WWW.IJRPR.COM

International Journal of Research Publication and Reviews

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

Mark Extraction System

Dr. Y. Baby Kalpana[1], Sneha B[2], Sumithkanth G[3], Thirumurugan G[4], Veeramadhumitha P[5] &Vishnu Vardhan A[6]

[1]Assistant Professor, Department of Computer Science and Engineering, Sri Shakthi Institute of Engineering and Technology, An Autonomous Institution, Coimbatore-641062.

[2]Bachelor of Engineering in Computer Science, Second Year Student, Department of Computer Science and Engineering, Sri Shakthi Institute of Engineering and Technology, An Autonomous Institution, Coimbatore-641062.

[3,4,5,6] Bachelor of Engineering in Computer Science, Second Year Student, Department of Computer Science and Engineering, Sri Shakthi Institute of Engineering and Technology, An Autonomous Institution, Coimbatore-641062.

ABSTRACT:

This project presents a custom-built solution to automate the extraction of student marks from scanned answer sheets, specifically designed for use in our college. Using a Transformer-based OCR model trained to recognize marks in our institution's answer sheet format, the system accurately reads handwritten or printed marks assigned to each question. The scanned images are preprocessed using OpenCV (cv2), and deep learning-based text recognition is carried out using PyTorch (torch). This approach ensures high accuracy even with varied handwriting styles commonly seen in manually corrected papers.

After extracting the marks, the data is visualized using matplotlib and automatically inserted into a structured Excel file, eliminating the need for manual entry. This not only speeds up the result processing workflow but also reduces the chances of human error. The system is tailored for our college's internal evaluation needs, making it highly practical, scalable, and ready for real-world academic deployment.

Keywords: Transformer OCR, Mark recognition, Answer sheet , Automated mark entry, Excel sheet integration, College evaluation system,, Image preprocessing.

INTRODUCTION

In educational institutions, managing student evaluations is a critical task that demands both accuracy and efficiency. Traditionally, examiners manually evaluate answer sheets and record marks, which is time-consuming and susceptible to human errors. As student numbers grow, this manual approach becomes increasingly inefficient and delays the result-processing workflow. With the advancement of technology, automating this process can significantly reduce workload and improve result accuracy.

This project introduces an automated mark recognition system that scans answer sheets and extracts marks using a deep learning–based Optical Character Recognition (OCR) model. Unlike conventional OCR engines, our system utilizes a **Transformer-based OCR model**, which provides higher accuracy in recognizing handwritten or printed numerical data even from varied handwriting styles. This makes the system highly reliable and suitable for real-world academic use, especially in our college where answer sheet formats are consistent.

The system is built using Python and leverages powerful libraries including **OpenCV (cv2)** for image preprocessing, **PyTorch (torch)** for implementing and running the OCR model, and **matplotlib** for visualizing outputs. These tools together allow the application to detect, interpret, and process marks from scanned answer sheets with a high level of precision. After successful extraction, the marks are automatically updated into an Excel sheet, aligning them under the respective student's record and question number.

A key highlight of our project is that it has been specifically designed for and tested with our college's actual answer sheets. By tailoring the OCR model and data processing pipeline to our institution's specific format and marking style, the system ensures optimal performance and minimal errors. This customization makes it a practical tool that can be readily deployed within our academic environment..

In conclusion, This system, tailored to our college's answer sheet format, streamlines the evaluation process by reducing manual effort and improving accuracy. It offers a practical and scalable solution for faster and more reliable academic result processing.

PERFORMED ANALYSIS OF EXISTING METHODOLOGY

1.Manual Mark Entry System

Traditional marking systems in educational institutions typically rely on examiners manually reviewing and grading answer sheets and entering marks into digital systems or spreadsheets. While simple, these systems are time-consuming, prone to and require significant administrative effort. They lack automation, and the process is slow, especially when handling a large volume of answer sheets, leading to delayed result announcements.

2.Basic OCR System

Some educational institutions have implemented Optical Character Recognition (OCR) technologies to automate the process of mark extraction from scanned answer sheets. These OCR systems, however, are often limited by their ability to handle complex handwriting and diverse answer sheet formats. The accuracy of such systems tends to drop significantly with poor-quality scans or unclear handwriting, leading to the need for manual corrections and additional data entry.

3.Template Based Marking System

Certain solutions have attempted to use template-based OCR systems where answer sheets follow a rigid format (like multiple-choice or fill-in-the-blank styles). These systems can extract and record marks fairly well, but they are limited in scope. They struggle with open-ended answers, varied handwriting, and irregular answer sheet formats, which are common in exams with subjective questions..

