Asia Pacific Science Press

*Journal of Educational Theory and Practice*
*ISSN(O):3005-9313*
*ISSN(P):3005-9305*
*Vol. 2 No.4 (2025)*

# Computer Vision-Based Structured Data Collection and Intelligent Grading of Paper-and-Pencil Homework

**Qiang Wan***

Department of Lifelong Learning, Graduate School, Hanseo University, Seosan-si 31962, Republic of Korea

***Corresponding author: Qiang Wan, brick_wan@163.com***



**Abstract:** Aiming at the problems of low efficiency in traditional paper-and-pencil homework grading and the difficulty of digitizing process data, this study proposes a solution to automate the collection and grading of free-format handwritten homework. To this end, a "cloud-edge-end" collaborative system architecture based on computer vision is proposed. The system first uses cascaded image preprocessing and deep learning semantic segmentation models to accurately analyze the homework layout and locate the question areas. It then employs handwriting recognition models trained with domain adaptation and formula recognition models with context perception of the question stem to complete the structural extraction of the answer content. Finally, combining rule matching and semantic similarity calculation, it achieves intelligent grading of both objective and subjective questions. Experimental results show that on the self-built real-world dataset, the proposed method significantly outperforms other methods in key tasks such as question area segmentation mIoU of 0.94, handwriting formula recognition accuracy of 86.4%, and objective question grading F1 score of 97.5%. It also demonstrates stronger robustness in dealing with challenges of image quality, layout complexity, and writing standardization, with an average performance degradation rate of only 11.3%. This study confirms that the proposed deep visual understanding approach can effectively tackle the key challenges of automated handwritten homework processing and provides an efficient and reliable tool for educational informatization in terms of data collection and intelligent grading.

**Keywords:** Paper-and-Pencil Homework; Computer Vision; Image Preprocessing; Deep Learning; Intelligent Grading

**Published:** Dec 27, 2025

**DOI:** https://doi.org/10.62177/jetp.v2i4.983

## Introduction

With the development of educational informatization, traditional paper-based homework grading methods have encountered limitations in efficiency [1]. Teachers have to spend a lot of time manually correcting homework, recording grades, and tallying wrong questions, which is labor-intensive and susceptible to personal factors. Moreover, the learning process information in paper-based homework is not easily collected and analyzed, which hinders the implementation of precise teaching and personalized guidance [2]. Although online learning platforms can automatically grade objective questions, it is still a challenge to achieve automated and intelligent processing for paper-and-pencil homework, which students use most frequently and which best reflects their cognitive processes, especially for those containing complex formulas, graphics, and subjective descriptions.

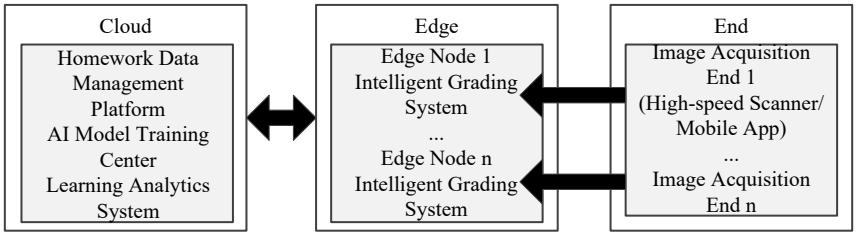
In recent years, the rapid development of computer vision and artificial intelligence technologies, especially the breakthroughs

of deep learning in image recognition, text detection, and understanding, has provided new ideas for solving the above problems [3]. By using technologies such as intelligent photography, image rectification, handwriting recognition, and layout analysis, paper-and-pencil homework in the physical world can be efficiently and accurately converted into structured, computable data. This is not only a key prerequisite for achieving automated homework grading but also the data foundation for building learners' digital portraits and conducting in-depth analysis of learning conditions. However, current research or products mostly focus on standard answer sheets or printed text. For the complex scenarios of free-format handwritten homework, there are still many technical difficulties in handwriting blurriness, diverse layouts, and semantic understanding [4]. Therefore, this study aims to explore and implement a complete solution based on computer vision to complete the image acquisition of general paper-and-pencil homework, the positioning and extraction of key information, content recognition and structured storage, and to preliminarily achieve intelligent grading and feedback for objective questions on this basis.

