Handwritten Conversion and Plagiarism Checker

Ms. Sai Nidhi K A
Department of Computer Science and Engineering
MBC College of Engineering
Peermade,Kerala ,India
sainidhika2001@gmail.com

Mr. Ephrem Varkey Oommen
Department of Computer Science and Engineering
MBC College of Engineering
Peermade, Kerala, India
ephremoommen2001@gmail.com

Ms. Aleesha Anna Thomas
Department of Computer Science and Engineering
MBC College of Engineering
Peermade, Kerala, India
annathomasaleesha@gmail.com

Prof. Aryalekshmi R
Assistant professor
Department of Computer Science and Engineering
MBC College of Engineering
Peermade, Kerala, India
aryalakshmir@mbcpeermade.com

Abstract—Using intelligence and a number of technologies, OCR (Optical Character Recognition) is a system that gathers information to turn an image into digital text. The system can be used to check for plagiarism in printed and handwritten documents as well as to help professors or students turn their handwritten notes into digital materials that can be easily shared with others. This platform enables the user to simply turn a camera-captured handwritten image into a paper by identifying it as being handwritten. The app collects data via the mobile device's camera. Document format is stored with the identified text. The key benefits of this application are its ability to check for plagiarism in hard copies, transform handwritten notes to documents, and do other tasks.

Keywords—Optical Character Recognition(OCR), Pytesseract, Android Studio.

I. INTRODUCTION

Nowadays, technology is changing the globe. Many people find it impossible to envision their beloved mobile gadget being without it even for a day. We utilize them for a variety of purposes, including information gathering, keeping in touch with friends and family, navigating new places, making decisions, and many more. But frequently we reach a point when we wish there was an application for a specific circumstance or need, but there isn't.

There is a huge need now for mobile devices to store any material that is currently available on paper, such as books or newspapers. It is already possible to store data by scanning the relevant text, however, this results in a picture that is useless for further processing. For a human, it is simple to recognize handwritten text, but for computer systems, it is a challenging operation. Numerous researchers have worked in this area, although they haven't always been completely accurate. The handwriting of many persons can be recognized by our eyes, but a machine cannot do this as readily. "Optical Character Recognition" is a technique that can help with this issue. 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. Researchers have been using artificial intelligence and machine learning methods to automatically evaluate printed and handwritten documents for the past ten years in order to digitize them. Using intelligence and a number of technologies, OCR (Optical Character Recognition) is a system that gathers information to turn an image into digital text. The system can be used to check for plagiarism in printed and handwritten documents as well as to help professors or students turn their handwritten notes into digital materials that can be easily shared with others. As one of the first steps in OCR recognition, documents are optically scanned first. The next step is recognition, which involves converting the images into character streams that represent the letters of words that have been identified. The last step is retrieving or saving the text that has already been transformed. Information from the retrieved text is used to convert the text. Once a printed or typed text image has been converted, other necessary activities are made simpler. A word or phrase can be used on a website, altered, and sent as an email, among other things.

The system can be used to check for plagiarism in printed and handwritten documents as well as to help professors or students turn their handwritten notes into digital materials that can be easily shared with others. Before printing and binding their theses, the majority of students use plagiarism detection software to check them. It's never certain whether pupils actually need to use plagiarism checkers. Verifying your academic paper for plagiarism is a must for any university's evaluation process, regardless of whether you utilized a plagiarism detector or not.

In this work, we use Python to create an Android application that can read handwritten writings and find instances of plagiarism in two different documents. The "Handwritten Conversion and Plagiarism Checking Application" platform allows users to transform handwritten text quickly and easily from an image that has been acquired with a camera into a document. The software takes input through the mobile device's camera. Document format is

stored with the identified text. The ability to turn handwritten notes into documents and check hardcopy for plagiarism are the key benefits of this tool. Using this feature through a single application is the project's goal. It could benefit for researchers, teachers, or students who do their work in handwriting. Thereby we can use this application for our day-to-day life.