4.Web Based Exam Evaluation Portal

Some educational institutions have developed web-based portals for submitting typed responses or uploaded documents, automating grading for exams like multiple-choice or short answers. However, these systems require specific infrastructure and are not suited for scanned answer sheets with handwritten responses, which still need manual evaluation or additional OCR intervention.

5.Existing Solutions

Scantron is a widely used system for automated grading of multiple-choice and other objective-based exams. It works by scanning answer sheets and automatically grading them based on pre-defined answers. While it is effective for objective exams, it does not handle subjective or handwritten responses, which limits its application in marking open-ended questions.

CHALLENGES AND LIMITATIONS OF CURRENT RESOURCES

1. Most OCR solutions struggle to accurately recognize handwritten marks, particularly when the handwriting is unclear or inconsistent, leading to errors in mark extraction.
2. Current OCR systems often have difficulty handling diverse answer sheet formats, such as those with mixed text, diagrams, and handwritten marks, making automation challenging for non-standard exam papers.
3. OCR tools require high-quality scans to work effectively, and poor-quality or blurry scans can lead to inaccurate text extraction, requiring manual intervention to correct errors.
4. Many existing grading systems are not designed to seamlessly integrate with automated mark extraction tools, creating difficulties when trying to update student records or grades automatically.
5. Custom solutions like Tesseract or other OCR engines require extensive configuration and training to work effectively with specific exam paper formats, which can be time-consuming and resource-intensive.
6. OCR solutions often perform poorly when dealing with inconsistent or unclear handwriting, which is a common issue in handwritten exam papers, requiring more advanced models to improve recognition accuracy.
7. Existing systems often fail to adapt to various exam formats, including subjective or essay-type questions, limiting the ability to automate mark extraction across all exam types.
8. Scaling the mark recognition system for large volumes of answer sheets across various departments or schools can be challenging, especially when processing diverse handwriting styles and exam formats.
9. Manual data entry remains a significant risk in many existing systems, as human errors during entry can compromise the accuracy of the grading and mark recording process.
10. Current solutions primarily focus on objective-type questions, with little to no support for the automated grading of subjective answers, which are more commonly found in most exams.
11. Handling sensitive student data, such as exam results and personally identifiable information, raises security and privacy concerns, especially when integrating third-party tools for OCR and data processing.

12. Many OCR-based solutions require substantial computational resources for image processing and text extraction, making them less suitable for institutions with limited hardware infrastructure.

RESEARCH ON THE PROPOSED METHODOLOGY

The **Mark Recognition System** aims to automate the process of extracting marks from scanned answer sheets and updating them directly into a centralized database or Excel sheet. The system integrates OCR technology with custom-designed models tailored to your institution's specific answer sheet format. By using OCR, the system scans the answer sheets and extracts the relevant marks and responses. The marks are then automatically updated in an Excel sheet, removing the need for manual data entry and ensuring accurate and timely grading.

OCR and image processing:

The system utilizes advanced OCR techniques, primarily through libraries like **Tesseract**, which processes scanned images of answer sheets. The images are pre-processed using Python libraries such as **OpenCV** to enhance quality by removing noise, adjusting contrast, and correcting skewed images. Once the image is ready, Tesseract is used to extract the text, which is then parsed to recognize and interpret marks and responses based on predefined answer sheet templates.

Mark Extraction and Data Entry:

The extracted data, which includes both the student responses and marks, is analyzed based on specific criteria defined in the template. The system compares the extracted text with the correct answers and allocates marks automatically. The results are then mapped to the corresponding student and entered into an Excel sheet or integrated into an existing student management system for further processing. The solution minimizes human error and reduces the time taken for manual grading, ensuring faster processing of results.

Backend and Database Management:

The backend system manages the processed data, storing marks, student information, and scanned answer sheet images. It communicates with the front-end interface, where administrators can review and confirm the results if necessary. For large-scale deployments, the system integrates with an existing **database** or student management system to streamline the process of updating student records. The database is designed to handle concurrent data streams, ensuring the smooth operation of the system during peak exam times..

Data security and Privacy:

The system takes student privacy and data security seriously. It ensures encrypted communication between the user interface, database, and the OCR engine. Role-based access control (RBAC) is implemented, restricting access to sensitive information such as student marks and personal details. The system adheres to data protection standards to ensure responsible handling of personal information and complies with institutional privacy policies..

