International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 06 | June - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Handwritten Text Recognition and Plagiarism Detection Using Machine Learning

Gayatri M. Rane[1], Shruti D. Anabhavane[2], Esha R. Mandavkar[3]

[1,2,3]*Post-Graduate Student, Master of Computer Application,* Finolex Academy of Management and Technology, Ratnagiri, Maharashtra, India.

Abstract - This review paper evaluates a recent study on a dual-purpose system for handwritten text recognition (HTR) and plagiarism detection, leveraging the Connectionist Temporal Classification (CTC) algorithm and cosine similarity. The reviewed work focuses on digitizing handwritten documents and ensuring content originality, with applications in academic and professional settings. This paper synthesizes the study's contributions, methodologies, and results, while situating it within the broader literature on HTR and plagiarism detection. Key insights include the system's high accuracy in recognizing complex handwriting and its effective use of cosine similarity for plagiarism detection. However, limitations such as challenges with cursive scripts and paraphrased content detection are identified. The review highlights research gaps and proposes future directions, including the integration of advanced deep learning models and multilingual datasets to enhance system robustness.

Key Words: Handwritten Text Recognition, Plagiarism Detection, Connectionist Temporal Classification, Cosine Similarity, Convolutional Neural Networks, Recurrent Neural Networks

1.INTRODUCTION

Handwritten Text Recognition (HTR) and plagiarism detection are critical technologies in the digital era, enabling the conversion of handwritten documents into machine-readable formats and ensuring content integrity. The reviewed paper proposes a system that combines HTR with plagiarism detection, utilizing the Connectionist Temporal Classification (CTC) algorithm and cosine similarity. This dual functionality addresses the growing need for automated document processing and academic integrity in sectors such as education, healthcare, and banking. The importance of this topic lies in its ability to streamline workflows, reduce manual effort, and safeguard intellectual property. This review aims to critically analyse the proposed system, evaluate its contributions against existing literature, and identify areas for improvement. The scope includes studies on HTR and plagiarism detection from 2020 to 2022, focusing on machine learning and deep learning approaches. The paper is structured as follows: a review methodology, thematic analysis of HTR and plagiarism detection techniques, discussion of research gaps, future directions, and a conclusion.

2. REVIEW METHODOLOGY

This review focuses on literature from 2020 to 2022, sourced from databases such as IEEE Xplore, Elsevier, and Springer. Keywords included "handwritten text recognition," "plagiarism detection," "CTC algorithm," "cosine similarity," "CNN," and "RNN." Inclusion criteria comprised peer-reviewed articles proposing machine learning or deep learning-based HTR and plagiarism detection systems. Exclusion criteria eliminated non-peer-reviewed sources and studies not addressing both HTR and plagiarism detection. The framework organizes studies by methodology (e.g., CTC-based, CNN-based) and application (e.g., academic, forensic). A total of 11 studies from the reviewed paper's bibliography were analyze, supplemented by additional relevant works.

2.1.Thematic Review Sections

2.1.1.Handwritten Text Recognition Techniques

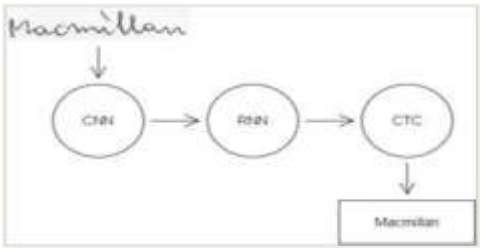

Figure 1: CNN-RNN-CTC Architecture for Handwritten Text Recognition [1]

Figure 1 illustrates the architecture used for Handwritten Text Recognition (HTR) in the reviewed system. The process begins with a handwritten input image, which is processed through a Convolutional Neural Network (CNN) to extract spatial features. These features are then passed to a Recurrent Neural Network (RNN) to capture sequential dependencies, and finally, the Connectionist Temporal Classification (CTC) algorithm decodes the sequence into readable text. This pipeline eliminates the need for pre-segmented data, enhancing recognition accuracy [1].Handwritten Text Recognition (HTR) involves converting handwritten images into digital text, a task complicated by diverse writing styles and scripts. The reviewed paper employs the CTC algorithm, which integrates Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to achieve high accuracy (85.43%) in recognizing complex handwriting [1]. This aligns with prior work by Guptha et al. [2], who used Long Short-Term Memory (LSTM) networks optimized with the Elephant Herding Optimization algorithm for cross-lingual character recognition, achieving robust performance across scripts. Similarly, Wang [3] proposed a style extractor network for fast writer adaptation, improving HTR accuracy for new handwriting styles. However, the reviewed system's use of CTC eliminates the need for pre-aligned datasets, a significant advantage over traditional methods like KNN and SVM, which achieved lower accuracies (71.15% and 80.39%, respectively) [1]. Studies by

