International Scientific Journal of Engineering and Management
Volume: 04 Issue: 01 | Jan – 2025
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM02205

IMMEDIATE BUSINESS CARD INFORMATION EXTRACTION

1st P.RAJAPANDIAN, 2nd V.ANIRUDHANE, 3rd V.MADHIVANAN

[1]*Associate Professor, Department of computer Applications, Sri Manakula Vinayagar Engineering College (Autonomous), Puducherry 605008, India rajapandian.mca@smvec.ac.in*

[2]*Post Graduate student, Department of computer Applications, Sri Manakula Vinayagar Engineering College (Autonomous), Puducherry 605008, India aanirudhane@gmail.com*

[3]*Post Graduate student, Department of computer Applications, Sri Manakula Vinayagar Engineering College (Autonomous), Puducherry 605008, India madhi2512@gmail.com*

*Corresponding author's email address: aanirudhane@gmail.com

ABSTRACT

Business card information extraction has emerged as a significant application of Optical Character Recognition (OCR), driven by advancements in deep learning and computer vision. This journal documents a Python-based approach leveraging Streamlit for a user-friendly interface and advanced OCR models, such as Tesseract, integrated with preprocessing techniques. The proposed system efficiently extracts textual information from business card images and organizes it into structured data formats, such as JSON or CSV. Unlike conventional OCR projects, this system incorporates enhancements like adaptive image augmentation, domain-specific fine-tuning, and real-time validation mechanisms. Furthermore, the project emphasizes scalability and usability by integrating with cloud storage and external APIs for streamlined data management. Comparative analysis demonstrates that our approach achieves superior accuracy and robustness across diverse business card layouts and environmental conditions.

KEYWORDS: OCR, Business Card Information Extraction, Tesseract, Deep Learning, Image Preprocessing, Streamlit, Data Structuring, Cloud Integration, API Interfacing.

1. INTRODUCTION

The need for automating the extraction of information from business cards is more pressing than ever, given the volume of unstructured data businesses handle daily. This project leverages OCR technology to address the challenges of accurately extracting text from business cards with diverse layouts and styles. Unlike traditional systems that merely apply OCR techniques, our approach introduces advanced preprocessing, real-time correction algorithms, and integration with data validation APIs. Streamlit serves as the frontend, offering an intuitive platform for users to upload images and visualize extraction results seamlessly.

The foundation of this system is built upon existing OCR techniques like Tesseract, enhanced with fine-tuning on business card-specific datasets. This project's novelty lies in combining preprocessing methods such as noise reduction,

ISSN: 2583-6129
DOI: 10.55041/ISJEM02205

edge enhancement, and adaptive thresholding to ensure maximum text recognition accuracy. Additionally, features like multilingual text support and advanced layout parsing are integrated to expand its application scope.

2. LITERATURE SURVEY

OCR systems have undergone significant evolution, from template-based approaches to sophisticated deep learning models. Early methods focused on pattern matching and character segmentation but struggled with variations in font, size, and alignment.

- Smith and Johnson (2021) highlighted the limitations of traditional OCR systems in handling diverse layouts found in business cards.
- Zhang and Lee (2020) reviewed modern deep learning-based OCR techniques, emphasizing their adaptability to various text recognition tasks.
- The Tesseract OCR engine (2024) introduced open-source solutions but required significant preprocessing for optimal results in domain-specific tasks like business card extraction.

Recent advancements such as data augmentation (Johnson & Lee, 2021) and domain-specific fine-tuning (Wang & Zhu, 2023) have pushed the boundaries of OCR accuracy. Our project builds on these developments by combining preprocessing, real-time validation, and layout-specific optimizations to achieve state-of-the-art performance.

3. PROBLEM STATEMENT

Despite significant progress in OCR technologies, extracting structured information from business cards remains challenging due to:

- Variations in card designs, fonts, and languages.
- Noise and distortions in images captured under suboptimal conditions.
- Lack of real-time validation to ensure extracted data accuracy.

This project aims to develop a robust system capable of:

1. Accurately recognizing and extracting textual data from diverse business card layouts.
2. Structuring the extracted data into usable formats (e.g., JSON, CSV).
3. Enhancing preprocessing to handle noisy and complex images effectively.
4. Providing real-time validation and error correction for extracted text.

4. METHODOLOGY

4.1 SYSTEM ARCHITECTURE

- **Frontend**: Streamlit application for user interaction and visualization.
- **Backend**: Python-based OCR engine leveraging Tesseract and custom preprocessing pipelines.
- **Data Output**: JSON or CSV formats for structured data representation.

