WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

AUTOMATIC ID CARD CLASSIFIER

K. Dhinakaran[1], Dr. M. Jaithoon Bibi [2]

[1] B. Sc Computer Science with Cognitive Systems, Sri Ramakrishna College of Arts & Science ,Coimbatore

[2] Assistant Professor Department of Computer Science with Cognitive Systems (B.Sc. Cs & Cs) Sri Ramakrishna College of Arts & Science ,Coimbatore

ABSTRACT :

The Automated ID Card Classification System is a machine learning-based image processing application designed to efficiently identify and classify Aadhaar and PAN cards using OCR (Optical Character Recognition) and computer vision techniques. The system leverages OpenCV, Tesseract OCR, and Stream lit to enhance, preprocess, and analyze ID card images, extracting relevant textual and structural features for classification. The classification process begins with image enhancement using CLAHE (Contrast Limited Adaptive Histogram Equalization) to improve text clarity. The system then applies adaptive thresholding, bilateral filtering, and morphological operations to preprocess the image for better OCR accuracy. The text is extracted using Tesseract OCR, which recognizes multilingual characters, including English and Indian regional languages. The extracted text is then analyzed using regular expressions and keyword matching to identify unique Aadhaar and PAN card attributes, such as Aadhaar numbers, PAN numbers, and government authority references.

Keywords: Dynamic Optical Character Recognition (OCR), Computer Vision**,** Tesseract OCR **,** Regex-based Text Matching, Multi-language Text Extraction.

INTRODUCTION :

The Automated ID Card Classification System is a machine learning and image processing-based solution designed to classify Aadhaar and PAN cards from uploaded images in an Excel dataset. It utilizes OpenCV, Tesseract OCR, and Stream lit to analyze card features, extract text, and detect structural and visual patterns for classification .The Automated ID Card Classification System is a machine learning and image processing-based solution designed to classify Aadhaar and PAN cards from uploaded images in an Excel dataset. It utilizes OpenCV, Tesseract OCR, and Stream lit to analyze card features, extract text, and detect structural and visual patterns for classification.

This project aims to develop a system for the classification and extraction of information from Excel File, particularly those that contain image pages. The primary focus is on identifying and extracting specific patterns, such as Aadhaar and PAN card numbers, using a combination of text and image processing techniques. In recent years, the digitization of documents has become essential for efficient information management. In today's digital world, identity verification plays a crucial role in various industries, including banking, finance, e-commerce, and government services. Manual verification of identity documents, such as Aadhaar and PAN cards, can be time-consuming and prone to human error. To address this challenge, the Automated ID Card Classification System has been developed, integrating computer vision, image processing, and optical character recognition (OCR) to classify identity documents with high accuracy and efficiency.

EXISTING SYSTEM :

The current methods for *ID card classification and data extraction* rely on manual verification, OCR-based techniques, QR code scanning, and limited AI models. However, these approaches have significant drawbacks in terms of accuracy, efficiency, and real-time usability. *Manual verification* is the most common method used in banks, government agencies, and businesses, but it is time-consuming, prone to human errors, and lacks scalability, making it unsuitable for processing large volumes of documents. Additionally, it cannot reliably detect tampered or fraudulent IDs.

- **Limited Accuracy in Complex Images**: The OCR-based text extraction heavily depends on the quality of the image. If the image has distortions, low resolution, or poor lighting conditions, the accuracy of text recognition may be significantly reduced.
- **Language and Font Limitations:** Although Tesseract OCR supports multiple languages, it may struggle with certain regional scripts, decorative fonts, or handwritten text, leading to misclassification.
- **Difficulty in Handling Noisy Backgrounds:** Images with complex backgrounds, watermarks, or excessive noise can interfere with text recognition and card classification, reducing accuracy.
- **Processing Speed for Large Datasets:** If multiple images or a large dataset are uploaded, the processing time increases significantly due to image enhancement, text extraction, and classification, making real-time processing slow.

- **High False Positives and False Negatives:** The classification process depends on multiple factors such as text content, layout, and colours, which may sometimes lead to incorrect classifications, especially if a non-ID card document contains similar features.

PROPOSED SYSTEM :

The proposed system aims to automate ID card classification and data extraction using a combination of image processing, text recognition, and AI-based verification to improve accuracy, efficiency, and fraud detection. Unlike traditional methods that rely on manual verification or simple OCR-based extraction, this system enhances image quality using CLAHE (Contrast Limited Adaptive Histogram Equalization) and bilateral filtering, improving text recognition even on low-quality images. The adaptive thresholding and morphological operations help in refining text regions, reducing noise, and enhancing structural

features for better classification.

