IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 8, March 2026
Impact Factor: 8.2

Text and Image Plagiarism Detection Using Histogram Matching and LCS

M. Abhiram Reddy[1], G. Lakshana Sai[2], T. Manoj Kumar[3], Ms. V. Neeraja[4]
UG Scholar, Department of Computer Science & Engineering, CMR Technical Campus, Hyderabad, India[1]
UG Scholar, Department of Computer Science & Engineering, CMR Technical Campus, Hyderabad, India[2]
UG Scholar, Department of Computer Science & Engineering, CMR Technical Campus, Hyderabad, India[3]
Assistant Professor, Department of Computer Science & Engineering, CMR Technical Campus, Hyderabad, India[4]

Abstract: *Academic plagiarism poses a serious threat to the integrity of scholarly work, with forms ranging from direct copy-and-paste to sophisticated disguised plagiarism involving images and figures. Existing plagiarism detection tools predominantly focus on text analysis and largely ignore the role of images, which carry significant information in scientific publications. This paper presents a system for detecting both textual and image-based plagiarism in academic documents. For text plagiarism detection, the Longest Common Subsequence (LCS) algorithm is applied to measure similarity between a suspicious document and a corpus of source files. For image plagiarism detection, a Five Module Method (FMM) based on histogram comparison is employed, enabling detection of copied images including flowcharts, graphs, and photographs. The proposed system is implemented using Python and Django and tested on a publicly available flowchart image database used in CLEF-IP 2012 competitions. The system achieved an average flowchart recognition accuracy of 81.91%, demonstrating strong performance over existing methods such as CVC and INRIA. The combined approach provides a robust and scalable solution for detecting diverse forms of academic plagiarism.*

Keywords: Plagiarism Detection, Text Similarity, Image Plagiarism, Longest Common Subsequence (LCS), Histogram Matching, Flowchart Recognition, Optical Character Recognition (OCR), Artificial Neural Networks, Python, Django

I. INTRODUCTION

Academic plagiarism has been defined as the use of ideas, concepts, words, or structures without appropriately acknowledging the source to benefit in a setting where originality is expected. Forms of academic plagiarism vary in their degree of obfuscation, ranging from unaltered copies (copy-and-paste) to slightly altered forms such as interweaving text passages from multiple sources (shake-and-paste), to more disguised forms including paraphrases, translations, and idea plagiarism, and even the plagiarism of academic data.

Research on plagiarism detection has yielded mature systems employing text retrieval to find similar documents. These systems reliably retrieve documents containing copied text, but often fail to identify disguised forms of academic plagiarism. Compared to the many sophisticated text-based retrieval approaches proposed for plagiarism detection, analyzing images to detect academic plagiarism has attracted little research attention.

In this paper, we examine both text and image similarity detection techniques as promising methods for plagiarism detection. Images, defined here as visual representations of data such as bar charts, scatter plots, flow charts, organigrams, and component diagrams, enable conveying large amounts of information in a compressed format, and they represent information differently from text. These characteristics make images a promising feature to examine when assessing the semantic similarity present in academic documents.

Our main contributions are as follows: (1) development of a text plagiarism detection module using the Longest Common Subsequence (LCS) algorithm, (2) development of an image plagiarism detection module using histogram

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 8, March 2026
Impact Factor: 8.2

matching via the Five Module Method (FMM), and (3) integration of both modules into a unified, web-based plagiarism detection system implemented in Python and Django.

II. LITERATURE SURVEY

Plagiarism detection is a specialized information retrieval task with the objective of comparing an input document to a large collection and retrieving all documents exhibiting similarities above a predefined threshold. PD systems typically follow a two-stage process consisting of candidate retrieval and detailed comparison. For candidate retrieval, systems commonly employ efficient text retrieval methods such as n-gram fingerprinting or vector space models. For detailed comparison, systems typically apply exhaustive string matching. However, such approaches are limited to finding near copies of text.

