International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 10 | Oct - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Handwritten Answer Sheet Digitization and Similarity-Based Copying Pattern Detection Using OCR and NLP

[1.]Prof. Prasad Mahamuni, [2.]Sandesh Patil,[3.]Rajvardhan Mohite, [4.]Tejas Borse, [5.]Dnyaneshwar Patil
Asst Professor, Student1, Student2, Student3, Student4
mahamuni.prasad@gmail.com, rajputsandesh726@gmail.com, mohiterajvardhan76@gmail.com, tejasborse4945@gmail.com, patildnyaneshwar416@gmail.com
MCA Department, ASM's Institute of Business Management and Research Chinchwad, Pune-19, Maharashtra, India

Abstract— Evaluation of handwritten examination answer sheets is a time-consuming task for teachers, particularly when a large number of students are involved. Manual evaluation requires teachers to read handwritten responses, compare them with reference answers, assign marks, and identify responses that may contain similar or potentially copied content. These activities can increase the workload of evaluators and make large-scale examination analysis difficult.

This paper presents a proposed intelligent examination-support system for digitizing handwritten answer sheets and assisting teachers in answer evaluation and copying-pattern detection. The system uses Optical Character Recognition (OCR) and Handwritten Text Recognition (HTR) techniques to convert scanned or image-based handwritten responses into machine-readable text. The extracted content is then processed using Natural Language Processing (NLP) and semantic similarity techniques. Student answers are compared with reference answers to generate answer-wise evaluation support and suggested marks. In addition, answers submitted by different students are compared to identify highly similar responses that may require further investigation.

Keywords— Handwritten Text Recognition, Optical Character Recognition, Natural Language Processing, Semantic Similarity, Automatic Answer Evaluation, Examination Support System, Copying Pattern Detection, Machine Learning

I. INTRODUCTION

A. Background

Examination evaluation is an important part of the academic process, but evaluating handwritten answer sheets manually can require considerable time and effort. The challenge becomes more noticeable when teachers need to evaluate a large number of answer sheets within a limited period. In addition to assigning marks, evaluators may need to determine whether an answer is correct, partially correct, incomplete, or unrelated to the question.

B. Scope and Objective

The proposed research focuses on developing an intelligent examination-support system for processing and analysing handwritten answer sheets. The system is designed to work with scanned or image-based answer sheets submitted by students. Teachers or authorized evaluators can upload the question paper, reference answers, and student answer sheets through the web application.

The main objectives of this review are:

- To digitize handwritten answer sheets using OCR and Handwritten Text Recognition (HTR) techniques.
- To convert handwritten responses into machine-readable text for further computational analysis.
- To preprocess and organize extracted text so that individual answers can be associated with their corresponding questions.
- To apply NLP techniques for cleaning, normalization, tokenization, and semantic processing of student responses.
- To compare student answers with reference answers using semantic similarity rather than relying only on exact keyword matching.
- To provide answer-evaluation support by identifying potentially correct, incorrect, partially correct, or incomplete responses.

C. Review Questions

RQ1: How can OCR and Handwritten Text Recognition techniques be used to digitize handwritten examination answer sheets?

RQ2: How can the extracted handwritten text be processed using NLP techniques for automated answer analysis?

RQ3: How can semantic similarity techniques be used to compare student answers with corresponding reference answers?

II. REVIEW METHODOLOGY

A. Literature Basis

The literature basis of this research includes previous studies related to handwritten text recognition, OCR, NLP, semantic similarity, and automated answer evaluation. These studies provide the technical foundation for converting handwritten responses into machine-readable text and analysing their meaning. Research on semantic similarity is considered for

comparing student answers with reference answers and with responses submitted by other students.

B. Thematic Classification

The two studies are grouped into two main themes.

Theme 1 — Handwritten Text Recognition and Digitization

This theme covers research dealing with the recognition and conversion of handwritten documents into digital text. It includes OCR, HTR, image preprocessing, handwriting segmentation, and recognition techniques.

The theme is directly related to the first stage of the proposed system, where scanned or image-based answer sheets are processed and converted into machine-readable text. The SRS identifies image quality, handwriting clarity, orientation, noise, and resolution as factors that can influence recognition performance.

Theme 2 — NLP-Based Answer Processing

This theme focuses on the application of Natural Language Processing to educational responses. It includes text preprocessing, normalization, tokenization, text representation, and extraction of meaningful information from student answers.