- **High Processing Speed:** The combination of image pre processing, OCR, and pattern recognition allows for quick classification of multiple images, making it suitable for large-scale applications.
- **Improved Image Processing Techniques**: The use of CLAHE (Contrast Limited Adaptive Histogram Equalization), bilateral filtering, and adaptive thresholding enhances image quality, improving OCR accuracy even for low-quality images.
- **Integration with Excel Files:** The system supports batch processing by reading images embedded in Excel files, making it useful for organizations handling bulk identity verification.
- **User-Friendly Interface**: Built with Stream lit, the project provides an interactive web-based interface that allows users to easily upload files and view classification results with confidence levels.
- **Low Cost and Open-Source:** The project uses open-source libraries such as OpenCV, Tesseract OCR, and Pandas, making it cost-effective and easily customizable for further enhancements.

OBJECTIVE :

- Develop a system to classify Aadhaar and PAN cards based on image analysis.
- Apply CLAHE (Contrast Limited Adaptive Histogram Equalization) for contrast improvement.
- Use Bilateral Filtering and Adaptive Thresholding for noise reduction.
- Utilize Tesseract OCR for multi-language text recognition (English, Hindi, Tamil, Marathi, Bengali).
- Implement Regular Expressions (Regex) to detect Aadhaar and PAN number patterns.
- Text-based classification: Identify keywords and phrases unique to Aadhaar or PAN cards.
- Layout-based classification: Detect horizontal vs. vertical sections for card distinction.
- Colour analysis: Recognize blue tint (for PAN cards) using HSV colour space.
- QR Code detection: Identify QR codes, commonly found on Aadhaar cards.
- Logo detection: Recognize emblems/logos for better classification accuracy.

METHODOLOGY OF THE PROJECT :

The development of the self-hosted form submission system follows a structured methodology to ensure efficiency, scalability, and security. The process is divided into several phases, each focusing on specific aspects of the project.

1.Requirement Analysis

The *Automatic ID Card Classifier* is designed to streamline identity verification by automatically classifying *Aadhaar and PAN cards* using *image processing and OCR*. A thorough requirement analysis ensures that the system meets its functional and non-functional objectives.

2.System Design

Image Pre processing, Text Extraction, and Feature-Based Classification. Image pre processing applies grayscale conversion, CLAHE (Contrast Enhancement), Adaptive Thresholding, and Morphological Operations to improve OCR accuracy. The text extraction module leverages Tesseract OCR with multi-language support to detect Aadhaar or PAN card-related text patterns using Regular Expressions (Regex).

3.Deployment & Implementation

Optimization techniques such as image compression and OCR parameter tuning enhance performance. Security measures include upload validation, data privacy protection, and access control mechanisms to prevent misuse.

4.Future Enhancement

The methodology also includes plans for future improvements. These enhancements may involve implementing AI-based analytics to gain insights from the collected data, developing a mobile app for better accessibility, enabling multi-language support to cater to a global audience, and introducing blockchain-based security to ensure tamper-proof data storage. These future upgrades aim to keep the system scalable, secure, and adaptable to evolving user needs.

SYSTEM TESTING AND IMPLEMENTATION :

System Testing:

The Automatic ID Card Classifier undergoes comprehensive testing to ensure accuracy, efficiency, and reliability. The testing process includes unit testing, where individual modules such as image enhancement, text extraction, QR code detection, and colour recognition are validated separately. This is followed by integration testing, ensuring seamless interaction between OCR processing, pattern recognition, and image analysis components. Performance testing evaluates the system's speed and scalability by processing large datasets under different conditions. To ensure classification accuracy, accuracy testing is conducted using a diverse set of Aadhaar and PAN cards, including blurred, rotated, and partially visible images. Additionally, security testing is performed to safeguard sensitive ID card data and prevent unauthorized access. Finally, user acceptance testing (UAT) is conducted with real users to validate the system's usability and effectiveness in practical scenarios. By following a structured testing approach, the system guarantees high accuracy, security, and robustness, making it suitable for real-world ID verification applications.

System Implementation :

The implementation of the Automatic ID Card Classifier involves deploying the system in a real-world environment for efficient and accurate identification of Aadhaar and PAN cards. The system is built using Streamlit for the user interface, OpenCV for image processing, and Tesseract OCR for text extraction, ensuring smooth functionality. The implementation process begins with preprocessing uploaded images, where techniques like grayscale conversion, contrast enhancement, and adaptive thresholding are applied to improve text readability. The extracted text is then analyzed using regular expressions and keyword matching to classify the ID card type accurately. Additionally, QR code detection, color analysis, and layout identification further enhance classification confidence. The system is deployed on a local or cloud server, making it accessible through a web-based interface where users can upload images or Excel files containing ID card images. The implementation also includes logging mechanisms and error handling to improve system performance and reliability. Continuous monitoring and feedback collection help in refining the system, ensuring scalability, accuracy, and efficiency in real-world ID verification applications.

