

VARNAONYX - OCR USING TESSERACT

Bukka Charitha Reddy¹, Bodramoni Eshika², J. Bhanu Teja Manikanta³, K. Shravan Kumar⁴

¹Student, Dept. of Electronics and Communication Engineering, MVSR Engineering College, Hyderabad, India

²Student, Dept. of Electronics and Communication Engineering, MVSR Engineering College, Hyderabad, India

³Student, Dept. of Electronics and Communication Engineering, MVSR Engineering College, Hyderabad, India

⁴Assistant Professor, Dept. of Electronics and Communication Engineering, MVSR Engineering College, Hyderabad, India

Abstract - This work introduces VarnaOnyx, an OCR based system that integrates Digital Signal Processing techniques to enhance text extraction from both scanned and handwritten documents. Conventional OCR approaches tend to produce lower accuracy when dealing with noisy or degraded input images. To address this, the proposed system applies preprocessing techniques such as grayscale conversion, noise filtering, and contrast enhancement before recognition using the Tesseract OCR engine. The proposed method enhances text visibility and improves recognition accuracy while keeping the processing pipeline efficient. The system also provides a user-friendly interface for uploading images and obtaining editable text. Experimental results demonstrate improved performance compared to direct OCR on raw images.

Key Words: Text Extraction, Optical Character Recognition (OCR), Document Parsing, Scanned Documents, Digital Documents, Large Language Models (LLMs)

1. INTRODUCTION

In recent years, the need for transforming physical documents into digital formats has increased significantly due to the growth of digital systems and data-driven applications. Optical Character Recognition (OCR) plays a key role in this transformation by enabling the extraction of textual information from images, scanned documents, and handwritten notes. This decreases manual data entry and improves efficiency in document management.

Despite advancements in OCR technologies, accurately recognizing text from real-world documents remains a challenge. Factors such as noise, low resolution, uneven lighting, skewed text, and complex backgrounds often affect recognition performance. Traditional OCR systems tend to perform well on clean, printed text but struggle when applied to degraded or handwritten inputs.

As a solution to these challenges, this paper explains VarnaOnyx, an enhanced OCR system that integrates Digital Signal Processing (DSP) techniques with the Tesseract OCR engine. The system focuses on improving input image quality through preprocessing methods such as filtering, noise reduction, and contrast enhancement before performing text recognition. By combining preprocessing techniques with OCR, the proposed approach aims to improve accuracy while maintaining a simple and efficient workflow. The system is designed to handle both scanned and digitally generated documents, making it suitable for various real-world applications including academic, administrative, and archival use.

2. LITERATURE REVIEW

Optical Character Recognition (OCR) refers to the process of converting textual content from images or scanned documents into editable digital text. OCR systems have evolved with the integration of machine learning techniques, improving their ability to recognize characters. However, their effectiveness is still influenced by the quality of the uploaded image. Issues such as noise, poor lighting, distortions, and background interference often reduce recognition accuracy, making preprocessing an important stage in OCR systems.

Previous studies have emphasized the role of preprocessing in enhancing OCR performance. One such work explored the use of techniques like smoothing filters and adaptive binarization to improve text visibility before

recognition. The study demonstrated that refining image quality prior to OCR significantly increases accuracy, particularly when dealing with degraded or low-resolution inputs.

Another research effort introduced a combined approach where image enhancement methods were integrated with conventional OCR engines. Techniques such as edge sharpening, contrast adjustment, and morphological processing were applied to better isolate textual regions. Although this method improved recognition results for complex images, it also added additional computational overhead.

A further study examined OCR systems that incorporate both preprocessing and post processing stages, including error correction using linguistic models. While this approach resulted in higher recognition rates, it required more processing power and was less efficient for lightweight applications. These observations signal that maintaining a balance between accuracy and system efficiency remains an important challenge in OCR system design.

3.METHODOLOGY

The methodology of The VarnaOnyx system is a step-by step processing pipeline designed to convert input documents into accurate, machine readable text. The system emphasizes improving input quality before recognition and ensuring efficient extraction of textual information.

