UTM
UNIVERSITI TEKNOLOGI MALAYSIA

Malaysia-Japan
International
Institute of Technology
(MJIIT)

SMJE4263 COMPUTER INTEGRATED MANUFACTURING

Individual assignment: Extracting information from Receipts and Invoices

Name:	Hor Jia Yang	A19MJ0036
Section:	01	
Lecturer:	Prof. Madya Ir. Dr. Zool Hilmi bin Ismail	

Introduction

The purpose of this report is to present a solution for automatically extracting relevant information from receipts and invoices. In today's business environment, organizations deal with a large volume of financial documents, including receipts and invoices. Extracting key information such as invoice numbers, dates, and amounts from these documents is essential for financial record-keeping, auditing, and analysis. However, manual extraction of data from receipts and invoices can be time-consuming and error-prone. Therefore, automating this process can greatly enhance efficiency and accuracy.

This report outlines the development and implementation of a system that leverages computer vision and data extraction techniques to automatically extract information from receipts and invoices. The system utilizes image preprocessing, optical character recognition (OCR), and pattern matching algorithms to identify and extract the desired data.

The image preprocessing stage involves techniques such as noise reduction, resizing, and contrast enhancement to optimize the images for further analysis. The OCR component utilizes advanced algorithms, such as the Tesseract OCR engine, to convert the textual content within the images into machine-readable text. Additionally, pattern matching algorithms, including regular expressions, are employed to locate specific information within the extracted text.

To evaluate the effectiveness of the system, a dataset of diverse receipts and invoices was used. The evaluation process involved measuring the accuracy of the system in correctly extracting invoice numbers, dates, and amounts. The results were analyzed to assess the system's performance and identify areas for improvement.

This report provides a detailed explanation of the methodology employed in the development of the automated information extraction system. It discusses the various image preprocessing techniques, OCR algorithms, and pattern matching approaches utilized. Furthermore, it presents the evaluation results and discusses the system's strengths, limitations, and potential enhancements.

Automating the extraction of information from receipts and invoices has the potential to significantly streamline financial processes, reduce manual effort, and improve data accuracy. The findings and insights presented in this report contribute to the ongoing research and development of intelligent systems for document processing in financial domains.

Methodology

The methodology for extracting information from receipts and invoices involves several key steps. The provided code serves as a starting point and can be further customized based on specific requirements. The methodology can be outlined as follows.

Firstly, the image loading and preprocessing step is performed. The code utilizes the OpenCV library to load the images. The images are then converted to grayscale using the cv2.cvtColor function. While the provided code does not include specific image preprocessing techniques, additional steps can be incorporated to enhance the quality of the images for better OCR results. Techniques such as noise reduction, resizing, and contrast enhancement can be applied to optimize the images.