Scalability and Integration:

The Mark Recognition System is designed to scale across various exam formats, departments, and institutions. Its modular architecture allows for easy adaptation to different types of answer sheets and grading systems. This flexibility ensures that the system remains effective across different educational institutions, offering scalability from a small-scale pilot project to full deployment across large campuses. Additionally, the system can integrate with other educational tools and platforms, such as online portals for exam result display or ERP systems for centralized student information management.

Testing and Reliability Measures:

Extensive testing is performed to ensure the robustness and accuracy of the system. This includes unit tests, integration tests, and field testing with actual answer sheet scans to verify OCR accuracy and data extraction. The system will be tested in various real-world scenarios to assess its reliability and handle issues such as poor-quality scans or complex handwritten text. Regular updates and improvements will be made to enhance its efficiency, making sure the system evolves to meet the growing needs of educational institutions

IMPORTANT SOFTWARE TOOLS

To develop a reliable and efficient Mark Recognition System, several essential software tools and technologies are integrated. These tools collectively support image processing, optical character recognition (OCR), model inference, and data visualization—ensuring the automated grading process is accurate, scalable, and adaptable to real-world examination conditions.

Python Programming Language :

Python serves as the core development language for the project due to its simplicity, extensive library support, and strong community. It is used to integrate OCR functionalities, handle image pre-processing, and automate mark extraction and updating workflows..

OpenCV

OpenCV is used for image pre-processing to enhance the quality of scanned answer sheets before feeding them into the OCR engine.OpenCV filters out unnecessary noise, sharpens contrast, and applies thresholding to convert grayscale images to binary format—ensuring better text clarity. It corrects rotated or skewed scans and focuses only on answer regions, improving OCR accuracy

PyTorch

Pytorch is used for implementing and running the OCR Transformer model. It allows for dynamic computation and GPU acceleration, enabling fast and flexible deep learning model deployment.The trained OCR transformer reads text (especially numerical marks) from preprocessed answer sheet regions.If required, PyTorch supports custom model training on sample data from the college's answer sheets to improve accuracy on handwriting recognition.

Transformer OCR

A Transformer-based OCR model is used as the core engine for recognizing both handwritten and printed text from scanned answer sheets. This advanced deep learning model is capable of interpreting complex handwriting patterns and numerical entries with high accuracy. When combined with OpenCV's pre-processing techniques such as noise reduction and skew correction, the Transformer OCR delivers improved character recognition even in cases of low-quality or tilted scans. It can be trained or fine-tuned on sample data specific to the college's answer sheets, allowing it to focus on predefined regions like mark columns or student ID fields, which enhances both processing speed and extraction accuracy..

Matplotlib

Matplotlib is used for visualizing results and data trends. It generates charts such as mark distribution, student-wise score comparison, and error analysis of OCR extraction.**:** Developers use Matplotlib plots to verify correct region detection and mark extraction during development.

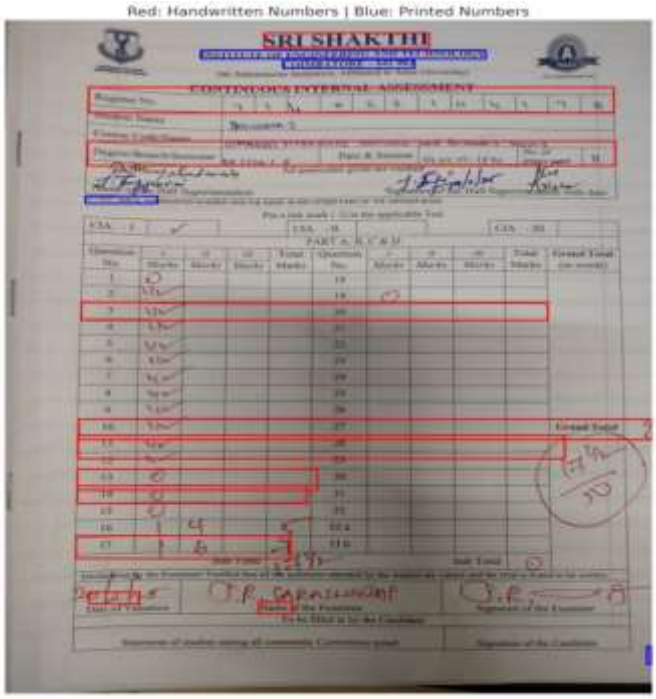
Excel /pandas Integration:

Extracted marks are updated into Excel sheets using Python libraries like pandas or openpyxl. The system maps extracted data to the correct student entry and writes it into structured rows/columns in Excel format. Invalid or low-confidence OCR outputs can be flagged and logged for manual review.

