# Intelligent Document Processing

Ameya Padwad
*Northeastern University*
*Boston, MA*
padwad.a@northeastern.edu

Risa Samanta
*Northeastern University*
*Boston, MA*
samanta.r@northeastern.edu

Ashutosh Rane
*Northeastern University*
*Boston, MA*
rane.as@northeastern.edu

Sanjana J Dhangundi
*Northeastern University*
*Boston, MA*
dhangundi.s@northeastern.edu

***Abstract*— This paper proposes an intelligent document processing system that uses the VGG architecture, which is a Convolutional Neural Network (CNN), to classify images of documents into four categories: receipts, technical papers, newspapers, and book covers. Following categorization, the system uses Optical Character Recognition (OCR) in order to extract information based on the category. For receipts, the system identifies total expenses, and date; for technical papers, it extracts titles, authors, and conference titles; for newspapers, it captures the name and date; and for book covers, it recognizes titles and authors. In the future, this model can be customized to operate with various document types and extract different types of information. The resulting system could be deployed across organizations to facilitate efficient and rapid document processing.**

***Keywords*— Object Detection, Image Classification, Optical Character Recognition, CNN, VGG***

## I. Introduction

The processing of document images using computer vision is a critical task in numerous applications, from administrative automation to academic research. Traditional methods often rely on generalized OCR techniques that lack the specificity required for various document types, leading to suboptimal text extraction. Furthermore, the process is usually slow and inefficient. This research aims to address these limitations by employing a CNN-based classification system combined with an OCR framework. This approach is designed to tailor text extraction techniques according to the document type, potentially revolutionizing how data is extracted from diverse document formats.

## II. Literature Survey

The research paper, *"An approach towards Real-Time Object Detector Using Open CV"* was proposed in 2022 by M. Kumar and R. Bhatt. In this study, the authors outline a strategy for deploying a Yolo algorithm using OpenCV and Python to create a system that can identify items such as people, bicycles, cars, buses, and boats from an image or a stream of images that are fed to it in the form of previously recorded video or real-time input from a camera. Object detection accurately draws bounding boxes around discovered objects, providing information about the location of the requested object in the video or image. [1]

In 2020, R. Mittal and A. Garg presented *"Text extraction using OCR: A Systematic Review"* introduces the concept of Optical Character Recognition, explains the process of extraction, presents the latest techniques, technologies, and current research in the area. This paper discusses modern systems including popular engines like Tesseract and applications in domains such as healthcare and banking. It elaborates on the entire OCR process from image acquisition to post-processing, emphasizing improvements in feature extraction and classification techniques. Furthermore, it explores ongoing challenges in OCR accuracy, particularly with complex scripts and languages, and suggests potential areas for future research, including integration with AI and AR to enhance functionality and application scope. [2]

The *"Images Classification of Dogs and Cats using Fine-Tuned VGG Models"* was a project presented by Mahardi, I. -H. Wang, K. -C. Lee and S. -L. Chang in 2020. This paper presents the development of an image classifier for identifying various breeds of dogs and cats using fine-tuned VGG16 and VGG19 models. The classifiers were trained on a dataset comprising 21 different breeds, yielding high training and validation accuracies around 98.5%, with slightly lower testing accuracies of 83.68% for VGG16 and 84.07% for VGG19. Fine-tuning pre-existing models proved efficient, handling the challenge of limited data availability and the complexity of breed differentiation. VGG19 slightly outperformed VGG16 in testing accuracy, though VGG16 had a higher recall rate for some breeds. The paper also examines prediction speeds, with both models predicting new images in approximately four seconds, highlighting the practicality of deploying these models in real-time applications. [3]

The research paper, "*PaddlePaddle: An Open-Source Deep Learning Platform from Industrial Practice*" authored by Y J Ma, D H Yu, T Wu, was published in 2021. It provides a comprehensive overview of PaddlePaddle's capabilities as a fully functioning open-source deep learning platform. Special attention is given to the platform's proficiency in the field of Optical Character Recognition (OCR). The paper delves into the technical mechanisms by which PaddlePaddle supports complex OCR tasks, such as scene text recognition and character segmentation. This is facilitated by PaddlePaddle's sophisticated neural network architectures and its robust distributed training system, which is adept at handling large-scale data processing. [4]

## III. Problem Formulation

In this paper, we propose to employ a model that will do the following:

- It will receive an input image and transform it into a standard format using object detection and skewing and warping techniques.
- It will then classify it using VGG 16, which is a Convolutional Neural Network (CNN), into one of the four document categories:
  - Receipts
  - Technical Papers
  - Newspapers
  - Book Covers

- Depending on the category of the document, important textual information related to a category will be extracted with the use of Optical Character Recognition (OCR):
  - For receipts: Individual Item price, Total expenses and date of transaction
  - For technical papers: Title, Author and Conference Title
  - For Newspapers: Name of the newspaper and Date
  - For Book Covers: Title and Author

## IV. Technical Approach

There are three parts to the approach in this paper: (a) Image Processing and Object Detection to remove noise from the input image and detect the area of interest; (b) Image Classification to determine the category of document the image fits best in; and, (c) Optical Character Recognition (OCR) to extract important information depending on the category of the document.