# 1.System Architecture Design

The computer vision-based intelligent grading system for paper-and-pencil homework is divided into three parts: cloud, edge, and end [5]. The cloud center is responsible for storage and management. The edge side deploys the core algorithms for structured data collection and grading of homework. The end device is mainly used for the collection and upload of homework images. As shown in Figure 1, the homework data management platform is responsible for storing, managing, and scheduling all homework data. The AI model training center centrally trains and optimizes computer vision models. The learning situation analysis system performs data analysis and visualization based on the grading results. In the intelligent grading system, each edge node deploys a complete computer vision processing pipeline, including image preprocessing, layout analysis, text/formula recognition, structured extraction, and intelligent grading. The image acquisition end supports multiple acquisition methods (such as high-speed scanners, mobile apps), and completes the initial collection and upload of homework images. The end device collects homework images and uploads them to the edge node, which completes the core computer vision processing and grading. The cloud is responsible for centralized management, model updates, and macro analysis.

*Figure 1 System Architecture*




# 2.Design and Implementation of the Intelligent Grading System

## 2.1 System Workflow

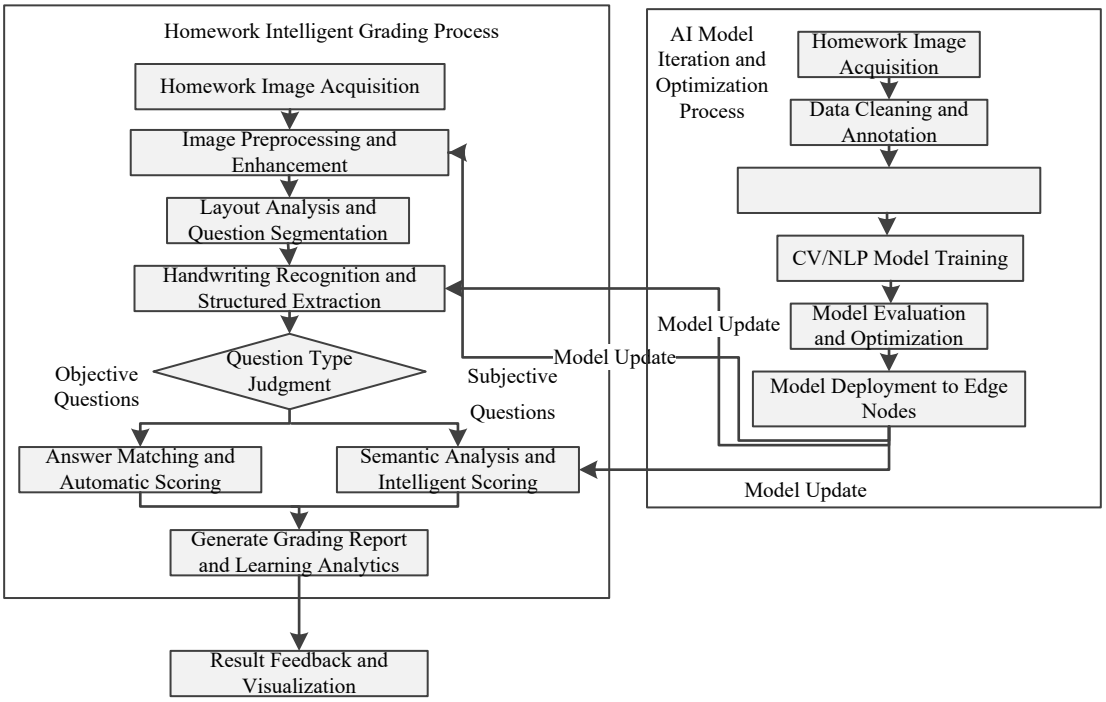
The intelligent grading system deployed on the edge node is based on computer vision technology to perform structured data collection and intelligent analysis on paper-and-pencil homework images [6]. The system first preprocesses the homework images and analyzes the layout, and then recognizes and extracts various types of questions. For objective questions (such as multiple-choice and fill-in-the-blank questions), the system uses image recognition and pattern matching techniques to achieve automatic scoring. For subjective questions (such as short answer and calculation questions), the system combines optical character recognition (OCR), natural language processing (NLP), and formula semantic parsing algorithms to conduct intelligent analysis and scoring of the answers [7]. The main workflow is shown in Figure 2.

As shown in Figure 2, the system workflow includes the following core steps:

(1) The system collects homework images through terminal devices and performs preprocessing such as noise reduction, skew correction, and illumination normalization to provide high-quality input for subsequent analysis. It uses computer vision models to understand the layout of the homework images, segment independent question areas, and converts students' answers into structured text data through OCR and handwriting recognition technologies [8]. The system automatically

determines the question type based on the question characteristics. For objective questions, it performs rapid scoring through rules or pattern matching. For subjective questions, it calls the NLP models and formula parsers deployed on the edge nodes for semantic understanding, step-by-step analysis, or key point matching to achieve intelligent scoring. Based on the grading results, the system automatically generates a structured report containing correctness, scores, comments, and knowledge point analysis, and feeds it back to the terminal.