Hemanth and Jayasree [4] and Mahapatra [5] also highlight CNN-RNN combinations and hybrid KNN-SVM models, but these often require extensive preprocessing, unlike the CTC-based approach.

A critical insight is that CTC's ability to handle unaligned sequences makes it more efficient for real-world applications, such as digitizing historical documents or processing forms in banking [6]. However, the reviewed system struggles with highly cursive or poorly written scripts, a limitation also noted in Shivakumara et al. [7], who addressed text line segmentation in struck-out documents. This suggests that while CTC-based models are promising, their performance depends on dataset diversity and preprocessing quality.

2.1.2.Plagiarism Detection Methods

$$similarity(A,B) = cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sum_{i=1}^{n} B_i^2}$$

Figure 2: Cosine Similarity Formula [1]

The cosine similarity formula is defined as:
similarity (A, B) = cos(θ) = (A · B) / (||A|| ||B||) = (Σ A_iB_i) / sqrt (Σ A_i^2 * Σ B_i^2),

where A and B are vectors representing text features. This approach evaluates the angle between two vectors, measuring their orientation rather than magnitude, making it effective for detecting textual similarity in plagiarism detection systems [1].

Plagiarism detection in digitized handwritten documents is a growing area of research, particularly in academic settings. The reviewed paper uses cosine similarity to measure textual overlap, achieving reliable detection of direct copying [1]. This method is computationally efficient and scalable, as it evaluates document similarity regardless of size, aligning with findings by Bhole and Bhagwat [8], who compared plagiarism detection models for online studies. Chinedu and Ikerionwu [9] further emphasize cosine similarity's effectiveness in detecting source code plagiarism, noting its low false-positive rate. However, the reviewed system's reliance on cosine similarity limits its ability to detect paraphrased or semantically altered content, a gap also identified in [9].

Alternative approaches include Java-based string manipulation for low false-positive rates [10] and deep learning-based methods like Autoencoders for feature extraction [4]. These methods offer improved detection of complex plagiarism but require significant computational resources. The reviewed system's strength lies in its integration of HTR and plagiarism detection, enabling seamless processing of handwritten documents, but its reliance on cosine similarity alone restricts its applicability to sophisticated plagiarism cases.

2.1.3.Comparative Analysis of Algorithms

The reviewed paper compares CTC, CNN, SVM, and KNN algorithms, with CTC outperforming others in accuracy (85.43%) and precision (94.80%) [1]. This is consistent with Sreeraj et al. [1], who reported improved performance of machine learning models in related tasks like cervical spondylosis detection, suggesting CTC's versatility in sequence prediction. However, earlier studies like Memon and Uddin [6] highlight SVM's robustness in OCR tasks, though with lower accuracy (80.39%) compared to CTC. KNN's lower performance (71.15%) reflects its sensitivity to dataset size and quality [5]. The reviewed system's use of synthetic data to augment training datasets [2] enhances its performance, but its reliance on standardized image sizes (128x32 pixels) may limit scalability for larger documents.

3.DISCUSION AND RESEARCH GAPS

The reviewed system demonstrates significant advancements in HTR and plagiarism detection, particularly through its use of CTC and cosine similarity. Its dual functionality addresses practical needs in academic and professional settings, reducing manual labour and enhancing content integrity. The CTC algorithm's ability to handle unaligned sequences is a key strength, supported by literature emphasizing CNN-RNN models [3][4]. However, the system's limitations include reduced accuracy for cursive or poorly written scripts and the inability of cosine similarity to detect paraphrased content. These align with broader challenges in HTR, where dataset diversity and script complexity remain barriers [7]. In plagiarism detection, the literature highlights a need for semantic analysis to address advanced plagiarism techniques [9]. Additionally, Rakholia [11] notes that deep learning approaches, such as those using convolutional and recurrent architectures, can further improve HTR accuracy by leveraging larger and more diverse datasets, suggesting a potential enhancement for the reviewed system's robustness.

