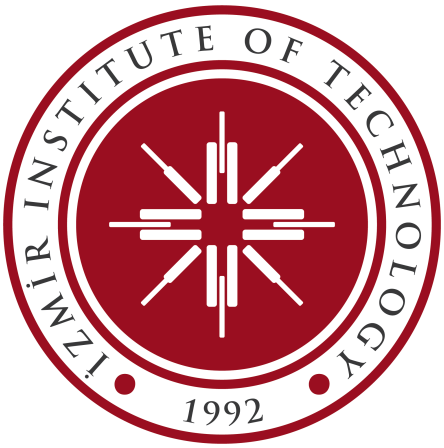
SEDS536 Image Recognition
Project Report
Spring 2025

Project Title: Handwritten Exam Paper Recognition Using Open-Source OCR Models

January 4, 2026



- Mustafacan Koç - 323011014

Abstract

Manual grading of handwritten exam papers is time-consuming, error-prone, and difficult to scale in educational settings. This project presents an automated optical character recognition (OCR)–based system designed to extract and interpret handwritten content from student exam papers in order to support automated grading and rapid feedback. The proposed approach applies image preprocessing techniques such as paper detection, cropping and noise reduction to improve input quality, followed by segmentation of relevant answer regions. Handwritten text recognition is performed using state-of-the-art OCR models optimized for handwritten data. The extracted text is then normalized to reduce recognition errors and enable further evaluation or comparison with expected answers. The system focuses on short answers and numeric responses commonly found in exam papers, where automation provides the greatest benefit. The results demonstrate that combining targeted preprocessing with modern OCR models can significantly improve recognition reliability for handwritten exam documents. This application aims to reduce instructors' grading workload, increase consistency in evaluation, and enable faster feedback for students, illustrating the practical potential of OCR technologies in educational assessment workflows.

1 Introduction

Handwritten examinations remain widely used in educational institutions due to their flexibility, low cost, and suitability for assessing problem-solving and short-answer responses. Despite the increasing adoption of digital learning platforms, many instructors continue to rely on paper-based exams, especially in large classes and formal assessments. However, evaluating handwritten exam papers is a labor-intensive and repetitive process that requires significant time and effort from instructors. As class sizes grow, manual grading becomes difficult to scale and may lead to delays in providing feedback to students. In addition, prolonged grading processes can introduce inconsistencies caused by fatigue and subjective judgment. These limitations highlight the need for supportive technological solutions. Automating parts of the grading workflow, particularly the extraction of handwritten answers, offers the potential to reduce instructor workload and improve efficiency while preserving the familiar handwritten exam format used in many educational settings.

This work focuses on the problem of automatically extracting handwritten short-answer and numeric responses from student exam papers. Unlike printed documents, handwritten exams exhibit significant variability in writing styles, character shapes, spacing, and alignment, which makes reliable text recognition challenging. These difficulties are further amplified when exam papers are captured as images under non-ideal conditions, such as uneven lighting, shadows, or perspective distortions. As a result, standard optical character recognition (OCR) systems that perform well on printed or structured documents often fail to generalize to handwritten exam content. The problem addressed in this project is therefore not generic document OCR, but the robust recognition of handwritten exam responses under realistic and unconstrained conditions commonly encountered in educational settings.

In practical educational scenarios, exam papers are often captured using smartphone cameras rather than flatbed scanners, resulting in images that vary widely in quality and appearance. Such images commonly suffer from uneven illumination, shadows cast by the environment or the device itself, perspective distortion caused by angled capture, and background clutter surrounding the paper. These factors introduce noise that significantly degrades the performance of standard OCR systems, which typically assume clean, well-aligned document inputs. Without targeted preprocessing and document normalization, OCR models may incorrectly detect text regions or misinterpret handwritten characters. These challenges make handwritten exam OCR a non-trivial task and highlight the necessity of specialized processing steps to handle the variability present in real-world exam data. To address these challenges, this project proposes an end-to-end OCR-based pipeline designed for handwritten exam papers. The system integrates image preprocessing and page correction with text region detection, handwritten text recognition, and evaluation against ground-truth answers. The approach relies on existing open-source OCR models for handwritten text, emphasizing system integration and practical evaluation rather than the development of new recognition models.