### *A. Image processing and Object Detection*

1. Edge Detection

Given an image of a document, we first apply edge detection to locate the document's boundaries. This step employs a custom model built on the YOLOv8 framework, known for its real-time object detection efficiency. The model is fine-tuned on a labeled dataset comprising various document images with annotated edges. It outputs bounding box coordinates ($x_i$, $y_i$), which are then used to identify the document's contour in the image space.

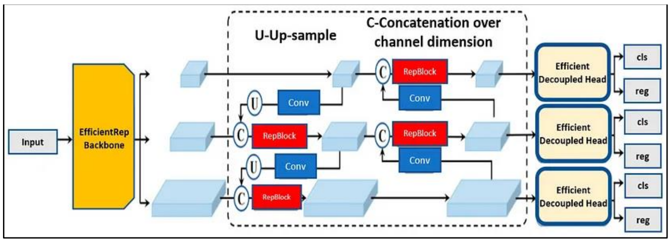


***Fig. 4.1:*** *YOLOv8 Architecture*

2. Geometric Transformation

With the detected edges, the image is subjected to a series of geometric transformations to correct for perspective and skew. We utilize Hough Transform techniques [5], which involve the following steps:

i. Grayscale Conversion: The RGB image is converted to grayscale to reduce computational complexity using cv2.COLOR_BGR2GRAY

ii. Canny Edge Detection: The Canny edge detector is applied to the grayscale image to obtain binary edge-maps.

iii. Line Detection: Using the Probabilistic Hough Transform, the line segments $L_i$ in the edge-maps are detected.

iv. Intersection Computation: The intersections $P_i$ of the detected line segments are calculated to find the document's corner points.

v. Perspective Warping: A perspective warp is applied using the corner points to transform the document into a standardized view:

$$getPerspectiveTransform(P_{source} , P_{target})$$

where $P_{source}$ are the corner points in the original image, and $P_{target}$ are the corner points of the desired standardized image dimension.

vi. De-skewing: If necessary, further affine transformations are implemented to correct for any remaining skew in the document's text lines.

### *B. Image Classification*

Image classification using Convolutional Neural Networks (CNN) is a very efficient technique due to their ability to learn and extract intricate features from raw image data automatically. This research employed the VGG-16 model for the task of image classification.

The VGG-16 model is a convolutional neural network (CNN) architecture proposed by the Visual Geometry Group (VGG) at the University of Oxford. It is characterized by its depth, consisting of 16 layers, including 13 convolutional layers and 3 fully connected layers. VGG-16 is renowned for its simplicity and effectiveness, and its ability to achieve strong performance on computer vision tasks, including image classification and object recognition. The model's architecture features a stack of convolutional layers followed by max-pooling layers, with progressively increasing depth.

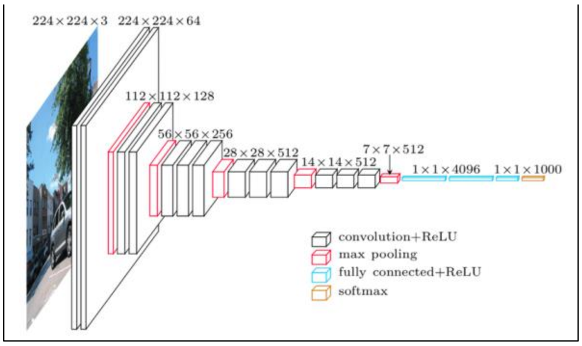


***Fig. 4.2:*** *VGG Architecture*

### *C. Optical Character Recognition*

Following the classification of documents into distinct categories our system leverages Optical Character Recognition (OCR) to extract textual content from the identified documents. This crucial stage employs the open-source PaddleOCR framework, renowned for its high accuracy and efficiency in text detection and recognition.

PaddleOCR provides a robust solution for detecting text within diverse document formats. It outputs the textual content along with bounding boxes and coordinates, which delineate the text's location on the document. This spatial data is essential for the subsequent information extraction process.

Upon the OCR process's completion, the system retrieves bounding boxes containing the coordinates of the detected text regions. These coordinates are instrumental in isolating text blocks specific to each document type, enabling targeted text extraction based on the document's classification.