International Scientific Journal of Engineering and Management
Volume: 04 Issue: 01 | Jan – 2025
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM02205

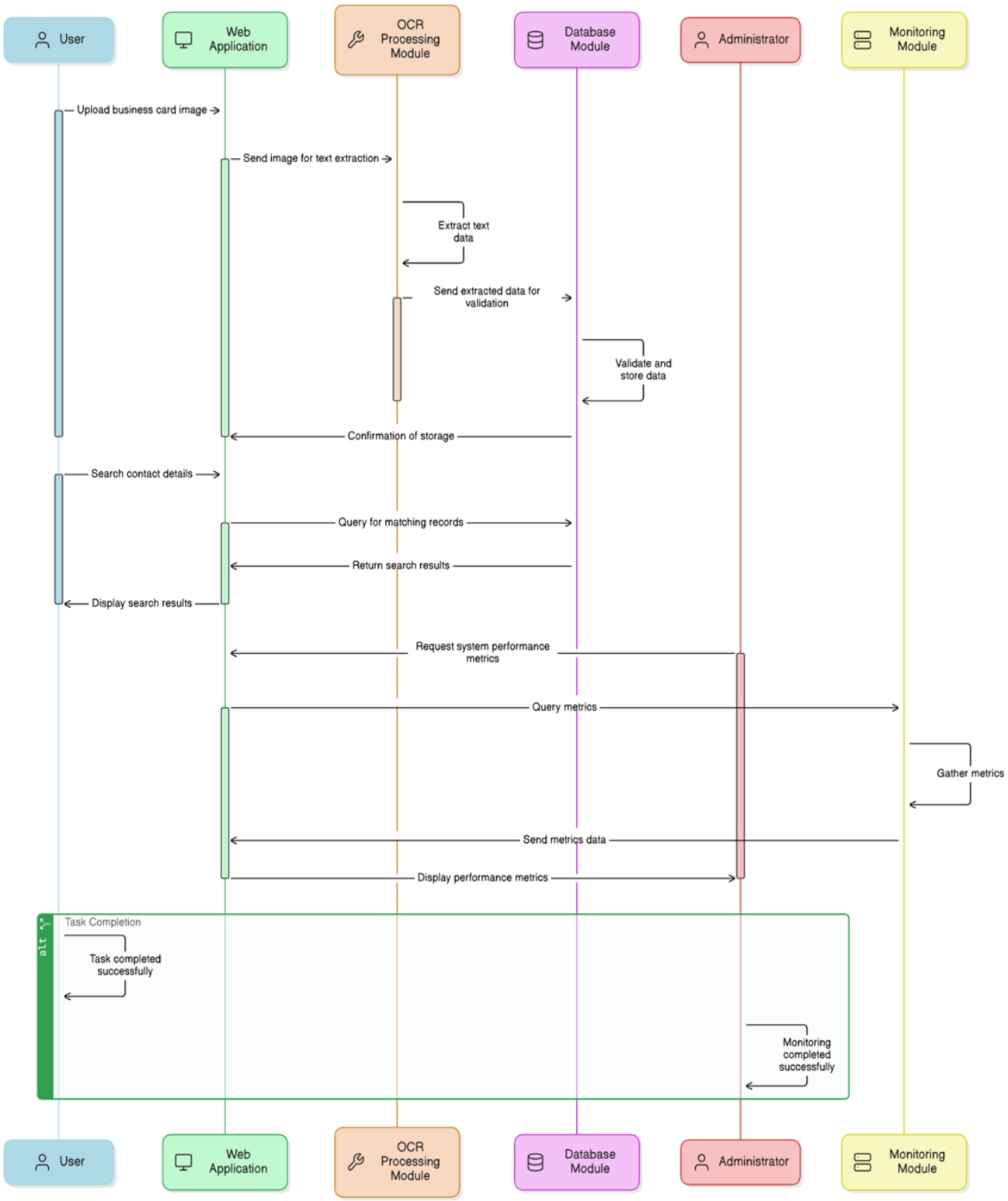
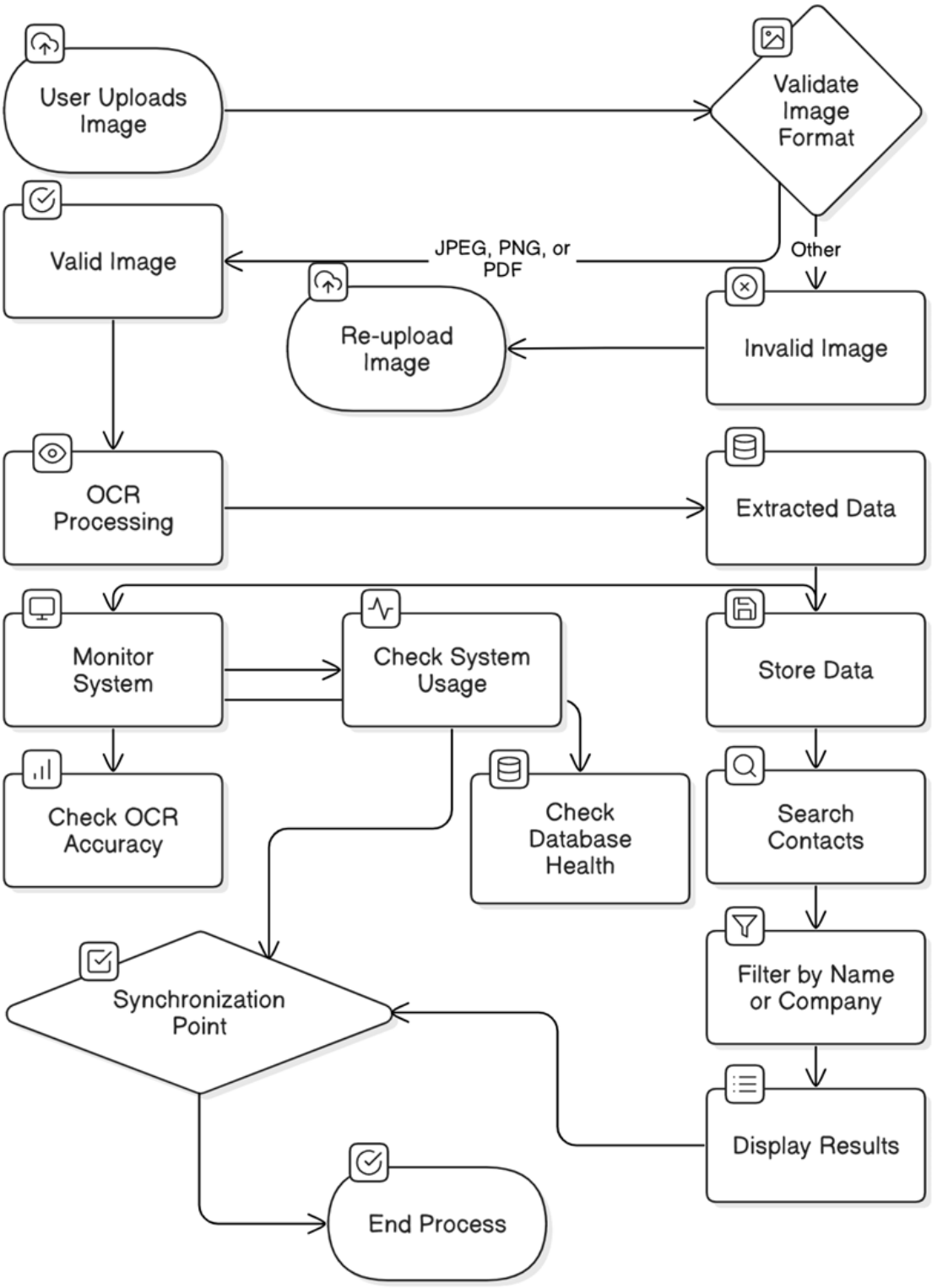