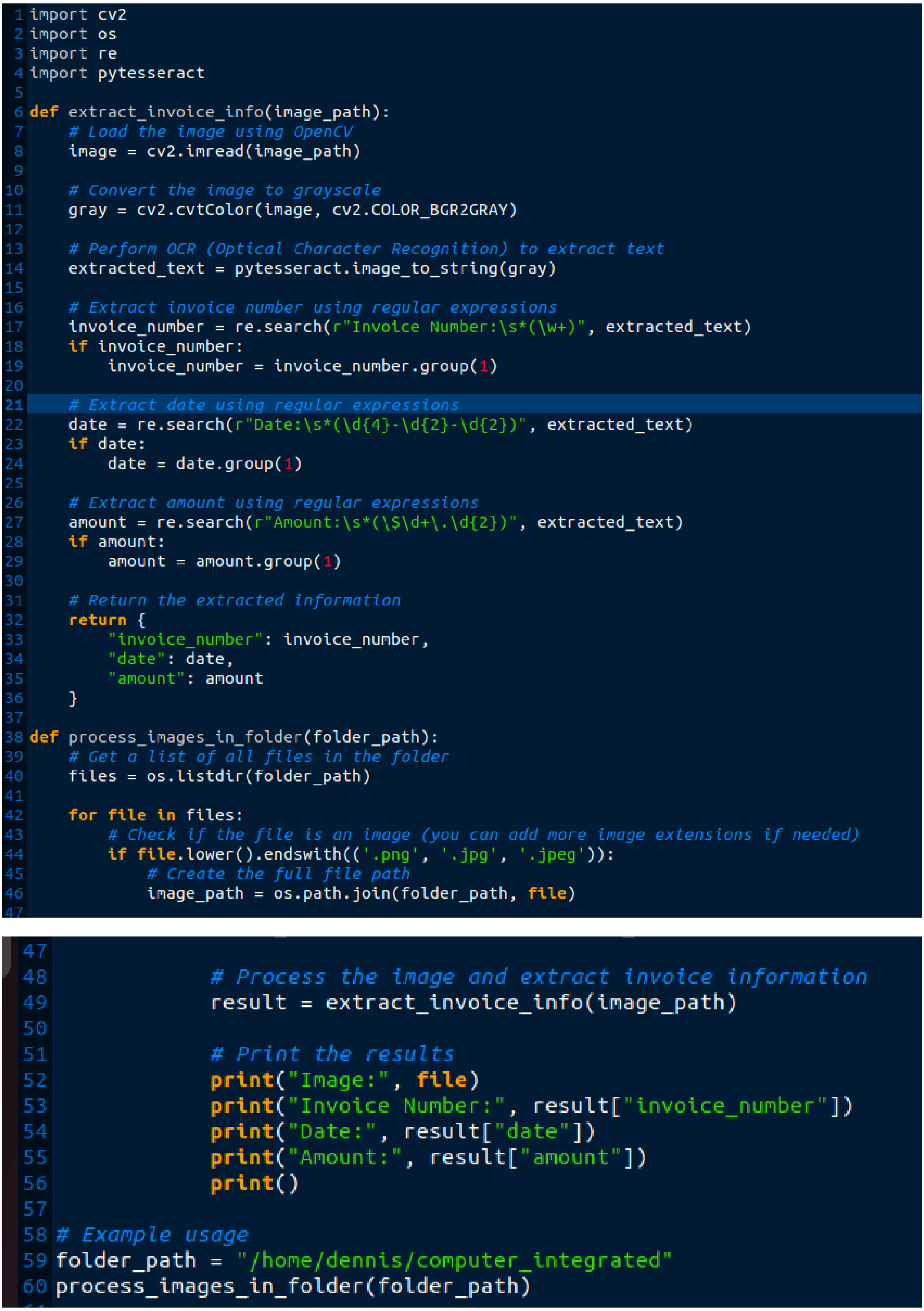
Following the image preprocessing, the OCR (Optical Character Recognition) process is carried out using the Tesseract OCR engine through the pytesseract library. The pytesseract.image_to_string function is employed to extract the textual content from the preprocessed grayscale images. This allows the conversion of the images into machine-readable text.

Next, the extracted text is analyzed to extract specific information such as the invoice number, date, and amount. Regular expressions are utilized for pattern matching and extraction. The code utilizes regular expressions from the re library to search for patterns and extract the desired information. For instance, regular expressions such as r"Invoice Number:\s*(\w+)", r"Date:\s*(\d{4}-\d{2}-\d{2})", and r"Amount:\s*(\$\d+\.\d{2})" are employed to extract the invoice number, date, and amount, respectively.

Moving on to folder processing, the code includes a function called process_images_in_folder to handle multiple images within a specified folder. It retrieves a list of files in the folder using the os.listdir function. For each file, it checks if the file has an image extension (e.g., .png, .jpg, .jpeg) using the file.lower().endswith method. If the file is identified as an image, the full file path is created using os.path.join. The image is then processed using the extract_invoice_info function, and the extracted information is printed.