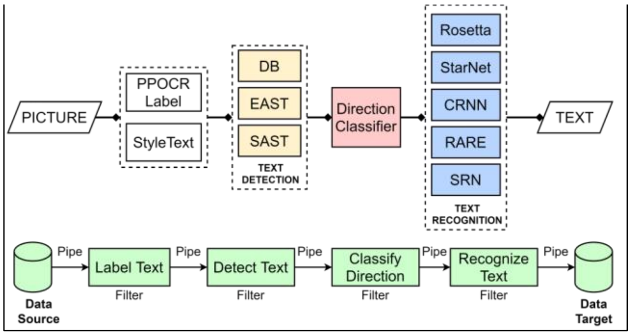


***Fig. 4.3:*** *PaddleOCR Working Procedure*

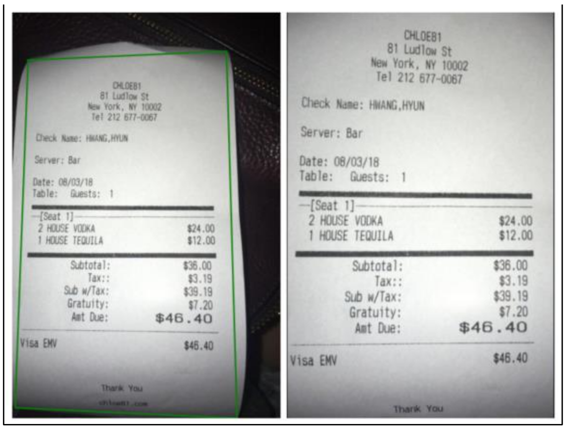
Utilizing the document class information obtained from the initial classification phase, our system applies custom regular expression (regex) patterns, and positional data to parse and extract relevant information from the text. These patterns are tailored to the specifics of each document type:

- Books: The focus is on extracting significant identifiers such as the book's title and author from the cover, which is often prominently displayed and follows predictable formatting.
- Newspapers: The extraction targets the newspaper's name and date, which is easily identified thanks to the standardized layout typically used in newspaper headers.
- Receipts: The system parses dates, individual prices, and the total amount using maximum of all prices detected, the location of the text (left for type of product, right for amount) and regex.
- Technical Papers: In the case of technical papers, the extraction targets are the title, author(s), and the conference or publication title. These details are also extracted using a combination of the text's position and regex patterns.

The OCR module functions seamlessly within the broader system architecture. Post-document classification, the identified document type (receipt, newspaper, book) dictates the specific regex pattern applied. This methodical approach ensures that the extraction process is both accurate and relevant to the document's content. The integration of PaddleOCR not only enhances the accuracy of text detection but also contributes significantly to the overall effectiveness of the information extraction process.

## RESULTS

Figure 5.1 shows the original and processed images of a receipt. The left image displays the receipt in its original environment with background elements, and the right image showcases the result after applying a Hough transform for perspective correction and noise reduction to isolate the receipt for enhanced clarity and OCR performance.



CHLOE81
81 Ludlow St
New York, NY 10002
Tel 212 677-0067

Check Name: HWANG,HYUN

Server: Bar

Date: 08/03/18
Table: Guests: 1

| | |
|---|---|
| -[Seat 1]- | |
| 2 HOUSE VODKA | $24.00 |
| 1 HOUSE TEQUILA | $12.00 |
| Subtotal: | $36.00 |
| Tax:: | $3.19 |
| Sub w/Tax: | $39.19 |
| Gratuity: | $7.20 |
| Amt Due: | $46.40 |
| Visa EMV | $46.40 |

Thank You

***Fig. 5.1:*** *Edge detection and warping*

For image classification using VGG, we were able to achieve a training accuracy of 88.18% and a validation accuracy of 75.10%. Figure 5.2 shows the training and accuracy curves.

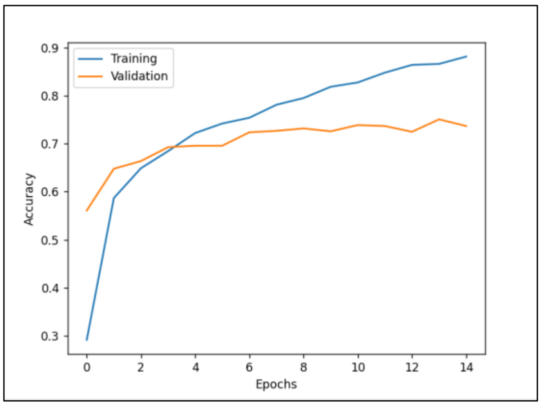
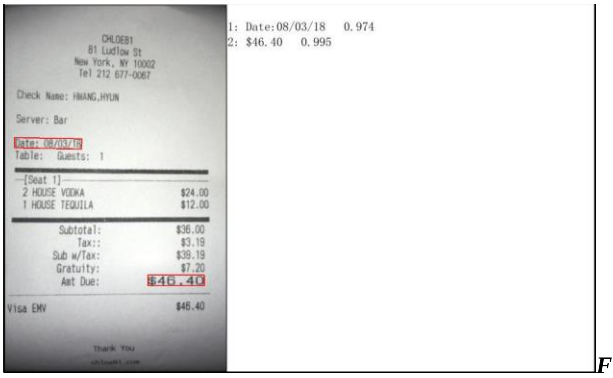


