Review-1 Report

Improving Text Extraction Accuracy with Image Preprocessing

By,
KOMALI BEERAM (160118733067)
SOUMYA VEMURI (160118733081)

Department of Computer Science and Engineering,
Chaitanya Bharathi Institute of Technology
(Autonomous),
(Affiliated to Osmania University, Hyderabad)
Hyderabad, TELANGANA (INDIA) –500 075
[2021-2022]

ABSTRACT

Digitization allows us to immortalize a physical entity by creating a digital representation of it on our devices. It saves us time in manually sifting through physical storage units such as albums and notebooks and provides us with programs to manage and secure our data. We often take images of Receipts or Invoices, Identity Cards, and nutritional labels to save a copy of their details. This can be taken a step further by automating the process of information extraction and documentation.

Advancements in computer vision have provided us with the expertise to create tools for text detection and extraction. But it is still an ongoing challenge because documents with unstructured layouts, poor image quality, and noise around the text yield very low accuracy in text extraction results. Conquering this challenge would require the image to be highly enhanced through pre-processing techniques such as Brightness Correction, Contour Detection, Skewness Correction, Morphology, and Binarization. A mechanism made from the best combination of image pre-processing techniques prior to text extraction can improve text accuracy to a large extent.

Table of Contents

1. INTRODUCTION

In today's day and age, an increase in demand for digitization has fueled a massive growth in technology and communication and the use of printed materials such as books and papers has significantly reduced. Also, it is easier to organize digitized data and analyze them for various purposed with many advanced techniques like artificial intelligence etc. To translate physical and handwritten documents into digital copies, optical character recognition (OCR) has come into the sight of researchers and since its first advent it has undergone significant changes in methodology and made considerable progress towards its goal.

1.1. Problem Statement

Historical OCR engines have their accuracy lying between 70-80% for a high-quality image at page level. That means in a page of 100 words 70-80 words are accurate. This will lead to significant inaccuracies if used on a large volume of sensitive documents.

Through this project we aim to improve the accuracy of the result by implementing an ideal combination of image preprocessing techniques prior to text extraction.

1.2. Methodology

I. **Document Scan and File Upload**
User scans their document and uploads it to the system interface. Documents include invoices, receipts, nutrition labels, book covers etc.

II. **Image Preprocessing Pipeline**
Scanned documents undergo the best combination of image preprocessing techniques for maximum image text enhancement. Image Preprocessing Techniques include

   a. Lighting Correction
   Lighting correction corrects light variation produced by surface relief or document curvature. When a thick book is scanned, the shadow of the binding will appear on the image. This technique allows obtaining light uniformity and eliminate such shadows, whether vertical or horizontal.

   b. Contour Detection
   Contours can be explained simply as a curve joining all the continuous points along the boundary, having the same color or intensity. The contours are a useful tool for shape analysis and object detection and recognition. For better accuracy, binary images are used.

   c. Scaling of Image

Ensure that the images are scaled to the right size which usually is of at least 300 Dots Per Inch. Keeping DPI lower than 200 will give unclear and incomprehensible results while keeping the DPI above 600 will unnecessarily increase the size of the output file without improving the quality of the file. Thus, a DPI of 300 works best for this purpose.

d. Skewness Correction
   When the optical axis of the camera is not perpendicular to the text plane, Perspective distortion occurs. Text boundaries lose rectangular shapes and characters distort, decreasing the performance of recognition models trained on undistorted samples.

e. Noise Removal
   Noise can drastically reduce the overall quality of the OCR process. It can be present in the background or foreground and can result from poor scanning or the poor original quality of the data.

f. Binarization
   This step converts a multicolored image (RGB) to a black and white image. There are several algorithms to convert a color image to monochrome image, ranging from simple thresholding to more sophisticated zonal analysis.

g. Contrast Correction
   Low contrast can result in poor OCR. Increase the contrast and density before carrying out the OCR process. Increasing the contrast between the text/image and its background brings out more clarity in the output.

III. **Text Extraction**
   The image goes through an Optical Character Recognition (OCR) engine that’s been built to recognize printed text in paper documents, handwritten characters, and text elements in the image.

IV. **Information Extraction**
   Information Extraction involves extracting meaningful information from raw text data into a structured format.

1.3. Outline of Report

For performance of recognition, we will simulate the proposed text recognition system to various characters in English language and calculate the recognition rate or accuracy which is expected to improve from the current OCR engines.

1.4. Scope of Project

The scope of our project on a grid infrastructure is to provide an efficient and enhanced software tool for the users to perform document image analysis, document processing by reading and recognizing the characters in research, academic, governmental, and business organizations that are having large pool of documented, scanned images. Irrespective of the size of documents and the type of characters in documents, the product is recognizing them, searching them, and processing them faster according to the needs of the environment.

1.5. Organization of Report

The report is divided into two parts. Each part deals with different aspects of our project. Each part has various chapters explaining in detail.

- Part 1: Introduction
  This part summarizes about the project in brief. It includes the problem statement, methodology, and outline of our results and the scope of the project.
- Part 2: References
  This part has a consolidated list of papers that we referred to develop the project.

Although the document may be read from front to back for a complete understanding of the project, it was written in sections and hence can be read as such. For an overview of the document and the project itself, refer to Introduction.

2. REFERENCES

[1]. Yang Zhang, Hao Zhang, HaoranLi, “Event Info Extraction from Flyers”, 2021.

[2]. Mittal, Rishabh; Garg, Anchal, “Text extraction using OCR: A Systematic Review”, 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), 2020, pp 357–362.

[3]. Dan Sporici, Elena Cușnir and Costin-Anton Boiangiu, “Improving the Accuracy of Tesseract 4.0 OCR Engine Using Convolution-Based Preprocessing”, International Journal of Innovative Technology and Exploring Engineering, Volume-12 Issue-5, May 2, 2020.

[4]. Sanjeev Kumar, Mahika Sharma, Kritika Handa, Rishika Jaiswal, “Improve OCR Accuracy with Advanced Image Preprocessing using Machine Learning with Python”, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-7, May 2020.