II. PROPOSED METHODOLOGY

The OCR (Optical Character Recognition is the technology that collects data to convert the image into digital text using intelligence and a variety of technology. Documents are first optically scanned as one of the first steps in OCR recognition. The OCR can help teachers or students to convert their handwritten material into digital material that can be easily distributed to others and it can also use to check the plagiarism of both handwritten and printed documents. This platform gives access to the user to recognize the handwritten from the captured image through a camera and it easily converts into a document.

Handwritten characters are read using the Pytesseract module in the proposed work. A Python optical character recognition (OCR) tool is called Pytesseract, also known as Python-tesseract. It can read and recognise text on photos, licence plates, and other visual media. To interpret the words from the provided image in this case, we'll utilise the tesseract package. Using tfidf vectorization, we turn papers into vectors for the purpose of detecting plagiarism between two documents. The cosine similarity of these vectors yields the plagiarism score. For user convenience, we will also incorporate these tools into an Android application.

A. Login

The interface used by providers of authentication technology is described by the term "Login Module." Applications have login modules put in to offer a certain kind of authentication.

Authentication technology providers implement the Login Module interface while applications write to the Login Context API. The Login Module(s) to be used with a specific login application are specified by a Configuration. The web application's front end is a login module. It is a piece of software that enables users to log in to a system or application by establishing their identity using a username and password. The module normally offers a user interface for inputting the login information and then verifies that the user is permitted to access the system by comparing the entered information to a database or other authentication method. It is a necessary element of many software programmes, especially those that demand secure access to sensitive information or functionality. These apps can make sure that only permitted users are given access to the system by asking users to authenticate themselves with a login module.

A graphical user interface or command-line interface is often used by the login module to request users to input their username and password. The module checks the user's credentials against a database or another authentication method after they have been entered to see if they are valid for access to the system. The additional authentication procedures can add an additional layer of security and aid in preventing unauthorized access to sensitive systems and data. Username and password are utilized in this paper's login interfaces for current users, and registering a new user who will likely use the application is a possibility as well. The various methods by which a user can log in to a system or application are referred to as login options. These choices often consist of:

1)Password and user name: The user provides their username and password to access the system under this most popular type of authentication. This approach is popular since it's easy to use and apply, but it's also the least secure.

2)Social media login: A few systems allow users to sign in with their Google or Facebook accounts. Users who don't want to make a new account for each system they use can take use of this feature, however it might also present privacy issues. The procedure through which a user creates a new account with a system or application is referred to as the register login option. The normal steps in this procedure include entering the user's personal data, such as name, email address, and date of birth, selecting a username and password, and accepting the terms of service and privacy statement of the system. The user can access the system by entering their selected username and password after completing the registration process. Before allowing access to specific features or services, the system could ask the user to confirm their email address or other forms of communication.

For systems and programmes, social media platforms, and email services that require user accounts to access particular features or services, the register login option is crucial. But, it's crucial to put in place the right security measures during the registration process to prevent unwanted access and preserve user privacy. It enables the system to authenticate and approve users. The user's complete information will be kept in a safe database. The register login option is a crucial part of many systems and applications, giving users a safe and convenient way to create and manage their accounts. The login module is a crucial part of many software applications, providing a secure and reliable mechanism for users to authenticate themselves and gain access to the system.

B Text recognition using OCR

OCR (Optical Character Recognition), a technology that allows computers to extract text from scanned photographs or other digital sources, is used for text recognition. An image of a printed or handwritten document is analysed using OCR software, which then recognises the individual characters and words. As a result, the image can be transformed by the computer into searchable, editable text that can be edited. In many different applications, including document management, data entry, and digital archiving, OCR technology is frequently employed. For instance, paper documents can be scanned and converted into digital files using OCR software, enabling users to swiftly and conveniently search and obtain the text content of the papers.

Here, Python-tesseract is utilised as a tool for Python's optical character recognition (OCR). . In other words, it will identify and "read" any text that is contained in photos. Google's Tesseract-OCR Engine is wrapped in Python-tesseract. It is an OCR engine that can be used to extract text

from photos and other sources and is opensource. With a straightforward and user-friendly API, Pytesseract enables developers to include Tesseract OCR into their Python projects. Installing Tesseract OCR on your computer is a prerequisite for using Pytesseract. The Python package manager, pip, can then be used to install the pytesseract package. Once installed, you may use the pytesseract module's functions to conduct OCR on photos or other sources by importing it into your Python programmer.