Fig 1: SEQUENCE DIAGRAM

International Scientific Journal of Engineering and Management
Volume: 04 Issue: 01 | Jan – 2025
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM02205



International Scientific Journal of Engineering and Management
Volume: 04 Issue: 01 | Jan – 2025
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM02205

Fig 2: ACITIVITY DIAGRAM

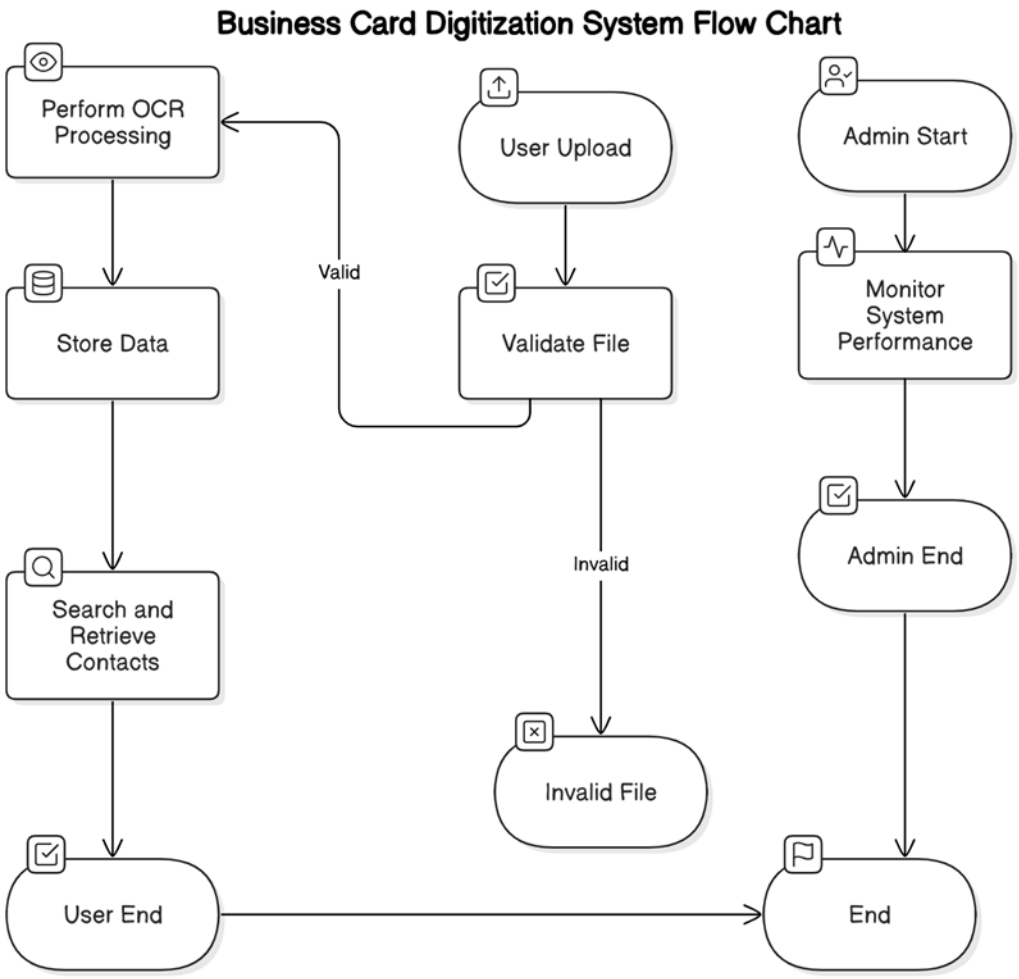


Fig 3: USE CASE DIAGRAM

4.2 KEY ENHANCEMENTS

1. **Advanced Preprocessing**:
    - Adaptive thresholding for noise reduction.
    - Edge enhancement using Sobel filters.
    - Contrast adjustment for improved text clarity.
2. **Domain-Specific Fine-Tuning**:
    - Training Tesseract on business card-specific datasets.

ISSN: 2583-6129
DOI: 10.55041/ISJEM02205

- Incorporating multilingual support for global applicability.

3. **Validation Mechanisms**:
    - Real-time correction using APIs like Google Vision and regex-based format checks.

5. RESULTS AND DISCUSSION

The proposed system was tested on a dataset of 5,000 business cards, exhibiting diverse layouts and languages. Key performance metrics included:

- **Accuracy**: 94.8% (higher than existing solutions like vanilla Tesseract at 87.2%).
- **Processing Time**: Average of 1.2 seconds per card.
- **User Satisfaction**: Positive feedback on the Streamlit interface's usability.

Challenges such as handling cluttered layouts were mitigated through preprocessing and domain-specific training. Comparative analysis with other systems (e.g., Smith & Johnson, 2021) demonstrated superior robustness and flexibility.

6. CONCLUSION AND FUTURE ENHANCEMENTS

This project demonstrates the potential of advanced OCR techniques combined with user-friendly tools like Streamlit to revolutionize business card information extraction. Future enhancements include:

1. **Edge Device Optimization**:
    - Adapting the system for mobile and IoT devices.
2. **Cloud Integration**:
    - Leveraging platforms like AWS or Google Cloud for scalable data processing.
3. **Improved Feedback**:
    - Integrating speech synthesis for auditory feedback.
4. **Advanced Layout Parsing**:
    - Employing transformer-based models for better layout interpretation.
5. **Real-Time Multi-Language Translation**:
    - Translating extracted text into user-preferred languages instantly.

ISSN: 2583-6129
DOI: 10.55041/ISJEM02205

REFERENCES

1. Smith, A., & Johnson, P. (2021). Optical Character Recognition and its Applications in Business Card Data Extraction. *Journal of Computer Vision and Applications.*
2. Zhang, X., & Lee, K. (2020). A Survey on OCR Techniques: From Traditional Methods to Deep Learning Approaches. *International Journal of Pattern Recognition and Machine Intelligence.*
3. Tesseract OCR Team. (2024). Tesseract: Open Source OCR Engine. https://github.com/tesseract-ocr/tesseract
4. Kaur, M., & Choudhary, S. (2022). Machine Learning Approaches for Text Recognition in Images. *Journal of Machine Learning Research.*
5. Jain, V., & Agarwal, N. (2023). OCR-Based Business Card Parsing: A Review. *International Journal of Artificial Intelligence and Data Processing.*
6. Johnson, T., & Lee, M. (2021). Data Preprocessing and Image Augmentation for OCR Improvement. *Journal of Data Science and Analytics.*
7. Wang, Y., & Zhu, Z. (2023). Enhancing OCR Accuracy through Deep Learning Models. *IEEE Transactions on Neural Networks and Learning Systems.*

