2013, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2013, pp. 251–259. doi:10.1007/978-3-642-41181-6_26.

[20] I. Chtourou, A. C. Rouhou, F. K. Jaiem, S. Kanoun, ALTID : Arabic/Latin Text Images Database for recognition research, in: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), IEEE Computer Society, 2015, pp. 836–840. URL: https://www.computer.org/csdl/proceedings-article/icdar/2015/07333879/12OmNwB2dXI. doi:10.1109/ICDAR.2015.7333879.

[21] H. Luqman, S. A. Mahmoud, S. Awaida, KAFD Arabic font database, Pattern Recognition 47 (2014) 2231–2240. URL: https://linkinghub.elsevier.com/retrieve/pii/S0031320313005463. doi:10.1016/j.patcog.2013.12.012.

[22] B. Bataineh, A Printed PAW Image Database of Arabic Language for Document Analysis and Recognition, Journal of ICT Research and Applications 11 (2017) 199–211. doi:10.5614/itbj.ict.res.appl.2017.11.2.6.

[23] M. H. Yahia, H. Al-Muhtaseb, Bpti: Bilingual (Arabic/English) Printed Text Images Dataset for Recognition Research, 2022. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4007916.

[24] M. Badry, H. Hassan, H. Bayomi, H. Oakasha, QTID: Quran Text Image Dataset, International Journal of Advanced Computer Science and Applications (IJACSA) 9 (2018) 385–391. URL: https://thesai.org/Publications/ViewPaper?Volume=9&Issue=3&Code=IJACSA&SerialNo=51. doi:10.14569/IJACSA.2018.090351, number: 3 Publisher: The Science and Information (SAI) Organization Limited.

[25] I. Alsheikh, M. Mohd, A Quranic Dataset for Text Recognition, in: Proceedings of the 1st International Conference on Informatics, Engineering, Science and Technology, Bandung, Indonesia, 2019. doi:10.4108/eai.18-7-2019.2287842.