This project contributes an applied and systematic investigation of handwritten exam paper recognition. It presents the design and implementation of a complete OCR pipeline tailored to exam scenarios, covering image preprocessing, region localization, text recognition, and evaluation. In addition, the project evaluates multiple open-source OCR models for handwritten text within the specific context of exam papers, providing comparative insights into their behavior under realistic conditions. Finally, the analysis highlights both the strengths of the proposed approach and its current limitations, offering a clear assessment of the challenges that remain for reliable handwritten exam OCR in practical educational settings.

2 Literature Review

Optical character recognition (OCR) and automated grading of handwritten exam papers have received increasing attention as educational institutions seek to reduce the manual effort involved in assessment. Prior work in this area can be broadly grouped into handwritten text recognition models, OCR-based exam evaluation systems, and recent approaches that integrate large language models into grading workflows.

Pangestu et al. Pangestu et al. (2023) investigated handwritten text recognition using a CNN–BLSTM–CTC architecture for line-level handwriting recognition. Their system was evaluated on IAM handwritten text images and achieved a test accuracy of 50.21% with a character error rate (CER) of 8.65%. While their results demonstrate the feasibility of neural handwriting recognition, the reported accuracy highlights the inherent difficulty of recognizing unconstrained handwritten text, even under relatively controlled dataset conditions.

Several studies have extended OCR systems toward automated grading applications. Lekshmy et al. Lekshmy et al. (2024) proposed a CNN-based OCR pipeline designed for structured exam response cards, leveraging fixed printed marks to guide segmentation. By training separate recognition models for digits, letters, and true/false responses, they reported recognition accuracies exceeding 98% across multiple answer types. However, their approach relies on highly structured input formats and predefined layouts, which limits its applicability to unconstrained exam papers captured under real-world conditions.

More recent work has explored hybrid OCR and natural language processing (NLP) pipelines for grading handwritten responses. Pawar et al. Pawar et al. (2025) introduced an OCR-based answer extraction system followed by multi-stage text correction and semantic grading using BERT embeddings and cosine similarity. Similarly, Mulla et al. Mulla et al. (2025) proposed an AI-powered evaluation framework combining OCR, NLP, and machine learning techniques to automate grading while emphasizing secure data handling. These approaches demonstrate the potential of combining recognition and semantic evaluation but remain sensitive to OCR errors introduced during handwriting recognition.

The use of large language models in exam grading has also been investigated. Liu et al. Liu et al. (2024) evaluated GPT-4 for grading semi-open handwritten university-level mathematics exam responses and reported average grading accuracies between 0.59 and 0.62. Their study showed that large language models can assist human graders but also revealed significant limitations, particularly in handwritten mathematical expression recognition and in extracting reliable text from full exam pages. These findings suggest that robust OCR and segmentation remain critical prerequisites for downstream grading.

In parallel, research has focused on improving handwritten exam recognition itself. Shi et al. Shi et al. (2025) proposed a multimodal handwritten exam text recognition framework that classifies handwriting types and applies specialized recognizers, achieving 86.63% accuracy on a heterogeneous dataset combining exam images with standard handwriting benchmarks. Their results highlight the benefits of context-aware and content-specific recognition strategies but require complex model architectures and diverse training data.

From a model perspective, Li et al. Li et al. (2022) introduced TrOCR, a Transformer-based end-to-end OCR architecture that integrates vision and language modeling without relying on external language models. TrOCR has since become a widely used open-source baseline for handwritten text recognition and serves as the foundation for many applied OCR pipelines.