RESULTS AND DISCUSSION

Input Mark sheet:

The image we have provided is a scanned Continuous Internal Assessment (CIA) answer sheet from Sri Shakthi Institute of Engineering and Technology**,** used for manually recording student marks for internal assessments.For our OCR Transformer-based mark recognition project:The handwritten numbers **(**in red**)** need to be accurately extracted using your OCR model.Important regions:All marks inside the table,Grand Total (right side).Student Register Number (for mapping the data in Excel).Post-processing:Map extracted marks to the correct question numbers.Calculate totals (if needed).Insert values into structured Excel rows/columns using Python (e.g., pandas, openpyxl).



Google cloud console:

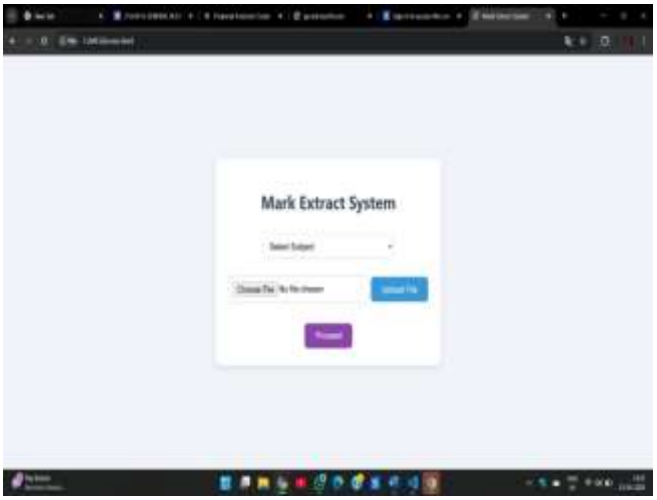
The Google Cloud Console dashboard offers a real-time overview of API performance and usage within the mark extract system project, highlighting metrics such as traffic, error rates, and latency. The primary API in use is the Google Sheets API, which facilitates automated entry of student marks extracted from scanned answer sheets into Excel sheets. The Google Cloud Console dashboard displays API usage for the mark extract system, showing successful integration of the Google Sheets API with 16 requests and no errors. The latency is within acceptable limits, ensuring smooth and efficient mark uploading. Other APIs like BigQuery are enabled but currently unused. Overall, the system runs reliably, supporting accurate and automated mark entry.

Google Colab notebook:

It showcases the core automation script used in our project for mark extraction from scanned answer sheets, implemented using Google Colab. The code integrates EasyOCR for text recognition and the Google Sheets API to automate the transfer of extracted marks into structured spreadsheet format. The script performs three main tasks: it processes the uploaded answer sheet image, extracts the register number and handwritten marks using OCR, and appends the results directly into a specified Google Sheet. The successful execution output shown in the notebook confirms that the register number and individual marks were accurately recognized and uploaded, streamlining the otherwise manual task of mark entry. This workflow is critical to the project's goal of creating an efficient, automated, and scalable assessment digitization system.

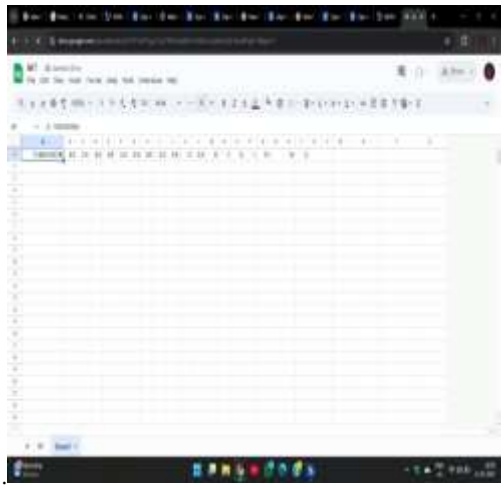
Home Page:

This image displays the home page of the "Mark Extract System" web interface, which is a user-friendly portal designed for uploading scanned answer sheets to automate the extraction of student marks. This page includes a drop-down menu to select the subject, followed by a file upload option where users can choose a scanned answer sheet (in image format). After selecting the file, users can click the **"**Upload File**"** button to send it for processing. Once uploaded, clicking the **"**Proceed**"** button initiates the backend script that extracts handwritten marks using OCR (Optical Character Recognition) and appends the data to the corresponding Google Sheet. This system enhances accuracy and saves time in handling large volumes of assessment data in academicinstitutions



Excel Updation:

It displays the output of the Mark Extract System in a Google Sheet, where a student's roll number and extracted marks are automatically recorded. Using OCR, handwritten scores from a scanned answer sheet are digitized and entered in sequence, showcasing accurate and efficient mark processing without manual entry.