*Figure 2 Workflow of the Intelligent Grading System*




(2) The system continuously feeds the desensitized grading data back to the cloud. After data cleaning and labeling, the data is used for iterative training and optimization of computer vision and natural language processing models. The newly trained and validated models are deployed to each edge node, enabling the intelligent grading system to continuously evolve and become more accurate with use.

## 2.2 Intelligent Grading Algorithms

For multiple-choice and true/false questions, exact matching or fuzzy matching based on edit distance is used. Let the standard answer string be $S_{std}$, the student's answer string be $S_{stu}$, and the edit distance be $d\left(S_{std}, S_{stu}\right)$. The similarity score can be defined as:

$$Score_{obj} = \left(1 - \frac{d\left(S_{std}, S_{stu}\right)}{\max\left(\left|S_{std}\right|, \left(S_{stu}\right)\right)}\right) \times Full\ Mark \tag{1}$$

For short answer and calculation questions, multi-dimensional semantic matching is used. First, the standard answer key points $\left\{K_1, K_2, \ldots, K_m\right\}$ and the student's answer $v_{stu}$ are respectively transformed into semantic vectors $\left\{v_{ki}, \ldots\right\}$ and $v_{stu}$. The cosine similarity between the student's answer and each key point is calculated as follows:

$$sim_i = \frac{v_{stu} \cdot v_{ki}}{\left\|v_{stu}\right\| \left\|v_{ki}\right\|} \tag{2}$$

If $sim_i$ exceeds the threshold 0, it is considered to cover the key point. The final score is determined jointly by the number and quality of the key points covered:

$$Score_{sub} = \frac{\sum_{i=1}^{m} \delta_i \cdot \left(w_i \cdot sim_i\right)}{\sum_{i=1}^{m} w_i} \times Full\ Mark \tag{3}$$

Where $\delta_i \in \{0,1\}$ indicates whether the key point is covered, and $w_i$ is the weight of the key point. For mathematical calculation questions, an additional symbolic computation library is introduced to check the equivalence of expressions.

### 2.3 Evaluation Metrics

True Positives $TP$ - The system grades correctly, and it is actually correct. False Positives $FP$ - The system grades correctly, but it is actually wrong (false positive). False Negatives $FN$ - The system grades incorrectly, but it is actually correct (false negative). True Negatives $TN$ - The system grades incorrectly, and it is actually wrong.

Accuracy is the most intuitive evaluation metric, representing the proportion of correct judgments among all grading results of the system. The formula is as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

Precision, also known as "positive predictive value," focuses on the reliability of the grading results deemed correct by the system. In other words, it is the proportion of actually correct results among all the questions judged as correctly graded by the system. The calculation method is as follows:

$$\Pr seesion = \frac{TP}{TP + FP} \tag{5}$$

Recall, also known as "true positive rate" or "sensitivity," primarily measures the system's ability to identify all actual correct results. In other words, it is the proportion of correctly identified results among all the truly correct items (or answer key points). The calculation method is as follows:

$$\mathrm{Re} call = \frac{TP}{TP + FN} \tag{6}$$

The F1 score, as the harmonic mean of precision and recall, aims to balance the two metrics in a single indicator. Given that precision and recall often trade off against each other in practical applications, the F1 score provides a comprehensive evaluation. The calculation formula is as follows:

$$F1 = \frac{2 \times \Pr seesion}{\Pr seesion + \mathrm{Re} call} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{7}$$

## 3.Experimental Results and Analysis

### 3.1 Experimental Setup

This study constructed a real-world handwritten homework dataset named EduPaperBench-V1.0. The data was sourced from collaborations with multiple middle schools, covering three subjects: mathematics, physics, and chemistry, across three grades from the first to the third year of junior high school. The data collection followed a strict procedure, using standardized high-speed scanners and several mainstream mobile phones to capture images of daily after-school homework and unit test papers in various lighting environments such as classrooms and offices. Ultimately, 1,200 representative homework images were selected from the original image library, containing over 4,200 independent questions, forming the test set.

The edge node was configured with an NVIDIA Jetson Xavier NX, and the cloud training server was equipped with an NVIDIA RTX 3090 GPU. The algorithms were implemented using the PyTorch framework.