As an illustration, the following line of code demonstrates how to utilise Pytesseract to extract text from an image file:

```
import pytesseract
 from PIL import Image
# Load the image
image = Image.open('example.png')
 # Perform OCR on the image
text = pytesseract.image_to_string(image)
# Print the recognized text
print(text)
```

The PIL (Python Imaging Library) module is used to import the "example.png" picture file, and the image to string function of pytesseract is used to conduct OCR on the loaded image and extract the text. The identified text is then printed to the console by the code. We strive to integrate handwritten character recognition using OCR because many people find it difficult and time-consuming to convert the handwritten into an editable form.

OCR software typically entails the following steps: Image preparation The image quality needs to be adjusted in this step to make it ideal for OCR analysis. This can entail eliminating any noise or distortion, boosting contrast, and adjusting any skew or perspective.

1) Text detection: The software then uses methods including edge detection, blob analysis, and template matching to pinpoint the areas of the image that are textfilled.

2)Character recognition: A database of recognised characters is searched for each character in the text sections by the software, which then analyses each character. Machine learning methods may be used in this situation to increase recognition accuracy.

3)Post-processing: The programme may carry out further processing stages, including spell-checking and formatting, after the text has been detected to fix any mistakes or inconsistencies. Steps in image processing:

a)upload a photo for handwriting recognition (scan image or upload from gallery)

b)Modify the chosen picture.

c)Using OCR, extract the text from the image and convert it into an editable document.

The level of OCR accuracy still varies according on the original image's quality, the difficulty of the text, and the particular OCR programme being used. OCR text recognition is a useful tool for digitising and managing text-based documents, and it's utilised extensively across many sectors, including government, banking, and healthcare.

B. *Plagiarism Checker*

The act of taking someone else's words or ideas without giving due credit or permission is known as plagiarism, and plagiarism checkers are tools that examine written content to find any instances of plagiarism. Writers, educators, and corporations can utilise plagiarism checkers to make sure that written work is unique and hasn't been plagiarised. Plagiarism checkers come in a variety of forms, from straightforward web-based solutions to complex software packages. Common characteristics of plagiarism detectors include:

1)Text comparison: To find any parallels or matches, the given text is compared to a database of previously published works.

2) Citation analysis: This entails reviewing the text for accurate source citations and identifying any that are absent or incomplete.

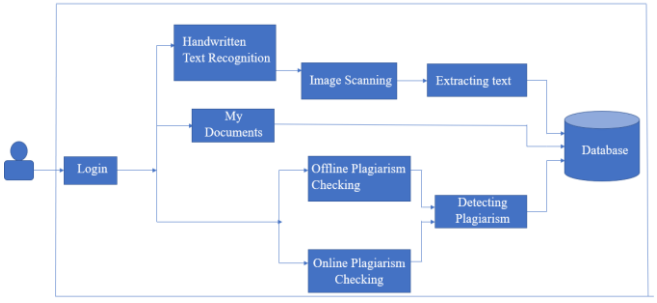
3) Paraphrasing detection: This entails identifying instances of paraphrasing or rephrasing, even when the words have been altered, in the text.

Turnitin, Grammarly, and Copyscape are a few wellknown plagiarism detectors. These tools employ a range of methods, such as text analysis, machine learning algorithms, and natural language processing, to find plagiarism.

We provide both online and offline plagiarism checkers as choices for our plagiarism-checking services. To verify plagiarism in an online setting, use an online plagiarism checker. In online plagiarism, we use the https://plagiarism.studyclerk.com/ platform for checking plagiarism. A popup menu will be available when the online plagiarism option is chosen, when clicking the link a browser will be opened. Through this site, we can upload our content as pdf, .doc docx, txt, rtf, and gds, and check for plagiarism. The majority of college students do not use a plagiarism checker once they finish editing their academic papers, Through this platform, we can upload the portion that we want to check. While clicking the online plagiarism option it directly directs to the site By scanning or uploading the image, it checks for plagiarism and notifies the user if it is found. In order to determine whether plagiarism has been discovered, an offline plagiarism checker compares two samples that have been posted through a gallery or scanned.