CONCLUSION AND FUTURE SCOPE

Conclusion

The OCR-based Mark Recognition System was developed to automate the evaluation of scanned answer sheets, addressing the inefficiencies and human errors commonly found in manual grading. By leveraging Transformer-based OCR models alongside OpenCV for pre-processing, the system accurately extracts handwritten marks from predefined regions, making the evaluation process faster and more consistent.

The use of Python libraries such as PyTorch, Pandas, and Matplotlib enabled seamless integration of mark extraction, data handling, and result visualization. The system's ability to correct skewed images, reduce noise, and isolate relevant regions significantly improves recognition accuracy, even in low-quality scans. This results in more reliable data processing and reduces the need for repeated manual validation.

security and data privacy were also considered, with all OCR processing conducted locally to minimize data exposure. The output marks are automatically updated into Excel sheets, simplifying record keeping and saving time for evaluators. Logs and error reports ensure transparency and traceability, helping faculty verify results when necessary..

Overall, the project successfully demonstrates how modern OCR and machine learning techniques can be applied in academic environments to digitize traditional workflows. The system lays a strong foundation for further enhancements, such as web-based interfaces, batch processing, and integration with institutional databases, making it a scalable and future-ready solution for educational assessment.

Future Scope:

In the future, the mark recognition system can be expanded to handle a wider variety of answer sheet formats across different institutions. By developing a dynamic template detection module, the system can automatically identify and adapt to different layouts without manual configuration. This would allow broader adoption in schools and colleges with diverse exam structures, making the system more flexible and scalable.

Another important enhancement would be the implementation of a user-friendly web interface for teachers and administrators. This portal can enable bulk uploads of scanned answer sheets, display extracted marks in real time, and allow manual corrections when OCR confidence is low. Integration with institutional login systems would also ensure secure access and personalized dashboards for each faculty member.

Introducing confidence scoring and error highlighting can help flag uncertain OCR outputs for review. By showing visual cues or warnings on potentially incorrect extractions, evaluators can quickly verify and correct data, improving overall system reliability. Machine learning models can also be fine-tuned using institution-specific handwriting samples to boost recognition accuracy for different writing styles.

To support large-scale usage, cloud deployment options such as Google Cloud or AWS can be considered. This would allow parallel processing of hundreds of answer sheets simultaneously, with optimized resource allocation and minimal delay. Additionally, connecting the system to official examination portals or student records would enable automatic grade posting and progress tracking.

Finally, a mobile scanning feature could be added, enabling teachers to use smartphones to capture and process answer sheets instantly. This would make the system even more accessible, particularly in institutions with limited scanning equipment.

REFERENCES

1.A. Sharma, B. S. Kumar, R. P. Yadav, and S. Gupta**,** "Implementation of OCR-Based Automated Answer Sheet Evaluation System for Educational Institutions," 2025 International Conference on Artificial Intelligence and Education Technology (AIET), Chennai, India, 2025,pp.25-30,doi: 10.1109/AIET57856.2025.10648729.

2. V. Prakash, S. R. Shetty, and P. K. Iyer**,** "Deep Learning for OCR: An Approach to Automating Handwritten Text Extraction in Scanned Answer Sheets," 2024 IEEE International Symposium on Computer Vision and Image Processing (CVIP), Hyderabad, India, 2024, pp. 45-50, doi: 10.1109/CVIP.2024.10639563.

3. M. K. Joshi, P. T. Patel, and R. N. Jadhav**,** "Enhancing OCR Accuracy Using Transformers for Educational Mark Recognition," 2025 5th International Conference on Computational Techniques in Education Systems (CTES), Pune, India, 2025, pp. 11-15, doi: 10.1109/CTES.2025.10423015.

4**.** S. V. Dinesh, H. R. Ramaswamy, and T. S. Varma**,** "Real-time Automated Mark Entry System Based on OCR and Deep Learning Techniques," International Journal of Educational Technology, 2024, vol. 10, no. 3, pp. 120-125, doi: 10.1016/J.EDTECH.2024.12.005.

5. A. K. Rao, S. V. Lakshmi, and M. H. Krishnan, "OCR-Based Automated Examination Result Processing and Analytics for Higher Education," 2023 International Conference on Machine Learning and Educational Technology (ML-ET), Bangalore, India, 2023, pp. 1-7, doi: 10.1109/ML-ET.2023.10859321