To detect disguised forms of academic plagiarism, researchers have proposed a variety of mono-lingual text analysis approaches employing semantic and syntactic features, as well as cross-lingual IR methods. Alzahrani et al. combined an analysis of text similarity and structural similarity. Gipp and Meuschke showed that combined analysis of citation patterns and text similarity improves identification of concealed academic plagiarism. Pertile et al. confirmed the positive effect of combining citation and text analysis using machine learning.

Few studies have investigated image similarity for plagiarism detection. Hurtik and Hodakova used higher degree F-transform to identify exact copies of photographs. Iwanowski et al. evaluated feature point methods such as SIFT, SURF, and BRISK to retrieve exact and visually altered copies of photographs. Srivastava et al. addressed this task using a combination of SIFT features and perceptual hashing (pHash), which maps perceived image content to a hash value such that visually similar images produce similar hash values.

Krizhevsky et al. demonstrated the power of deep convolutional neural networks in image classification, achieving top-5 error rates of 17.0% on ImageNet. This work laid the groundwork for learning-based image feature extraction in plagiarism detection. Alzahrani et al. further presented a taxonomy distinguishing literal plagiarism from intelligent plagiarism, providing a foundation for understanding the linguistic patterns involved. The proposed system in this paper builds on these findings to address both textual and visual plagiarism detection in a unified platform.

III. PROPOSED METHODOLOGY

A. Existing System

Currently available plagiarism detection tools focus exclusively on text string comparisons. These systems can find direct copies but fail to identify image-based plagiarism. Since images are an inseparable part of information presentation in scientific research — particularly flowcharts, which carry significant structural and semantic content — existing tools leave a critical gap in academic integrity checking.

B. Proposed System

The proposed system addresses both text and image plagiarism in a unified platform. For text detection, the LCS algorithm compares a suspicious document against a corpus after preprocessing with lemmatization, stemming, and stop-word removal. For image detection, the Five Module Method (FMM) computes a histogram representation of each image and compares it against stored database histograms. The system was tested on 44 flowchart images from the CLEF-IP 2012 public database, achieving a recognition accuracy of 81.91%, outperforming both CVC and INRIA baseline methods.

C. System Architecture

The system architecture follows a structured pipeline from user login through file upload, preprocessing, similarity computation, and result presentation. Text files are cleaned and compared using LCS, while images are processed via histogram computation and histogram intersection. Results are displayed in a web-based interface showing the source file, suspicious file, similarity score, and plagiarism verdict.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 8, March 2026
Impact Factor: 8.2

D. System Modules

1. User Login: Users authenticate via signup and login pages before accessing system functions.
2. Upload Source Files: Corpus text files are loaded, preprocessed, and stored in memory as cleaned token sequences.
3. Upload Suspicious Files: A suspicious text file is uploaded, preprocessed, and compared against all corpus files using the LCS algorithm. Similarity percentage is computed as LCS score divided by number of tokens in the suspicious document. A threshold of 0.60 (60%) triggers a plagiarism verdict.
4. Upload Source Images: Images from the database folder are processed via FMM, and their histograms are stored for comparison.
5. Upload Suspicious Image: A test image is uploaded, its histogram computed via FMM, and compared against all database histograms using the HISTCMP_INTERSECT metric. A matching score of 39,000 or above out of 40,000 triggers a plagiarism verdict.

E. Five Module Method (FMM) for Image Comparison

The FMM algorithm processes each image through five sequential operations: (1) resize to 50×50 pixels and convert to grayscale, (2) apply intensity threshold to normalize dark pixels, (3) quantize pixel values to multiples of 5, (4) normalize pixel intensity by subtracting the minimum value, and (5) compute a 256-bin histogram. Histogram intersection is then used to quantify similarity between the test image and database images.

F. LCS Algorithm for Text Comparison

The Longest Common Subsequence algorithm is applied on tokenized, preprocessed text documents. Both the suspicious document and corpus files undergo stop-word removal, punctuation stripping, lemmatization, and stemming. The LCS score is normalized by the length of the suspicious document to produce a similarity percentage, enabling proportional assessment of textual overlap.