It's crucial to remember that no plagiarism detector is perfect, and certain cases of plagiarism could go unnoticed. Moreover, plagiarism checkers should be used as a tool to help identify plagiarism rather than as a substitute for appropriate citation and good writing habits. The writer must still make sure that their writing is authentic and properly cited.

III. . IMPLEMENTATION



First, Android Studio is being used to create an Android app. The user has the option to take a picture using the GUI. Those using this application for the first time need to register. Each user must register their own information. The database will be used to hold all user information, and it can be safeguarded. In the registration, the screen requires more fields to register that is followed by a name, phone number, address, username, and password.

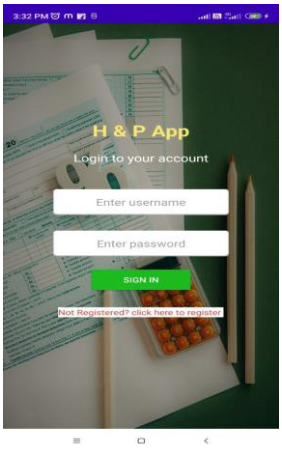


Fig 2. The Login Page of app

Fig 3. Option for Register

The OCR engine receives the binary picture after that for further processing. The system is built to recognise handwritten letters as input data quickly and accurately, modify the output of that recognition, and save the edited output in document format. Also, a choice to share the system's output, which comes in document format and displays the findings of the plagiarism check. The user can capture the handwritten document using the Interface. The user has two options for loading images: camera capture or gallery loading. The user can capture the handwritten document using the Interface. The user has two options for loading images: camera capture or gallery loading. There are four icons in the sub-window. presenting the handwritten text recognition first, then the Offline plagiarism checker, my document,and finally the Online plagiarism checker. The image is chosen for processing when the camera or gallery icon is clicked. Cropping and rotating the image after opening it can be done by pressing the done button, which sends the altered image to the OCR engine.

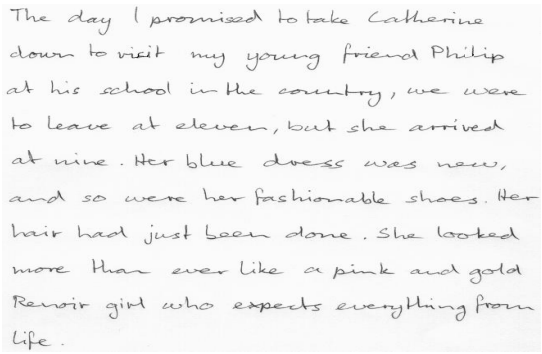


Fig 4. Test Image

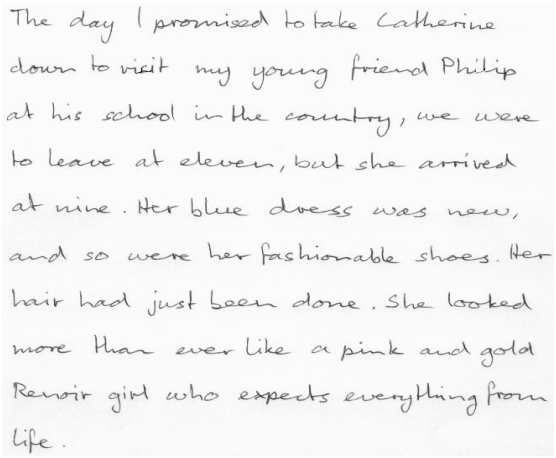


Fig 5. After Removal of Noise

The text converted from the image is displayed on the screen. This text can be changed. After that, there is a choice to save the converted text in document format (.doc). This database stored document file, which is composed of transformed text, will be used later. Once the procedure is complete, the document file can be opened, edited, and saved whenever you like in the future. This document file can also check plagiarism.