A notable gap is the system's lack of support for multilingual scripts, limiting its applicability in linguistically diverse regions. Additionally, the reliance on offline processing, while beneficial for low-connectivity areas, restricts real-time applications. Conflicts in the literature arise regarding optimal algorithms, with some studies favouring hybrid KNN-SVM models [5] over CTC for specific datasets, suggesting context-dependent performance. The reviewed system's integration of HTR and plagiarism detection is innovative, but its scalability and robustness require further validation across larger, diverse datasets. Incorporating deep learning advancements, as discussed in Rakholia [11], could address these scalability issues by optimizing feature extraction for varied handwriting styles.

4.FUTURE RESEARCH DIRECTIONS

Future research should focus on:

1. **Multilingual Support:** Expanding datasets to include regional scripts like Hindi and Marathi, as suggested in the reviewed paper, to enhance global applicability [2].
2. **Dataset Augmentation:** Leveraging generative adversarial networks (GANs) to create diverse synthetic datasets, improving model generalization [2].
3. **Improved HTR Robustness:** Incorporating transfer learning and transformer architectures to handle cursive and low-quality handwriting, building on insights from [3]. Additionally, adopting deep learning techniques from Rakholia [11], such as enhanced convolutional architectures, could improve recognition accuracy across diverse scripts.
4. **Real-Time Processing:** Developing online HTR systems to support dynamic applications, such as real-time form processing in banking.
5. **Advanced Plagiarism Detection:** Integrating Siamese LSTM or transformer-based models to detect semantic plagiarism, addressing limitations of cosine similarity [9].

5.CONCLUSIONS

This review highlights the strengths and limitations of a novel system for handwritten text recognition and plagiarism detection using CTC and cosine similarity. The system achieves high accuracy (85.43%) in recognizing complex handwriting and effectively detects direct plagiarism, offering practical value in academic and professional contexts. However, challenges with cursive scripts and paraphrased content detection underscore the need for further development. By addressing these gaps through multilingual datasets, advanced algorithms, and real-time capabilities, the system can evolve into a robust tool for diverse applications. This review underscores the significance of integrating HTR and plagiarism detection, paving the way for future innovations in automated document processing and content integrity.

6.REFERENCES

[1] M. Sreeraj, J. Joy, M. Jose, M. Varghese, T. J. Rejoice, "Comparative analysis of Machine Learning approaches for early-stage Cervical Spondylosis detection," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 6, pp. 3301–3309, 2022.

[2] N. Guptha, G. Megharaj, "Cross lingual handwritten character recognition using long short term memory network with aid of elephant herding optimization algorithm," *Pattern Recognition Letters*, vol. 159, May 2022.

[3] Z.-R. Wang, "Fast writer adaptation with style extractor network for handwritten text recognition," *Neural Networks*, vol. 147, December 2021.

[4] G. R. Hemanth, M. Jayasree, "CNN-RNN based handwritten text recognition," *ICTACT*, vol. 12, pp. 2457–2463, October 2021.

[5] K. K. Mahapatra, "Handwritten character recognition using KNN and SVM based classifier over feature vector from autoencoder," *CCIS*, vol. 1240, pp. 1415–1419, 2020.

[6] J. Memon, M. Uddina, "Handwritten optical character recognition (OCR)," *IEEE Access*, vol. 8, pp. 142642–142668, 2020.

[7] P. Shivakumara, T. Jain, T. Lu, "Text line segmentation from struck-out handwritten document images," *Expert Systems with Applications*, vol. 210, July 2022.

[8] B. P. Bhole, S. Bhagwat, "Study and comparison of plagiarism detection model for online studies," *ICAAIC*, vol. 10, pp. 1154–1158, 2022.

[9] C. O. Chinedu, U. Ikerionwu, "Plagiarism detection systems," *IJSRP*, vol. 4, pp. 72–77, 2020.

[10] S. K. Bera, S. Kunda, "Distance transform based textline extraction from unconstrained handwritten document images," *Expert Systems with Applications*, vol. 186, pp. 72–77, July 2021.

[11] B. Rakholia, "Handwritten character recognition using deep learning," *International Journal of Recent Technology and Engineering*, vol. 8, pp. 5815–5819, 2020.