In addition to OCR and NLP-based grading pipelines, Ghadekar et al. Ghadekar et al. (2025) proposed an automated handwritten exam evaluation framework that integrates optical character recognition with machine learning techniques and Retrieval-Augmented Generation (RAG). Their system applies OCR to extract handwritten answers and employs machine learning-based answer mapping, grammar and context checks, and block-diagram evaluation. When predefined model answers are unavailable, a RAG-based mechanism is used to generate reference answers dynamically. This approach highlights the potential of retrieval-augmented methods to increase flexibility in automated grading systems, while still depending heavily on the reliability of the initial OCR stage.

Overall, existing literature demonstrates that while high recognition accuracy is achievable in controlled or highly structured settings, handwritten exam OCR remains challenging under realistic conditions involving smartphone-captured images, diverse handwriting styles, and unconstrained layouts. Many systems either rely on strong layout assumptions or require extensive model training and dataset curation. In contrast, the present project focuses on evaluating and integrating existing open-source OCR models within a practical end-to-end pipeline for handwritten exam papers, emphasizing robustness, reproducibility, and an honest assessment of current limitations.

3 Methodology

3.1 System Overview and Data Acquisition

The proposed system follows an end-to-end pipeline for extracting handwritten content from exam papers and evaluating recognition performance. The overall workflow consists of page segmentation and normalization, text region detection, handwritten text recognition, answer structuring, and quantitative evaluation. The system is designed as an evaluation-focused pipeline and does not involve training or fine-tuning of recognition models.

The dataset used in this study consists of 90 images of handwritten exam papers captured using smartphone cameras. The images were collected from 10 different individuals, each providing their own handwriting, in order to reflect variability in writing styles. To further increase diversity, each exam paper was photographed multiple times under unconstrained conditions, including variations in lighting, camera angle, and background. This setup aims to approximate realistic data acquisition scenarios commonly encountered in educational environments.

All OCR models employed in the system are pre-trained and used as-is. As a result, the experimental setup is purely evaluative, focusing on analyzing the behavior and limitations of existing open-source OCR models when applied to handwritten exam papers captured under real-world conditions.

3.2 Image Preprocessing and Page Segmentation

Due to the unconstrained nature of smartphone-captured exam images, a dedicated preprocessing and page segmentation stage is applied before text detection and recognition. The goal of this stage is to isolate the exam paper

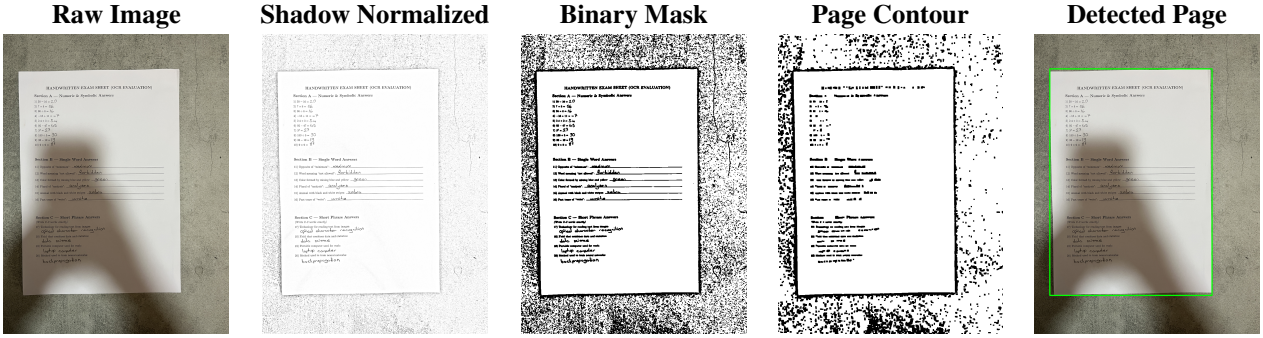


Figure 1: Visualization of the page segmentation process.

from the background, reduce illumination artifacts, and obtain a clean, cropped page image suitable for downstream OCR processing.