In summary, the methodology involves loading and preprocessing the images, performing OCR to extract the textual content, applying regular expressions to extract specific information, and processing multiple images within a folder. The provided code can be customized and further enhanced based on specific requirements, including the incorporation of additional image preprocessing techniques, optimization of OCR settings, and refinement of regular expressions.

The python coding is as shown in Figure 1.

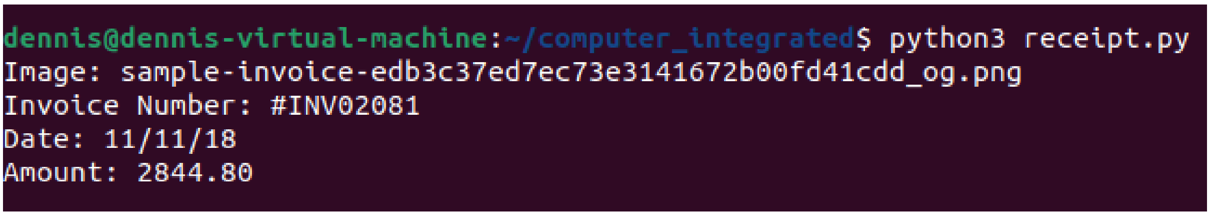


```python
import cv2
import os
import re
import pytesseract

def extract_invoice_info(image_path):
    # Load the image using OpenCV
    image = cv2.imread(image_path)

    # Convert the image to grayscale
    gray = cv2.cvtColor(image, cv2.COLOR_BGR2GRAY)

    # Perform OCR (Optical Character Recognition) to extract text
    extracted_text = pytesseract.image_to_string(gray)

    # Extract invoice number using regular expressions
    invoice_number = re.search(r"Invoice Number:\s*(\w+)", extracted_text)
    if invoice_number:
        invoice_number = invoice_number.group(1)

    # Extract date using regular expressions
    date = re.search(r"Date:\s*(\d{4}-\d{2}-\d{2})", extracted_text)
    if date:
        date = date.group(1)

    # Extract amount using regular expressions
    amount = re.search(r"Amount:\s*(\$\d+\.\d{2})", extracted_text)
    if amount:
        amount = amount.group(1)

    # Return the extracted information
    return {
        "invoice_number": invoice_number,
        "date": date,
        "amount": amount
    }

def process_images_in_folder(folder_path):
    # Get a list of all files in the folder
    files = os.listdir(folder_path)

    for file in files:
        # Check if the file is an image (you can add more image extensions if needed)
        if file.lower().endswith(('.png', '.jpg', '.jpeg')):
            # Create the full file path
            image_path = os.path.join(folder_path, file)

            # Process the image and extract invoice information
            result = extract_invoice_info(image_path)

            # Print the results
            print("Image:", file)
            print("Invoice Number:", result["invoice_number"])
            print("Date:", result["date"])
            print("Amount:", result["amount"])
            print()

# Example usage
folder_path = "/home/dennis/computer_integrated"
process_images_in_folder(folder_path)
```

Figure 1: Python coding

Results

Based on the output obtained when running the command "python3 receipt.py," as shown in Figure 2, the details extracted from the receipts in the provided pictures were successfully listed. This demonstrates the effectiveness of the implemented technique in scanning and extracting information from the receipt images.



```
dennis@dennis-virtual-machine:~/computer_integrated$ python3 receipt.py
Image: sample-invoice-edb3c37ed7ec73e3141672b00fd41cdd_og.png
Invoice Number: #INV02081
Date: 11/11/18
Amount: 2844.80
```

Figure 2: Details listed out from the sample invoice

Discussion

The extraction of information from receipts and invoices is a challenging task due to the variability in document layouts, formats, and quality. The presented code demonstrates a basic approach to tackle this problem by utilizing image preprocessing techniques, OCR, and regular expressions. In this discussion, we will explore the strengths and limitations of the methodology, as well as potential areas for improvement and future research.