TABLE I
THEMATIC DISTRIBUTION OF THE REVIEWED LITERATURE

Ref.	Study	Primary Theme	Main Contribution
[1]	Text-to-Text Semantic Similarity for Automatic Short Answer Grading — Mohler and Mihalcea	Semantic Similarity & Automatic Answer Evaluation	Explores semantic similarity between student answers and reference answers for automatic short-answer grading.
[2]	The Eras and Trends of Automatic Short Answer Grading — Burrows, Gurevych and Stein	Automatic Short Answer Grading	Reviews major approaches and developments in automatic short-answer grading, including NLP, machine learning, and semantic-based methods.

III. SERIOUS GAMES FOR MORAL AND ETHICAL LEARNING

A. Educational Approach

The proposed system follows an AI-assisted educational approach for evaluating handwritten examination answers.

It uses OCR/HTR to digitize handwritten responses and NLP techniques to analyse the extracted content. Student answers are compared with reference answers using semantic similarity to provide evaluation support and suggested marks. The system keeps the teacher involved in reviewing and finalizing the evaluation, particularly for subjective or unclear answers.

B. Learning Objective

The primary objective is to support teachers in evaluating handwritten answers by combining OCR/HTR, NLP, and semantic similarity techniques. The system aims to compare student responses with reference answers and generate answer classifications and suggested marks.

C. Learning Outcomes

The proposed system converts handwritten responses into digitized text that can be processed and analysed electronically using NLP and semantic similarity techniques. It provides answer-wise evaluation information, including suggested marks and answer classifications, while also generating similarity scores for comparing responses submitted by different students.

D. Significance

The proposed system is significant because it integrates handwritten answer-sheet digitization, NLP-based analysis, semantic evaluation, and similarity detection within a single examination-support workflow. By automating repetitive processing tasks, the system can assist teachers in handling large numbers of handwritten responses more efficiently and provide additional information for reviewing answer quality and potentially similar responses. However, the system is intended as an evaluation-support tool rather than a replacement for human judgement.

IV. DIGITAL GAME-BASED LEARNING FOR DIGITAL CITIZENSHIP

A. Educational Approach

The proposed system uses an AI-assisted approach to analyse digitized student answers using Natural Language Processing and semantic similarity techniques. The extracted responses are compared with reference answers to identify relevant concepts and provide evaluation support.

B. Digital Etiquette Learning

The proposed system applies Natural Language Processing techniques to the text extracted from handwritten answer sheets. The responses are cleaned, normalized, and processed to identify relevant information and concepts before being compared with reference answers. This approach allows the system to analyse student responses based on their textual content rather than relying only on exact word matching.

C. Learning Motivation

The proposed system is designed to reduce the repetitive workload involved in handwritten answer evaluation by providing automated analysis support. By generating extracted answers, suggested marks, answer classifications, and similarity scores, the system can help

teachers review examination responses more efficiently. The automated results are intended to support the evaluation process while allowing teachers to verify and modify the results when necessary

D. Student Engagement

The proposed system can improve the overall examination evaluation process by providing structured digital feedback about student answers. Extracted responses, suggested marks, answer classifications, and similarity scores give teachers a clearer view of student performance. This information can help teachers identify answers that require additional attention and support more consistent review of examination responses.

V. COMPARATIVE SYNTHESIS

A. Complementary Roles

The proposed system combines multiple technologies that address different stages of handwritten answer evaluation. OCR/HTR is used for handwritten text digitization, while NLP and semantic similarity support answer evaluation and comparison. Similarity analysis further helps identify highly similar responses between students, while teacher verification provides human judgement for final evaluation.

B. Comparison of Educational Focus

Table II. Comparison Of The Two Game-Based Learning Approaches

Aspect	Handwritten Answer Evaluation	Similarity-Based Copying Detection
Main purpose	Support evaluation of student answers	Identify highly similar student responses
Input	Student answer and reference answer	Answers submitted by different students
Primary technology	OCR/HTR, NLP and semantic similarity	NLP and semantic similarity
Main output	Answer classification and suggested marks	Similarity score and flagged responses
Human involvement	Teacher reviews and modifies results	Teacher verifies flagged similarity cases

C. Comparison of Outcomes

The answer-evaluation component focuses on determining how closely a student's response relates to the corresponding reference answer and provides suggested evaluation information. The similarity-detection component focuses on comparing responses between students and identifying highly similar content. Therefore, the two components have different objectives but operate on the same digitized and processed answer data.

D. Overall Relationship Between the Studies

The components work together as a continuous examination-support workflow. OCR/HTR converts hadwritten answers into digital text, NLP prepares and analyses the extracted responses, semantic similarity supports answer comparison, and the evaluation and similarity modules generate useful results for the teacher.

VI. RESEARCH GAPS AND PROPOSED SOLUTIONS

A. Limited Participant Diversity

Existing studies on automated answer evaluation and handwritten text processing may be conducted using specific datasets, educational settings, or limited groups of learners. This can make it difficult to determine whether the proposed approaches perform consistently across different handwriting styles, subjects, educational levels, and answer formats.