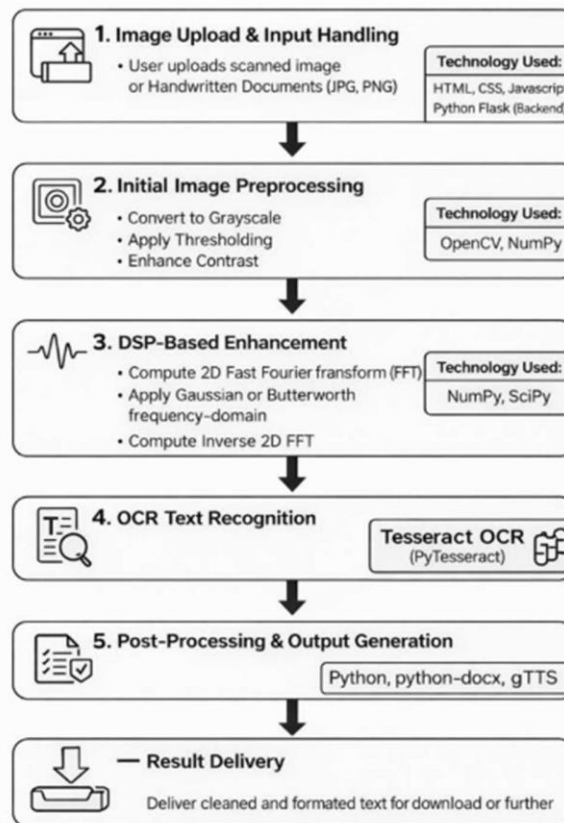


Fig-1: Algorithm of VarnaOnyx OCR with DSP based preprocessing

The VarnaOnyx system presents an effective approach to document digitization by integrating Digital Signal Processing (DSP) techniques with an OCR engine. The proposed framework enhances input image quality through preprocessing steps such as noise reduction, filtering, and contrast improvement, which significantly improves text recognition accuracy. By combining these techniques with the Tesseract OCR engine, the system

is able to handle a variety of document types, including scanned, handwritten, and digitally generated inputs. The implementation of a web-based interface further improves usability, allowing users to easily upload documents and obtain editable text. Experimental observations demonstrate that the preprocessing stage plays a crucial role in reducing recognition errors and improving output quality compared to direct OCR processing. Despite its effectiveness, the system has certain limitations when dealing with highly complex handwriting and heavily degraded documents. Future work can focus on integrating deep learning-based OCR models, incorporating language-based error correction, and extending support for multiple languages. Additionally, optimizing the system for real-time processing and mobile platforms can further enhance its practical applicability.

In conclusion, VarnaOnyx provides a balanced and efficient solution for OCR based document processing, offering improved accuracy, flexibility, and usability for real-world applications. The process starts with data acquisition, where the user uploads documents in the form of images or PDFs through a web-based interface. These inputs may include scanned documents, handwritten notes, or digitally generated files.

Once the input is received, it undergoes a preprocessing stage aimed at enhancing image quality. This step includes grayscale conversion, noise reduction, filtering, and contrast enhancement. Digital Signal Processing (DSP) techniques such as Gaussian and Butterworth filtering are applied to reduce noise and improve clarity. Additionally, operations like binarization and skew correction help in aligning the text and improving readability.

After preprocessing, the refined input is passed to the text extraction stage. For image-based inputs, the Tesseract OCR engine is used to recognize and convert text into editable format. For digitally generated documents, parsing techniques are used to directly extract textual content. This approach ensures that both scanned and digital documents are handled efficiently.

The extracted text is then subjected to post-processing, where minor refinements are applied to improve readability and formatting. This includes the removal of unwanted characters and basic structuring of the output text. Finally, the system produces an output stage, where the processed text is displayed to the user through the interface. The user can review, edit, and export the text for further use. This structured workflow ensures improved accuracy while minimizing manual effort.

4. SYSTEM ARCHITECTURE

The VarnaOnyx system is designed using a modular and layered architecture to ensure efficient processing, scalability, and improved accuracy. The overall framework is divided into four main components: the Input Module, Image Preprocessing Module, OCR Engine Module, and Output Module. These components work together in a sequential pipeline to transform various document formats into machine-readable text, as illustrated in Fig. 1.

The process starts with the Input Module, where users upload documents in formats such as JPG, PNG, or PDF through a web-based interface. The system can handle both scanned documents, including handwritten and printed text, as well as digitally generated files. This flexibility allows the system to be applied across different types of document sources.