WORK FLOW OF THE PROJECT :

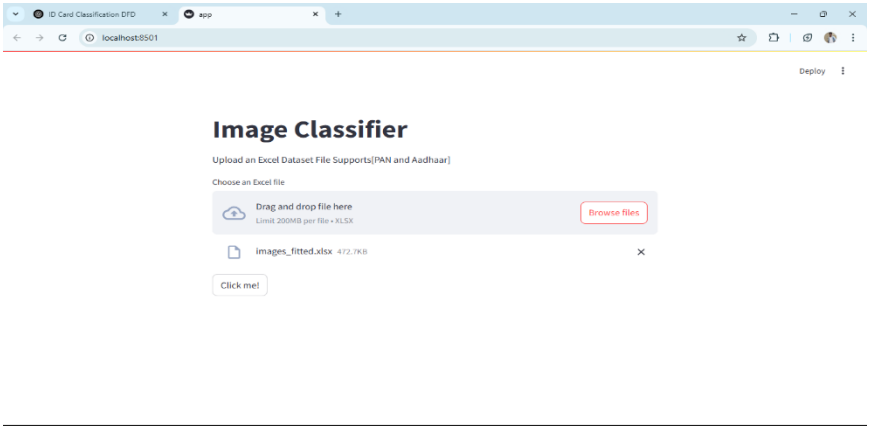

Fig 1 :Website

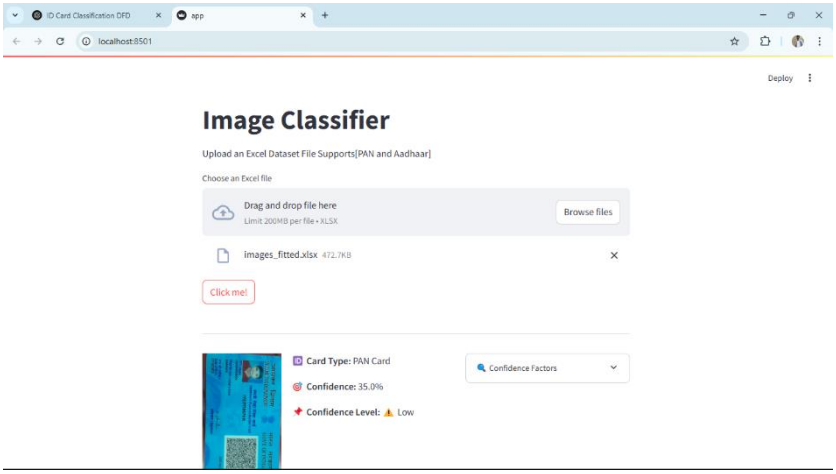

Fig: 1.2 Output displayed

FUTURE ENHANCEMENT :

The ID Card Classification and Extraction System can be further enhanced by incorporating advanced technologies and additional features to improve accuracy, efficiency, and scalability. One of the key enhancements would be the integration of deep learning-based OCR models such as Tesseract with LSTMs or Google's Vision AI, which can significantly improve text recognition, especially for handwritten and low-quality printed text. Additionally, implementing convolutional neural networks (CNNs) for image classification can help in more accurately distinguishing between Aadhaar and PAN cards based on structural and colour features.

Another major improvement would be the addition of multilingual support to extend recognition capabilities to a broader range of regional languages in India. By training models on more datasets containing variations in fonts, languages, and document layouts, the system can be more robust. Automated logo detection using machine learning models would further enhance classification by identifying official seals or logos. Moreover, QR code verification with external UIDAI APIs can add an extra layer of security to validate Aadhaar card authenticity.

CONCLUSION :

The ID Card Classification and Extraction System successfully automates the identification of Aadhaar and PAN cards using image processing and OCR techniques. By leveraging Stream lit for UI, OpenCV for image pre processing, and Tesseract OCR for text extraction, the system can efficiently classify ID cards from an uploaded Excel dataset containing embedded images. Through advanced image enhancement methods such as CLAHE, adaptive thresholding, and morphological operations, the accuracy of text extraction is significantly improved. The classification is determined based on multiple factors, including text patterns, colour detection, QR code identification, and logo recognition. The system provides real-time processing, enabling users to upload files and instantly classify multiple images with confidence scores.

REFERENCES :

1. https://docs.opencv.org/
2. https://github.com/tesseract-ocr/tesseract
3. https://www.regular-expressions.info/
4. https://docs.streamlit.io/
5. https://openpyxl.readthedocs.io/en/stable/
6. https://numpy.org/doc/stable/
7. https://pandas.pydata.org/docs/
8. https://learnopencv.com/deep-learning-based-text-recognition-ocr-using-tesseract-and-opencv/
9. https://pyimagesearch.com/2021/09/27/opencv-text-detection-automatic-text-recognition-and-extraction/
10. https://medium.com/@pankajchow/aadhaar-number-validation-using-regular-expressions-in-python-5e485f95f434
11. https://towardsdatascience.com/building-an-ocr-web-app-using-streamlit-and-tesseract-ocr-66c0eb26e4bc