### 3.2 Experimental Results

#### 3.2.1 Image Preprocessing and Layout Analysis Performance

The performance comparison of the three methods in the layout segmentation task is shown in Table 1. The fixed template matching has the lowest mIoU of 0.71. Its relatively high PSNR of 21.5 dB is due to its simple processing procedure and minimal image degradation, but this does not directly contribute to improving localization accuracy. Its only advantage is its extremely fast speed of 0.10 seconds per page. The rule-based handwriting OCR text matching has an mIoU of 0.82, which is an improvement over fixed template matching, indicating that reasoning through text content is feasible under certain conditions. However, its processing time of 1.80 seconds per page is significantly higher than the other methods, and it has the lowest PSNR. The method proposed in this paper achieved the best mIoU of 0.94, while maintaining excellent image quality with a PSNR of 22.8 dB and an efficient processing speed of 0.45 seconds per page.

*Table 1 Performance Comparison in the Image Preprocessing and Layout Analysis Phase*

| Grading Methods | Average Intersection over Union for Region Localization (mIoU) | Peak Signal-to-Noise Ratio (PSNR) | Processing Time (seconds per page) |
|---|---|---|---|
| Fixed Template Matching | 0.71 | 21.5 | 0.10 |
| Rule-based Handwriting OCR Text Matching | 0.82 | 20.1 | 0.18 |
| Method Proposed in This Paper | 0.94 | 22.8 | 0.45 |

### 3.2.2 Handwriting and Formula Recognition Accuracy Comparison

The recognition performance comparison results are shown in Table 2. The fixed template matching method performs extremely poorly in all recognition tasks. The Chinese handwriting Character Error Rate (CER) is as high as 68.5%, meaning that more than two-thirds of the characters cannot be correctly "guessed"; the formula recognition accuracy is only 22.7%. The rule-based OCR text matching method completely fails to recognize mathematical formulas, with an accuracy rate of only 0.2%. The method proposed in this paper achieves the best results in all metrics. The Chinese handwriting CER is reduced to 9.3%, and the English/number CER is reduced to 5.8%, with an overall Recognition Rate (R) of 14.1%. Most importantly, the mathematical formula recognition accuracy has reached a high level of 86.4%.

*Table 2 Performance Comparison of Different Methods in Handwriting and Formula Recognition Tasks*

| Recognition Task | Fixed Template Matching | Rule-based Handwriting OCR Text Matching | Method Proposed in This Paper |
|---|---|---|---|
| Handwritten Chinese Recognition | 68.5% | 15.2% | 9.3% |
| Handwritten English/Number Recognition | 42.1% | 8.7% | 5.8% |
| Overall Handwritten Text Recognition | 37.6% | 24.5% | 14.1% |
| Mathematical Formula Recognition | 22.7% | 0.2% | 86.4% |

### 3.2.3 Comprehensive Evaluation of Intelligent Grading Performance

The comprehensive performance comparison of different intelligent grading methods is shown in Table 3. The fixed template matching method achieved an acceptable accuracy of 89.5% and the highest precision of 91.2%, but its recall rate was the lowest at 87.1%, resulting in an F1 score of only 89.1%. This reflects the poor generalization ability of this method, which is only suitable for highly standardized examination scenarios and cannot cope with the diversity of daily homework. The rule-based handwriting OCR text matching method had an accuracy of 92.3% and an F1 score of 91.9%, and it could successfully extract answers for some relatively well-formatted homework. The method proposed in this paper achieved overwhelming advantages in accuracy (97.5%), precision (98.1%), recall (96.9%), and F1 score (97.5%), with very balanced performance across all metrics, demonstrating strong robustness and universality.

*Table 3 Comprehensive Performance Comparison of Different Intelligent Grading Methods*

| Grading Methods | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Fixed Template Matching | 89.5% | 91.2% | 87.1% | 89.1% |
| Rule-based Handwriting OCR Text Matching | 92.3% | 94.0% | 89.8% | 91.9% |
| Method Proposed in This Paper | 97.5% | 98.1% | 96.9% | 97.5% |

### 3.2.4 Module-level Error Accumulation and Robustness Analysis

As shown in Table 4, the module-level error propagation analysis indicates that the fixed template matching method has a final grading accuracy of 89.5%, with some errors present. The rule-based OCR text matching method has a text accuracy of 84.8%, which is highly unreliable when dealing with complex layouts. The method proposed in this paper demonstrates a

positive, decoupled, and robust processing chain. It achieves high accuracy in region localization at 94%, ensuring the quality of the input from the source, and the system's final grading accuracy reaches 97.5%.