The uploaded input is then passed to the Image Preprocessing Module, where the quality of the file is enhanced before text extraction. For scanned inputs, techniques such as binarization, noise removal, and skew correction are applied to improve readability.

Additionally, Digital Signal Processing (DSP) methods, including Gaussian and Butterworth filtering, are used to reduce noise and enhance clarity. For digitally generated documents, parsing and cleaning operations are performed to prepare the content for extraction.

Following preprocessing, the refined data is processed in the OCR Engine Module. The system utilizes the Tesseract OCR engine to convert image-based text into editable format. For digital documents, parsing tools are used to directly extract textual content and metadata. The outputs from both OCR and parsing processes are combined into a unified raw text format for further refinement.

Finally, the processed text is handled by the Output Module, where it is structured and displayed on to the UI. Basic post processing is applied so as to improve readability and remove unwanted elements. The final output is presented through the interface, encouraging users to edit, save, or export the extracted text. This structured workflow makes sure of efficient document digitization while minimizing manual intervention.

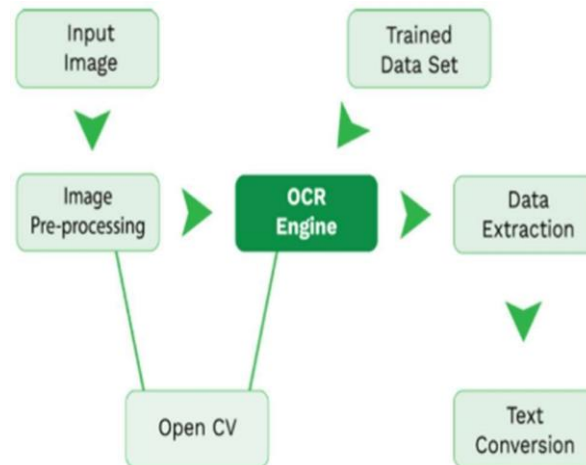


Fig -2: Overview of the Document Text Extraction Pipeline

5. EXPERIMENTAL SETUP AND EVALUATION

To understand the performance of The VarnaOnyx system, a series of experiments were conducted using various types of document inputs, including scanned images, handwritten notes, and digitally generated PDFs. The goal of the evaluation was to analyze the effectiveness of the preprocessing techniques and the overall accuracy of text extraction.

The technologies used in the system were Python, with libraries like OpenCV for image processing and Tesseract OCR for text recognition. A web-based interface was developed using Flask along with HTML, CSS, and JavaScript to allow users to upload documents and view extracted text. The experiments were performed on standard computing hardware to ensure practical applicability.

The evaluation process involved comparing OCR results of raw input images with those processed through the VarnaOnyx preprocessing pipeline. Key preprocessing techniques applied include grayscale conversion, noise removal, Gaussian filtering, Butterworth filtering, and thresholding. The steps were analyzed for their impact on improving text clarity and recognition accuracy.

Performance was assessed based on qualitative observations such as readability of extracted text, reduction in recognition errors, and overall output consistency. The results showed that preprocessing significantly improved OCR performance, especially for noisy and low-quality inputs. The system was able to produce more accurate and structured text compared to direct OCR processing.

Additionally, the system demonstrated flexibility in handling both scanned and digital documents by combining OCR and parsing techniques. This hybrid approach reduced processing errors and improved efficiency. Overall, the obtained experimental outcomes confirm that the integrated DSP-based preprocessing with OCR leads to better accuracy and usability in real-world document digitization tasks.

6. IMPLEMENTATION AND RESULTS

The VarnaOnyx system is a web-based application to provide an interactive and user-friendly platform for document processing. The frontend was developed using technologies like HTML, CSS, and JavaScript, while the backend was connected using the Flask framework in Python. This setup enables users to upload documents and view extracted text efficiently through a browser interface.

The system accepts input files in formats such as JPG and PDF. Once uploaded, the document is processed through the preprocessing pipeline, followed by OCR-based text extraction. The extracted text is then displayed on the interface, allowing users to review and make edits if necessary.