B. Need for Long-Term Evaluation

Most evaluation of automated examination-support systems may focus on immediate performance during a particular testing period. Such evaluation may not fully show how the system performs over repeated examinations, different answer-sheet conditions, and changing datasets.

C. Dependence on Evaluation Methods

The effectiveness of the proposed system depends on the methods used to evaluate OCR/HTR accuracy, answer evaluation, and similarity detection. Different evaluation tasks require different measures, and a single metric may not adequately represent the overall system performance.

D. Game Design and Learning Effectiveness

The effectiveness of the proposed system depends on the quality of its OCR/HTR, NLP processing, semantic similarity, and evaluation mechanisms. Poor handwriting, low-quality images, incorrect orientation, and noise can reduce text-recognition accuracy and subsequently affect answer evaluation and similarity analysis

E. Expansion Across Educational Domains

The proposed system is primarily designed for handwritten examination answer evaluation, but its underlying OCR/HTR, NLP, and semantic similarity techniques can be extended to different educational contexts. The same approach could be adapted for assignments, practice tests, descriptive assessments, and other handwritten academic responses.

VII. CONCLUSION

The proposed system combines OCR/HTR, NLP, and semantic similarity techniques to support the digitization and evaluation of handwritten examination answer sheets. It converts handwritten responses into machine-readable text and analyses them against reference answers to provide evaluation support and suggested

marks. The system also compares responses submitted by different students to identify highly similar answers that may require further examination. The automated results are intended to assist teachers rather than replace their judgement, particularly in cases involving subjective answers, unclear handwriting, or recognition errors. Overall, the proposed approach provides an integrated framework for handwritten answer analysis and similarity-based copying-pattern detection, while future evaluation across diverse datasets and educational contexts can help assess its practical effectiveness.

ACKNOWLEDGMENT

The authors sincerely acknowledge the guidance and support provided by Prof. Prasad Mahamuni and the MCA Department, ASM'S Institute of Business Management and Research, Chinchwad, Pune, throughout this research work. The authors are also grateful to the institution and all individuals who provided valuable suggestions and support during the development and documentation of the proposed system. Finally, the authors acknowledge the researchers whose previous work in OCR, HTR, NLP, semantic similarity, and automated answer evaluation contributed to the foundation of this research.

REFERENCES

[1] M. Mohler and R. Mihalcea, "Text-to-Text Semantic Similarity for Automatic Short Answer Grading," in Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009), Athens, Greece, pp. 567–575, 2009.

[2] S. Burrows, I. Gurevych, and B. Stein, "The Eras and Trends of Automatic Short Answer Grading," International Journal of Artificial Intelligence in Education, vol. 25, no. 1, pp. 60–117, 2015, doi: 10.1007/s40593-014-0026-8.

[3] M. Mohler, R. Bunescu, and R. Mihalcea, "Learning to Grade Short Answer Questions Using Semantic Similarity Measures and Dependency Graph Alignments," in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA, pp. 752–762, 2011.

[4] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, pp. 3982–3992, 2019, doi: 10.18653/v1/D19-1410.

[5] J. Memon, M. Sami, and R. A. Khan, "Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review," IEEE Access, vol. 8, pp. 142642–142668, 2020, doi: 10.1109/ACCESS.2020.3012542.

[6] "Handwritten Text Recognition Using Machine Learning," in Proceedings of the 2023 International Conference on Sustainable Emerging Innovations in Engineering and Technology (ICSEIET), Ghaziabad, India, 2023, doi: 10.1109/ICSEIET58677.2023.10303304.

[7] "Handwritten Recognition: A Survey," in Proceedings of the 2020 IEEE 4th International Conference on Image Processing, Applications and Systems (IPAS), Genova, Italy, 2020, doi: 10.1109/IPAS50080.2020.9334936.

[8] Z. Li, Y. Tomar, and R. J. Passonneau, "A Semantic Feature-Wise Transformation Relation Network for Automatic Short Answer Grading," in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021), pp. 6677–6690, 2021.

[9] J. Ferrero, L. Besacier, D. Schwab, and F. Agnès, "CompiLIG at SemEval-2017 Task 1: Cross-Language Plagiarism Detection Methods for Semantic Textual Similarity," in Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, Canada, pp. 109–114, 2017, doi: 10.18653/v1/S17-2012.

[10] M. Bexte, A. Horbach, and T. Zesch, "Similarity-Based Content Scoring—A More Classroom-Suitable Alternative to Instance-Based Scoring?" in Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), pp. 1892–1903, 2023.