Each input image is first converted from color to grayscale to simplify subsequent processing steps. To address uneven lighting and shadow effects, illumination normalization is performed by estimating the background illumination using a large-scale Gaussian blur and normalizing the image accordingly. Contrast enhancement is then applied using CLAHE to stabilize local intensity variations and improve the separability between foreground content and background regions.

Following normalization, multiple binarization strategies are employed to generate candidate representations of the page region. These include global thresholding (Otsu) and adaptive thresholding, as well as an edge-based pipeline using Canny edge detection followed by dilation and closing. Contours are extracted from both mask-based and edge-based representations and scored based on geometric properties such as area, rectangularity, and proximity to image borders. The most plausible page contour is selected using this scoring mechanism, with fallback to a rotated rectangle when needed.

Once the page contour is determined, the exam paper is cropped from the original image using an axis-aligned bounding region, producing a cleaned page image that excludes most background clutter. Intermediate preprocessing and segmentation results are optionally saved for debugging and inspection purposes. Figure 1 illustrates the main stages of the preprocessing and page segmentation pipeline.

3.3 Text Region Detection and Handwritten Recognition

After page segmentation and normalization, text regions are localized on the cropped exam page using an OCR-based detection approach. PaddleOCR is employed to detect text regions at the line level, returning quadrilateral polygons that correspond to individual text lines. Line-level detection is preferred over word-level detection, as it better preserves the spatial structure of handwritten exam responses and reduces fragmentation caused by irregular handwriting.

Each detected text polygon is subsequently normalized through a perspective transformation. This step warps the quadrilateral region into an upright rectangular crop, correcting residual skew and perspective distortion introduced during image capture. As a result, the handwritten text recognition models receive consistently aligned line images, which improves recognition stability across varying capture conditions.

Handwritten text recognition is performed using multiple open-source OCR models, including PaddleOCR, EasyOCR, and TrOCR. All models are used in their pre-trained configurations without any fine-tuning. PaddleOCR and EasyOCR provide line-level text strings for each detected text region, while TrOCR outputs a single line-level text prediction per crop using both base and large pre-trained variants. Each normalized text line is independently processed by all models, enabling a comparative evaluation of their recognition behavior on handwritten exam data.

3.4 Answer Structuring and Evaluation

Following handwritten text recognition, the extracted text lines are organized into structured answers corresponding to individual exam questions. Question detection is performed at a high level by identifying question indices within the recognized text using pattern-based matching. Detected question numbers serve as anchors, and subsequent text lines are associated with the nearest preceding question index to form a complete candidate answer. This approach allows multi-line responses to be grouped without relying on fixed layouts.

Evaluation is conducted using two complementary metrics: exact-match accuracy and character error rate (CER). Exact-match accuracy is computed at the question level and considers a prediction correct if the ground-truth answer appears as a substring within any of the aggregated candidate answers for a given model. This definition reflects the practical goal of detecting correct responses despite minor recognition inconsistencies. CER is calculated by measuring the normalized Levenshtein distance between the ground-truth answer and the closest matching candidate, providing a fine-grained assessment of character-level recognition errors.

All evaluations are performed on pre-trained models without any training or adaptation. Metrics are computed for each experimental run and subsequently aggregated across runs to obtain model-level performance summaries. This aggregation enables a stable comparison of OCR models under varying capture conditions while accounting for variability introduced by repeated image acquisition.

4 Experiments & Results

4.1 Experimental Setup

The experimental evaluation is conducted on a dataset consisting of 90 smartphone-captured images of handwritten exam papers. The images were collected from 10 different individuals to capture variability in handwriting styles. Each exam paper was photographed multiple times under unconstrained conditions, including variations in lighting, camera angle, and background. No data splitting into training or validation sets is performed, as the study focuses exclusively on evaluation using pre-trained OCR models.