When the user is select the plagiarism icon from the subwindow. There is a two options in it to check the plagiarism : online plagiarism checker and offline plagiarism checker. Online plagiarism checker is used to check plagiarism through online mode. By scanning or uploading the image it scans and report if plagiarism is detected. Offline plagiarism checker is used to check 2 samples which is uploaded through gallery or scanned, by comparing these two samples it will report if plagiarism is detected. It can also detect plagiarism in printed and handwritten texts. This application assists educators, researchers, and students in digitizing their documents and checking them for plagiarism. Doing all features under one roof is more practical and takes less time.

IV. RESULT

The image is taken with a camera or uploaded from a gallery in this system. Using an Android application that has been built, the image is turned into text. The text that has been transformed is saved as a document file that can be updated as needed in the future. The key issue is that since we are utilizing OCR to recognize handwritten characters, accuracy problems could arise. The handwritten text produced a test score of 94.1% correctness while the printed text produced 100% accuracy. By using an offline plagiarism checker we can detect the plagiarism of assignments, and projects of students. Mainly we are focusing this application on teachers and students, so it is completely free. The benefit of our approach is that character recognition doesn't require internet connectivity.

V. CONCLUSION

Using an Android app, this approach converts handwritten character recognition into editable text. The user is given the option to choose which portion of the image they want to convert after the image has been acquired by the camera and

imported into the Android app. The OCR engine continues processing and displays the transformed text on the screen. Document format is stored with the identified text. There is an option to alter the recognized text and save them. And also we can check the plagiarism in online and offline mode. The user does not need to manually enter the input language in this technique. We are currently working to recognize other languages in editable text. When text is printed instead of handwritten, accuracy is increased.

ACKNOWLEDGMENT

Would like to thank the authors of the study who provided the useful material as well as our Department of Computer Science and Engineering for their assistance and support throughout the project's full journey.

REFERENCES

[1] Thakare, S., Kamble, A., Thengne, V., & Kamble, U. R. (2018, December). Document segmentation and language translation using tesseract-ocr. In 2018 IEEE 13th International Conference on Industrial and Information Systems (ICIIS) (pp. 148-151). IEEE

[2] Mainkar, V. V., Katkar, J. A., Upade, A. B., & Pednekar, P. R. (2020, July). Handwritten character recognition to obtain editable text. In 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC) (pp. 599- 602). IEEE.

[3] Cheers, H., Lin, Y., & Smith, S. P. (2021). Academic source code plagiarism detection by measuring program behavioral similarity. IEEE Access, 9, 50391-50412.

[4] Ljubovic, V., & Pajic, E. (2020). Plagiarism detection in computer programming using feature extraction from ultrafine-grained repositories. IEEE Access, 8, 96505-96514.

[5] Wu, K., Fu, H., & Li, W. (2020, October). Handwriting Textline Detection and Recognition in Answer Sheet Composition with Few Labeled Data. In 2020 IEEE 11th International Conference on Software Engineering and Service Science (ICSESS) (pp. 129-132). IEEE.

[6] Ali, W., Ahmad, T., Rehman, Z., Rehman, A. U., Shah, M. A., Abbas, A., & Dustgeer, G. (2018, September). A Novel Framework for Plagiarism Detection: A Case Study for Urdu Language. In 2018 24th International Conference on Automation and Computing (ICAC) (pp. 1-6). IEEE.

[7] Acharya, M., Chouhan, P., & Deshmukh, A. (2019, December). Scan. it-Text Recognition, Translation and Conversion. In 2019 International Conference on Advances in Computing, Communication and Control (ICAC3) (pp. 1-5). IEEE.

[8] Mishra, P., Pai, P., Patel, M., & Sonkusare, R. (2020, November). Extraction of information from handwriting using optical character recognition and neural networks. In 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA) (pp. 1328-1333). IEEE.

[9] Ibrahin, A. S. B., Khalifa, O. O., & Ahmed, D. E. M. (2020, September). Plagiarism Detection of Images. In 2020 IEEE Student Conference on Research and Development (SCOReD) (pp. 183-188). IEEE.

[10] Parthiban, R., Ezhilarasi, R., & Saravanan, D. (2020, July). Optical character recognition for English handwritten text using recurrent neural network. In 2020 International Conference on System, Computation, Automation and Networking (ICSCAN) (pp. 1-5). IEEE.

[11] https://plagiarism.studyclerk.com/