Literature Review- Effective Image processing integrated plagiarism checker

Ronlo Thong  
*3rd year CSE*  
Bennett UniversityGreater Noida, Uttar Pradesh, India  
E20cse079@bennett.edu.in

*Abstract*—My Idea is to create a model which would be able to recognize texts from a camera hardware and check it for plagiarism at the same time. I believe the model is relevant because Most plagiarism websites only function for texts and text documents ,Further, even those websites which offer plagiarism checks on images, the images have to be scanned uploaded on the device first which at rush hours become inconvenient. My model will make plagiarism check easier as well as more convenient on Paper Documents/Printed document.

Introduction

The model is a combination of the plagiarism and text detection technology.

The Text detection technology deals with the problem of recognizing all kinds of different characters. Both handwritten and printed characters can be recognized and converted into a machine-readable, digital data format. Put simply, the image taken is processed, the characters extracted, and are then recognized.

While A plagiarism detector searches for similarities between your content and other texts using sophisticated database tools. Universities employ them to scan student assignments. Additionally, you can utilise paid plagiarism detectors to examine your own work before submission. Plagiarism detection tools index and scrape web content, searching your writing for similarities to previously published online materials. Keyword analysis highlights exact matches.

The model will able to use a camera hardware detect and convert to texts and check them for plagiarism.

Optical Character Recognition (OCR) of papers has tremendous practical value given the prevalence of handwritten documents in human exchanges. A discipline known as optical character recognition makes it possible to convert many kinds of texts or photos into editable, searchable, and analyzable data. In the past ten years, academics have developed systems that automatically evaluate printed and handwritten documents in order to convert them to electronic format. This review paper's goals are to present research directions and a summary of previous studies on character recognition in handwritten texts.

The use of ideas, concepts, words, or structures without properly citing the source to benefit in a situation where originality is expected has been defined as academic plagiarism. Academic plagiarism takes many different forms, varying in the degree of obfuscation from unaltered copies (copy and paste) to slightly altered forms, such as combining text passages from different sources (shake and paste), to disguised forms, such as paraphrases, translations, and idea plagiarism, and even the theft of academic data.

##### **Literature Review**

With the goal of comparing an input document to a huge collection and finding all documents that display similarities above a predetermined threshold, plagiarism detection is a specific Information Retrieval (IR) problem. Candidate retrieval and comprehensive comparison are typically the first two stages of the two-stage process used by PD systems [1]. The systems frequently use effective text retrieval techniques for candidate retrieval, like n-gram fingerprinting or vector space models [2]. The systems often use exhaustive string matching for the detailed comparison. Such methods, however, are only effective when looking for close replicas of a text. Researchers have proposed a range of mono-lingual text analysis techniques using semantic and syntactic aspects, as well as cross-lingual IR methods, to identify camouflaged types of academic plagiarism . Additionally, studies have demonstrated the benefits of hybrid techniques, or the combined analysis of text and other content aspects, for improving the re-trieval effectiveness for PD tasks.

The image analysis approaches are relatively resistant to changes in light or the addition of noise, and they are insensitive to simple.

A group of techniques known as perceptual hashing are used to translate the perceived content of pictures, videos, or audio files into a hash value (pHash) [4]. In contrast to cryptographic hashing, which produces significantly different hash values for tiny changes in the input, human perception of similar images likewise yields identical pHashvalues. The similarity of photos' pHash values can be used to quantify how similar they are. feature point techniques fail when picture elements, such as shapes, are rearranged.

For images predominantly comprising text, such as tables inserted as images, feature point techniques and perceptual hashing often also fail to generate meaningful similarities. In conclusion, earlier research on image-based PD proposed methods that reliably retrieve exact and cropped image copies as well as images that underwent one transformation. However, in practise, the feature points for individual learners are typically matched to multiple learners occurring in different places in the comparison document. This prevents the identification of meaningful clusters of matching features. The proposed methods frequently fail when applied to images that have undergone other modifications, such as rearranging shapes in the image, redrawing components of the image, or images that are primarily composed of text. These methods concentrate on photographs, for which they produce good results even when the photoquality is reduced or modified, such as by blurring.