***Fig. 5.2:*** *VGG training accuracy and validation accuracy vs. number of epochs*

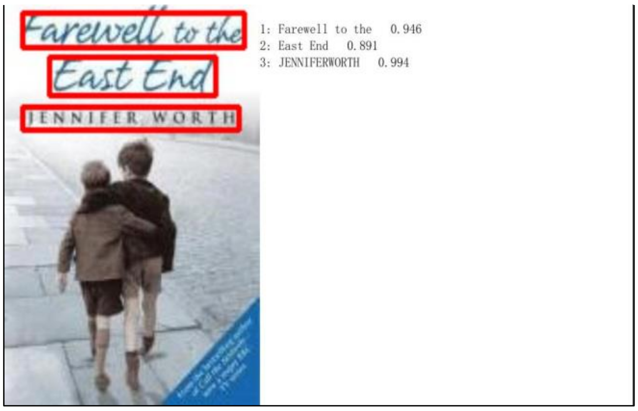
Once the image is correctly classified, the image is sent to OCR for extracting key information. Figure 5.3 illustrates the output of an OCR algorithm with key information highlighted: the date of transaction and the total amount due are enclosed in red boxes, with the extracted text and corresponding confidence scores displayed on the right.



```
1: Date:08/03/18    0.974
2: $46.40    0.995
```

***Fig. 5.3:*** *Output of PaddleOCR a receipt, highlighting the date and total amount*

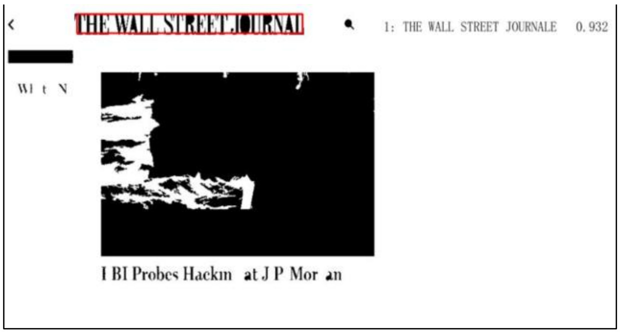
Figure 5.4 shows extracted text from a book cover using OCR. The image shows the book title and the author's name, both highlighted in red boxes to indicate the OCR's ability to recognize and digitize textual content from the cover with respective confidence scores displayed adjacent to the identified text.



```
1: Farewell to the    0.946
2: East End    0.891
3: JENNIFERWORTH    0.994
```

**Fig. 5.4:** *Output of PaddleOCR a book cover, highlighting the title and the author*

Figure 5.5 demonstrates the preprocessing and OCR steps taken to identify and highlight the newspaper title 'THE WALL STREET JOURNAL'. Initial preprocessing involved applying contour filtering based on the average height to enhance OCR recognition accuracy, as indicated by the confidence score



```
1: THE WALL STREET JOURNALE    0.932
```

**Fig. 5.5:** *Output of PaddleOCR a newspaper, highlighting the title*

## Conclusion and Future Scope

The intelligent document processing system developed in this paper had an accuracy of 75% for the image classification task. It also extracts relevant information with high efficiency and accuracy. This model excels in handling complex layouts and mixed scripts by focusing on specific areas of the text during the recognition process, thus enhancing the precision of extracted information.

In the future, the following enhancements can be made to this system:

- Adding an attention layer in both the classification and the OCR architectures have great potential to improve the accuracy of the system.
- This research focused primarily on pro-frontal images. Implementing more complex methods for skewing and warping can help increase the scope of the type of images captured.

## References

[1] M. Kumar and R. Bhatt, "An approach towards Real-Time Object Detector Using Open CV," 2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART), pp. 981-985.

[2] R. Mittal and A. Garg, "Text extraction using OCR: A Systematic Review," 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2020, pp. 357-362.

[3] Mahardi, I. -H. Wang, K. -C. Lee and S. -L. Chang, "Images Classification of Dogs and Cats using Fine-Tuned VGG Models," 2020 IEEE Eurasia Conference on IOT, Communication and Engineering (ECICE), Yunlin, Taiwan, 2020, pp. 230-233.

[4] Y J Ma, D H Yu, T Wu et al., "PaddlePaddle: An Open-Source Deep Learning Platform from Industrial Practice [J]", Frontiers of Data, vol. 1, no. 1, pp. 105-115, 2019.

[5] P. Ganesan and G. Sajiv, "A comprehensive study of edge detection for image processing applications," 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), Coimbatore, India, 2017, pp. 1-6.