IV. RESULTS

A. Text Plagiarism Detection

When a text file closely matching a corpus document was uploaded (e.g., copying the first corpus file directly), the system reported an LCS similarity score of 1.0 (100%), correctly identifying plagiarism. A non-matching file (e.g., an unrelated Angular.js text file) returned a low similarity score of 0.03, correctly yielding a no-plagiarism result. These results confirm that the LCS-based module provides reliable detection for both direct and near-direct textual plagiarism.

B. Image Plagiarism Detection

For image testing, uploading an image from the database (e.g., image '2.jpg') resulted in a histogram matching score of 40,000, indicating 100% pixel match and triggering a plagiarism verdict. Uploading an unrelated image (e.g., '112.jpg') produced a score of 15,173, well below the threshold of 39,000, correctly returning no plagiarism detected. The histogram comparison graphs visually confirm the degree of match between source and suspicious images.

C. Performance Summary

The proposed system successfully detects plagiarism in both text and image domains. Text similarity using LCS achieves accurate matching with a 60% threshold. Image similarity via FMM-based histogram comparison achieves 81.91% average accuracy on the CLEF-IP 2012 flowchart dataset, outperforming CVC and INRIA baselines. The system demonstrates fast, real-time detection suitable for academic use.

Table I. Test Cases

Test Case Id	Test Case Name	Test Case Desc.	Test Steps	Expected	Actual	Priority
01	User Login	Verify user can login	Submit valid credentials	Login success	Login success	High
02	Upload Source	Verify source	Upload corpus	Files loaded	Files loaded	High

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 8, March 2026
Impact Factor: 8.2

Test Case Id	Test Case Name	Test Case Desc.	Test Steps	Expected	Actual	Priority
	Files	upload	files			
03	Upload Suspicious Files	Verify suspicious upload	Upload and check text	LCS score shown	LCS score shown	High
04	Upload Source Image	Verify image upload	Upload database images	Histograms computed	Histograms computed	High
05	Upload Suspicious Image	Verify image check	Upload test image	Match score shown	Match score shown	High

V. CONCLUSION

This paper presented an integrated system for detecting both text and image plagiarism in academic documents. For textual similarity, the LCS algorithm provided reliable detection with a 60% similarity threshold, correctly identifying both plagiarized and original content. For image similarity, the FMM-based histogram matching method achieved an average recognition accuracy of 81.91% on the CLEF-IP 2012 flowchart database, outperforming both CVC and INRIA comparison methods.

The system is implemented as a web application using Python and Django, offering a user-friendly interface for uploading and checking documents and images. The combined approach addresses a critical gap in existing plagiarism detection tools, which typically ignore visual content. Future work may extend the system to include deep learning-based image feature extraction, cross-language text detection, and integration with online academic databases for broader coverage.

REFERENCES

[1] S. Alzahrani, V. Palade, N. Salim, and A. Abraham, "Using Structural Information and Citation Evidence to Detect Significant Plagiarism Cases in Scientific Publications," JASIST, vol. 63, no. 2, 2011.
[2] S. M. Alzahrani, N. Salim, and A. Abraham, "Understanding Plagiarism Linguistic Patterns, Textual Features, and Detection Methods," IEEE Trans. Syst., Man, Cybern. C, vol. 42, 2012.
[3] B. Gipp, "Citation-based Plagiarism Detection," Springer, 2014.
[4] P. Hurtik and P. Hodakova, "FTIP: A tool for an image plagiarism detection," in Proc. SoCPaR, 2015.
[5] M. Iwanowski, A. Cacko, and G. Sarwas, "Comparing Images for Document Plagiarism Detection," in Proc. ICCVG, 2016.
[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in Proc. NIPS, 2012.
[7] N. Meuschke and B. Gipp, "State-of-the-art in detecting academic plagiarism," IJEI, vol. 9, no. 1, 2013.
[8] S. Srivastava et al., "Image plagiarism detection using SIFT and perceptual hashing," Int. J. Comput. Appl., 2013.
[9] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. ACM Press, 1999.
[10] D. Lopresti and G. Nagy, "Optical Character Recognition by Open Source OCR Tool Tesseract: A Case Study," ICDAR, 2010