A method called optical character recognition (OCR) transforms input text into a machine-encoded format [5]. OCR is now used to digitise typewritten materials as well as handwritten mediaeval manuscripts [6] and transform them to digital format [7]. Due to the fact that one no longer needs to sift through mountains of documents and files, retrieving the necessary information is now simpler. The needs of organisations for digital preservation of historical data, legal documents, and educational persistence, among others, are being met. The extraction of features and classification/discrimination of these features are what an OCR system relies on the most (based on patterns). As a branch of OCR, handwritten OCR is getting more attention. Based on the input data, it is further divided into offline system and online system . While online systems' nature of input is more dynamic and depends on the movement of a pen tip with a specific velocity, projection angle, position, and locus point, offline systems' nature of input is more static and takes the form of scanned images. As a result, an online system is thought to be more sophisticated and intricate because it fixes the issue with overlapping  search approach, which combines automatic and manual search. To find main studies and get a bigger picture, an automatic search was helpful. As a result, we expanded the review by adding more papers. As advised by Kitchenham et al. , the manual search approach was used to find references for the studies that the automatic search had turned up.

It used traditional databases, which have the most pertinent research articles, for automatic search. These databases include Scopus—Elsevier, IEEE Explore, ISI Web of Knowledge, and Springer.

After employing search terms to find the source data, the research articles were then analysed for data in order to determine their relevance to the study topics and inclusion.

Artificial Neural Networks (ANN) are made up of several processing units known as neurons and have an architecture that was inspired by biological neurons .Together, these processing units (neurons) model the input data and assign it to a predetermined class or label. Nodes are the primary component of neural networks (neuron). Each node's weights are modified in a supervised learning environment (training on labelled samples/data) to decrease the squared error on training samples.

***Conclusive Limitations/Drawbacks:***

OCR models fall short when extracting text from images shot in a variety of settings. An OCR tool finds the first character and traverses in a horizontal direction, searching for subsequent symbols. However, if the image is blurry, the font is unrecognizable, or the text is tilted, OCR fails to yield satisfactory results. The OCR annotation method finds the spaces between words that are still blank in handwritten text or cursive typefaces, identifying each letter as a distinct bounding box. The OCR model accepts that without these spaces, all the characters form an unified pattern and cannot be classified according to any character descriptions, OCR models often generate wrong results when dealing with noisy images. Your OCR model can get baffled between ‘8’ and ‘B’ or ‘A’ and ‘4’. The only way to tackle noisy images is to implement Deep Learning in your OCR solution or use de-noising image processing tools. Plagiarism checkers just look at the words in the work; they cannot tell if you have stolen an idea or piece of information if you haven't also stolen the words. This is a widespread issue in academia, which takes both this type of plagiarism and verbatim plagiarism very severely. Even while many plagiarism detectors try to distinguish between attributed use and unattributed use, it isn't always practicable due to the range of attribution forms. Additionally, many plagiarism detectors will flag matches that are merely coincidence because of how frequently specific English phrases occur.

##### **References**

[1]Benno Stein, Sven Meyer zu Eissen, and Martin Pohast. 2007. Strategies forRetrieving Plagiarized Documents. In Proc. SIGIR.

Norman Meuschke and Bela Gipp. 2013. State-of-the-art in detecting academic

plagiarism. IJEI 9, 1 (2013)

Norman Meuschke and Bela Gipp. 2013. State-of-the-art in detecting academic

plagiarism. IJEI 9, 1 (2013)

Norman Meuschke and Bela Gipp. 2013. State-of-the-art in detecting academic

plagiarism. IJEI 9, 1 (2013)

[2]Norman Meuschke and Bela Gipp. 2013. State-of-the-art in detecting academicplagiarism. IJEI 9, 1 (2013)

[3]Salha M. Alzahrani, Naomie Salim, and Ajith Abraham. 2012. UnderstandingPlagiarism Linguistic Paerns, Textual Features, and Detection Methods. In IEEETrans. Syst., Man, Cybern. C, Appl. Rev., Vol. 42

[4]I. K. Pathan, A. A. Ali and R. J. Ramteke, "Recognition of offline handwritten isolated Urdu character", Adv. Comput. Res., vol. 4, no. 1, pp. 117-121, 2012.

[5]C. C. Tappert, C. Y. Suen and T. Wakahara, "The state of the art in online handwriting recognition", IEEE Trans. Pattern Anal. Mach. Intell., vol. 12, pp. 787-808, Aug. 1990.

[6] M. Kumar, S. R. Jindal, M. K. Jindal and G. S. Lehal, "Improved recognition results of medieval handwritten Gurmukhi manuscripts using boosting and bagging methodologies", Neural Process. Lett., vol. 50, pp. 43-56, Sep. 2018.

[7] M. A. Radwan, M. I. Khalil and H. M. Abbas, "Neural networks pipeline for offline machine printed Arabic OCR", Neural Process. Lett., vol. 48, no. 2, pp. 769-787, Oct. 2018.

**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.**