Four open-source OCR models are evaluated in the experiments: PaddleOCR, EasyOCR, TrOCR-base, and TrOCR-large. PaddleOCR is used both for line-level text detection and text recognition, while EasyOCR and TrOCR are applied only at the recognition stage. All models are used in their pre-trained configurations without any fine-tuning or domain adaptation. Each detected text line is independently processed by all models to enable a fair comparative analysis under identical input conditions.

Model performance is assessed using three evaluation metrics: exact-match accuracy, character error rate (CER), and character-level accuracy. Exact-match accuracy measures whether the predicted answer contains the ground-truth answer as a substring at the question level. CER is computed as the normalized Levenshtein distance between the ground-truth answer and the closest matching prediction, providing a fine-grained measure of character-level errors. Character-level accuracy is derived as one minus the character error rate and offers an interpretable indicator of per-character recognition correctness. All metrics are computed per experimental run and aggregated across runs to obtain stable model-level performance estimates.

4.2 Quantitative Results

Table 1 summarizes the quantitative performance of the evaluated OCR models on the handwritten exam dataset. Results are reported using exact-match accuracy, character-level accuracy, and character error rate (CER), aggregated across all experimental runs.

Table 1: Quantitative performance of OCR models on handwritten exam images.

Model	Exact-Match Accuracy	Character-Level Accuracy	CER
PaddleOCR	0.6750	0.7343	0.2657
TrOCR (Base)	0.3276	0.4780	0.5220
TrOCR (Large)	0.4854	0.6503	0.3497
EasyOCR	0.1271	0.2576	0.7424

4.3 Discussion of Findings

The quantitative results indicate notable performance differences among the evaluated OCR models under unconstrained capture conditions. PaddleOCR achieves the highest exact-match and character-level accuracy, which can be attributed to its line-level detection and recognition pipeline that aligns well with the structured nature of exam responses. TrOCR-large consistently outperforms TrOCR-base, suggesting that larger transformer-based models benefit recognition quality when sufficient visual context is available, although they remain sensitive to noise and segmentation artifacts. EasyOCR exhibits the lowest performance across all metrics, indicating limited robustness to cursive handwriting and challenging illumination conditions present in the dataset. Overall exact-match accuracy values remain moderate, reflecting the difficulty of handwritten exam OCR when errors propagate from segmentation

and recognition stages. These findings highlight that recognition performance is strongly influenced by preprocessing quality and handwriting variability rather than model architecture alone. As no fine-tuning or parameter optimization was performed, the reported results should be interpreted as preliminary and indicative of baseline performance within the proposed pipeline.

5 Conclusion

This project investigated the feasibility of applying optical character recognition techniques to handwritten exam papers captured under real-world conditions. Manual grading of handwritten exams remains time-consuming and difficult to scale, motivating the need for automated or semi-automated solutions that can assist instructors without changing existing assessment practices. The primary objective of this work was to design and evaluate an end-to-end OCR-based pipeline capable of extracting and structuring handwritten exam responses from smartphone-captured images.