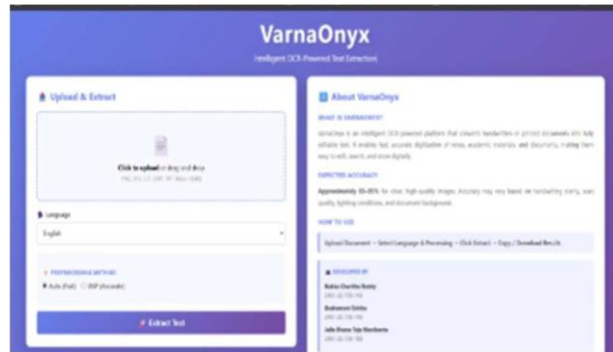


Fig -3: It shows the homepage of the VarnaOnyx system, where users upload documents. The interface is designed to be simple and accessible, ensuring ease of use.



Fig -4: It presents the output screen displaying the extracted text.

The results demonstrate that preprocessing techniques significantly improve text clarity and reduce recognition errors. Compared to direct OCR on raw input, the processed output is more accurate and structured.

The system was tested on different types of inputs, including handwritten notes, printed text, and low-quality scanned images. The results indicate that DSP-based preprocessing enhances OCR performance, particularly for noisy and distorted documents.

Overall, the implementation validates that VarnaOnyx provides an effective and practical solution for document digitization with improved accuracy and usability.

7. CONCLUSION AND FUTURE

The VarnaOnyx system presents an effective approach to document digitization by integrating Digital Signal Processing (DSP) techniques with an OCR engine. The proposed framework enhances input image quality through preprocessing steps such as noise reduction, filtering, and contrast improvement, which significantly improves text recognition accuracy. By combining these techniques with the Tesseract OCR engine, the system is able to handle a variety of document types, including scanned, handwritten, and digitally generated inputs. The implementation of a web-based interface further improves usability, allowing users to easily upload documents and obtain editable text. Experimental observations demonstrate that the preprocessing stage plays an important role in reducing recognition errors and improving output quality compared to direct OCR processing.

Despite its effectiveness, the system has few disadvantages when dealing with highly complex handwriting and heavily degraded documents. Future work can focus on integrating deep learning-based OCR models, incorporating language-based error correction, and extending support for multiple languages. Additionally, optimizing the system for real-time processing and mobile platforms can further improve its practical applicability.

In conclusion, VarnaOnyx provides a balanced and efficient solution for OCR based document processing, offering improved accuracy, flexibility, and usability for real-world applications.

ACKNOWLEDGEMENT

The authors would like to express their sincere thanks to Mr. K. Shravan Kumar, Assistant Professor at MVSR Engineering College, for his guidance, support, and vital suggestions throughout this work. His insights played a significant role in shaping the direction and quality of the study.

The author also extends thanks to the faculty and staff of the Department of Electronics and Communication Engineering for providing the necessary resources and assistance.

Additionally, appreciation is conveyed to the researchers and developers whose work in OCR, NLP, and document processing has contributed to the foundation of this project.

REFERENCES

- [1] Anakpluek, N., Pasanta, W., Chantharasukha, L., Chokratansombat, P., Kanjanakaew, P., & Siriborvornratanakul, T., "Improved Tesseract Optical Character Recognition Performance on Thai Document Datasets," *Big Data Research*, Vol. 39, 2025.
- [2] Guan, S., Lin, M., Xu, C., Liu, X., Zhao, J., Fan, J., Xu, Q., & Greene, D., "PreP-OCR: A Complete Pipeline for Document Image Restoration and Enhanced OCR Accuracy," *arXiv preprint*, 2025.
- [3] Sinha, R., & Rekha, B. S., "Digitization of Document and Information Extraction using OCR," *International Conference / Journal on Information Science and Engineering*, RV College of Engineering, Bangalore, India.
- [4] Smith, R., "An Overview of the Tesseract OCR Engine," *Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR)*, IEEE, 2007.
- [5] Gonzalez, R. C., & Woods, R. E., "Digital Image Processing," Pearson Education, 4th Edition, 2018.
- [6] Jain, A. K., Duin, R. P. W., & Mao, J., "Statistical Pattern Recognition: A Review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 1, 2000.
- [7] Otsu, N., "A Threshold Selection Method from Gray Level Histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 9, No. 1, 1979.