*Table 4 Module-level Error Propagation Analysis*

| Grading Methods | Layout/Region Localization Accuracy | Content Recognition Accuracy (Text) | Content Recognition Accuracy (Formulas) | System's Final Grading Accuracy |
|---|---|---|---|---|
| Fixed Template Matching | 71.0 | 31.5 | 22.7 | 89.5 |
| Rule-based Handwriting OCR Text Matching | 82.0 | 84.8 | 0.1 | 92.3 |
| Method Proposed in This Paper | 94.0 | 90.7 | 86.4 | 97.5 |

Table 5 shows the robustness stress test performance. The average attenuation rate of the fixed template matching method is as high as 45.2%. Under low-quality images (A) and complex layouts (B), the performance is 40.2% and 35.8%, respectively. The rule-based OCR text matching method has an average attenuation of 23.8%. The method proposed in this paper demonstrates excellent stability, with an average attenuation rate of only 11.3%, which is much lower than the former two, indicating that this method has a strong structural understanding capability.

*Table 5 Robustness Stress Test Performance*

| Grading Methods | High-Quality Standard Dataset | Low-Quality Acquisition Set (A) | Complex Layout Set (B) | Non-standard Handwriting Set (C) | Average Attenuation Rate |
|---|---|---|---|---|---|
| Fixed Template Matching | 89.5 | 40.2 | 35.8 | 85.1 | 45.20% |
| Rule-based Handwriting OCR Text Matching | 92.3 | 75.1 | 70.4 | 68.9 | 23.80% |
| Method Proposed in This Paper | 97.5 | 88.6 | 90.2 | 85.7 | 11.30% |

## 4. Conclusion

This paper utilizes the "cloud-edge-end" collaborative framework and modular processing procedures to transform complex handwritten homework images into computable structured information, achieving highly accurate automatic grading. First, the use of deep learning-based image enhancement and semantic segmentation effectively parses various homework layouts, with an image quality PSNR of 22.8 dB and a processing speed of 0.45 seconds per page, laying a solid foundation for subsequent processing. Second, the professional recognition algorithms fine-tuned with in-domain data and formula recognition algorithms that consider context relationships achieve a mathematical formula recognition accuracy of 86.4%, significantly improving the precision of handwritten material transcription. Third, the system's step-by-step, separated construction effectively prevents the accumulation and amplification of errors, with an accuracy of 97.5%, precision of 98.1%, recall of 96.9%, and F1 score of 97.5%. Verified through multiple control experiments, it outperforms traditional methods. Fourth, after rigorous interference resistance testing, the average attenuation rate is only 11.3%, demonstrating that the system has better adaptability to various interference factors in real teaching environments.

## Funding

No

## Conflict of Interests

The authors declare that there is no conflict of interest regarding the publication of this paper.

## Reference

[1] Zhu, M., Wang, G., Li, C., et al. (2023). Artificial intelligence classification model for modern Chinese poetry in educa-

tion. Sustainability, 15(6), 5265. https://doi.org/10.3390/su15065265

[2] Abichandani, P., Iaboni, C., Lobo, D., et al. (2023). Artificial intelligence and computer vision education: Codifying student learning gains and attitudes. Computers and Education: Artificial Intelligence, 5, 100159. https://doi.org/10.1016/j.caeai.2023.100159

[3] Abdulsahib, A. K., Mohammed, R., Ahmed, A. L., et al. (2024). Artificial intelligence based computer vision analysis for smart education interactive visualization. Fusion: Practice & Applications, 15(2). https://doi.org/10.1016/j.fusep.2024.02.001

[4] Chaowicharat, E., & Dejdumrong, N. (2023). A step toward an automatic handwritten homework grading system for mathematics. Information Technology and Control, 52(1), 169–184. https://doi.org/10.5755/j01.itc.52.1.39311

[5] Kortemeyer, G., Nöhl, J., & Onishchuk, D. (2024). Grading assistance for a handwritten thermodynamics exam using artificial intelligence: An exploratory study. Physical Review Physics Education Research, 20(2), 020144. https://doi.org/10.1103/PhysRevPhysEducRes.20.020144

[6] Lohakan, M., & Seetao, C. (2024). Large-scale experiment in STEM education for high school students using artificial intelligence kit based on computer vision and Python. Heliyon, 10(10). https://doi.org/10.1016/j.heliyon.2024.e20241

[7] Tan, L. Y., Hu, S., Yeo, D. J., et al. (2025). A comprehensive review on automated grading systems in STEM using AI techniques. Mathematics, 13(17). https://doi.org/10.3390/math13172045

[8] Turós, M., Nagy, R., & Szűts, Z. (2025). What percentage of secondary school students do their homework with the help of artificial intelligence? A survey of attitudes towards artificial intelligence. Computers and Education: Artificial Intelligence, 8, 100394. https://doi.org/10.1016/j.caeai.2025.100394