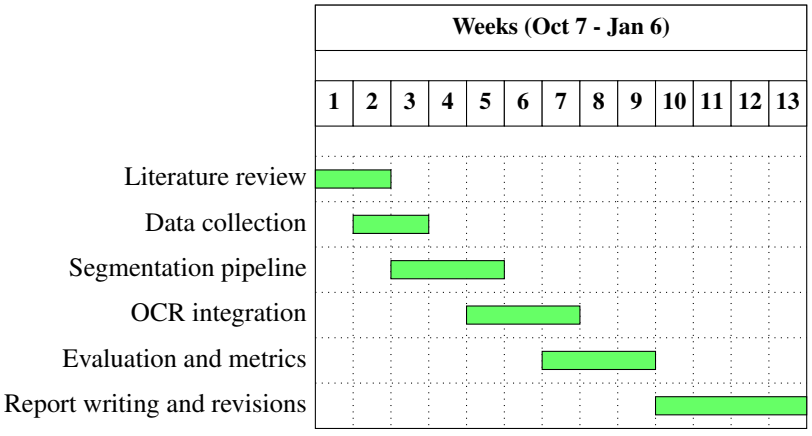
An evaluation-focused methodology was adopted, incorporating image preprocessing, page segmentation, line-level text detection, handwritten text recognition, answer structuring, and quantitative assessment. Multiple open-source OCR models were integrated into the pipeline, including PaddleOCR, EasyOCR, and TrOCR, all used in their pre-trained configurations without any fine-tuning. Experiments were conducted on a dataset consisting of handwritten exam images collected from multiple individuals under unconstrained lighting and capture conditions, reflecting realistic educational scenarios.

The preliminary experimental results highlight notable differences in performance across the evaluated models. PaddleOCR achieved the strongest baseline performance within the proposed pipeline, while transformer-based models demonstrated potential but remained sensitive to noise and segmentation artifacts. Overall recognition accuracy remained moderate, emphasizing the inherent difficulty of handwritten exam OCR when operating on noisy inputs and without domain adaptation.

Future work may focus on domain-specific fine-tuning of recognition models, and more advanced answer normalization techniques. Expanding the dataset to include a broader range of handwriting styles and exam formats may further support more comprehensive evaluation. Overall, this work demonstrates the practical challenges and potential of applying OCR technologies to handwritten exam assessment and provides a foundation for future improvements in automated grading systems.

6 Weekly Schedule/Project Plan

Weekly schedule from Oct 7 (Week 1) to Jan 6 (Week 13). Green = done, red = not done.



References

Ghadekar, P., O. Khanvilkar, J. Kharat, K. Joshi, T. Kulkarni, and Y. Kulkarni (2025). Automating the grading of handwritten examinations through the integration of optical character recognition and machine learning algorithms. *Journal of Integrated Science and Technology 13*(6), 1128. Article published 11-Apr-2025.

Lekshmy, P. L., S. Kayalvili, S. Velmurugan, B. Teja Sree, I. Kumari, and P. Karthik Kumar (2024). Optical character recognition (ocr) in handwritten characters using convolutional neural networks to assist in exam reader system. In *2024 2nd International Conference on Advancement in Computation & Computer Technologies (InCACCT)*, pp. 623–627. IEEE.

Li, M., T. Lv, J. Chen, L. Cui, Y. Lu, D. Florencio, C. Zhang, Z. Li, and F. Wei (2022). Trocr: Transformer-based optical character recognition with pre-trained models. *arXiv preprint arXiv:2109.10282*.

Liu, T., J. Chatain, L. Kobel-Keller, G. Kortemeyer, T. Willwacher, and M. Sachan (2024). Ai-assisted automated short answer grading of handwritten university level mathematics exams.

Mulla, A., A. Z. Karjal, H. Sulthana, M. S. S. Abbas, and M. H. Raza (2025, May). Ai-powered exam evaluation system for educational institutions. *International Journal of Multidisciplinary Research in Science, Engineering, Technology and Management 12*(5), 1587–1589.

Pangestu, J. S., K. Kho, D. Suryani, and L. Andika (2023). Implementation of handwriting recognition and answer evaluation with recurrent neural network. *Procedia Computer Science 227*, 779–784.

Pawar, A. H., V. Chavan, V. Pol, H. Shinde, and S. Thorat (2025). Automating handwritten answer evaluation: A deep learning and ocr integrated approach. *International Research Journal of Education and Technology 8*(4), 1120–1123.

Shi, H., Z. Zhu, C. Zhang, X. Feng, and Y. Wang (2025). Multimodal handwritten exam text recognition based on deep learning. *Applied Sciences 15*(16), 8881.