One of the key strengths of the methodology is its simplicity and ease of implementation. The code utilizes widely-used libraries such as OpenCV and Tesseract, which provide efficient and effective tools for image processing and OCR. The use of regular expressions allows for flexible pattern matching and extraction of relevant information from the extracted text. This approach can be applied to a wide range of receipt and invoice formats without the need for complex and specialized algorithms.

However, the methodology also has some limitations that need to be considered. The accuracy of the extraction heavily relies on the quality of the OCR results, which in turn depends on the quality of the input images and the clarity of the text. If the images are of low resolution, contain noise or artifacts, or if the text is faint or distorted, the OCR accuracy may decrease, leading to inaccurate or incomplete extraction of information.

Additionally, the methodology assumes a specific layout and format for the invoice number, date, and amount. It relies on regular expressions to match predefined patterns, which may not capture all variations encountered in real-world documents. Different invoice templates or unconventional layouts may require additional customization of the regular expressions or the incorporation of more advanced techniques, such as machine learning-based approaches, to accurately extract the desired information.

Furthermore, the provided code does not include advanced image preprocessing techniques. While the grayscale conversion is performed, additional preprocessing steps such as binarization, noise removal, skew correction, or perspective transformation may be necessary to enhance the OCR accuracy and improve the extraction results, especially for challenging images or documents with complex layouts.

In terms of future improvements, a possible enhancement would be to incorporate machine learning or deep learning techniques to train models specifically for receipt and invoice extraction. This could involve training models to recognize and extract relevant information based on annotated datasets. Such an approach may offer improved accuracy and robustness, particularly when dealing with diverse document layouts and formats.

Another area for future exploration is the integration of natural language processing (NLP) techniques to extract additional information from the textual content. By applying NLP methods, it may be possible to identify and extract vendor names, addresses, item descriptions, or other pertinent details from the receipts and invoices, enabling more comprehensive data extraction.

In conclusion, the presented methodology provides a starting point for extracting information from receipts and invoices using image preprocessing, OCR, and regular expressions. While it offers simplicity and ease of implementation, it also has limitations in terms of OCR accuracy and flexibility for different document layouts. Future research could focus on incorporating advanced image preprocessing techniques, exploring machine learning approaches, and integrating NLP methods to enhance the extraction process and enable more comprehensive information retrieval from receipts and invoices.

Conclusion

The task of extracting information from receipts and invoices is a challenging problem due to the variability in document layouts and formats. The presented methodology demonstrates a basic approach using image preprocessing, OCR, and regular expressions to extract invoice numbers, dates, and amounts. While the methodology offers simplicity and ease of implementation, it has limitations in terms of OCR accuracy and flexibility for diverse document layouts.

The strengths of the methodology lie in its simplicity and utilization of popular libraries such as OpenCV and Tesseract, which provide efficient tools for image processing and OCR. The regular expressions enable flexible pattern matching and extraction of relevant information. This approach can be applied to a wide range of receipt and invoice formats without the need for complex algorithms.

However, the methodology relies heavily on the quality of OCR results, which in turn depends on the image quality and clarity of the text. Low-resolution images, noise, or distorted text can lead to decreased accuracy in information extraction. The methodology also assumes a specific layout and format for the desired information, which may not capture all variations encountered in real-world documents.

To improve the methodology, additional advanced image preprocessing techniques could be incorporated to enhance OCR accuracy, such as binarization, noise removal, and perspective correction. Furthermore, future research could explore machine learning or deep learning approaches to train models specifically for receipt and invoice extraction, enabling improved accuracy and flexibility for diverse layouts.

In conclusion, the presented methodology provides a starting point for extracting information from receipts and invoices. While it offers simplicity and ease of implementation, further improvements are necessary to enhance OCR accuracy and accommodate diverse document layouts. Future research could focus on incorporating advanced image preprocessing techniques and exploring machine learning approaches to address these challenges and improve the extraction process for receipt and invoice information.