INFORMATION RETRIEVAL FROM BUSINESS RECEIPTS USING DEEP LEARNING

Parit Kansal, Shubh Gupta, Pankhuri Gupta

Department of Computer Science and Engineering

Harcourt Butler Technical University, Kanpur, Uttar Pradesh, India

ABSTRACT—This research presents a deep learning-based framework for intelligent document processing, combining Mask R-CNN for receipt segmentation and the Donut model for structured text extraction—eliminating the need for traditional OCR. A key contribution of this work is the generation of a synthetic dataset that simulates real-world document variations by overlaying business receipts onto diverse backgrounds with random transformations. This synthetic data enhances model robustness and generalization. The proposed system segments receipts from complex images and directly extracts essential information, such as company name, address, date, and total amount, in a structured format. By unifying segmentation and text extraction into an end-to-end pipeline, this approach offers a scalable, adaptable solution for automated financial document processing across various enterprise environments.

I. INTRODUCTION

Traditional document processing methods rely on OCR-based approaches, which often struggle with structured and semi-structured documents like invoices and receipts. This study addresses these challenges by leveraging deep learning techniques for end-to-end document understanding. Mask R-CNN, an advanced object detection and segmentation model, is used to extract receipts from images, while the Donut model—a Transformer-based document understanding network—automatically extracts structured text without OCR dependencies.

This research contributes to the field of document intelligence by integrating object segmentation and structured text extraction into a unified framework. The proposed system is evaluated on a dataset of business receipts, demonstrating its potential in financial automation, invoice management, and intelligent document processing.

Additionally, the modular design of the system allows easy adaptability to other document types, making it scalable and robust for broader real-world applications. This flexibility enhances the system's usability across diverse domains and reduces the need for task-specific re-engineering, ultimately leading to faster deployment and improved efficiency in enterprise workflows.

II. METHODS AND TECHNOLOGIES

The proposed methodology consists of four key stages: **dataset creation, preprocessing, segmentation using Mask R-CNN, and text extraction using the Donut model**. Each stage is carefully designed to ensure accurate receipt extraction and structured text retrieval.

A. Dataset Creation

To enhance the robustness of the model, a **custom dataset** was created, incorporating real-world variations in document images.

- **Synthetic Data Generation:** Business receipt images were cropped and randomly placed on diverse backgrounds (solid colors, patterned backgrounds, or real-world environments). Random rotations (±5°) were applied to simulate realistic document placement. The dataset For each synthetic image, a random number of business receipt images (ranging from 1 to 10) were selected from the CORD dataset, cropped, and randomly placed on diverse backgrounds such as plain white, solid colors, or real-world environments. To simulate realistic document placement, random rotations within ±5° were

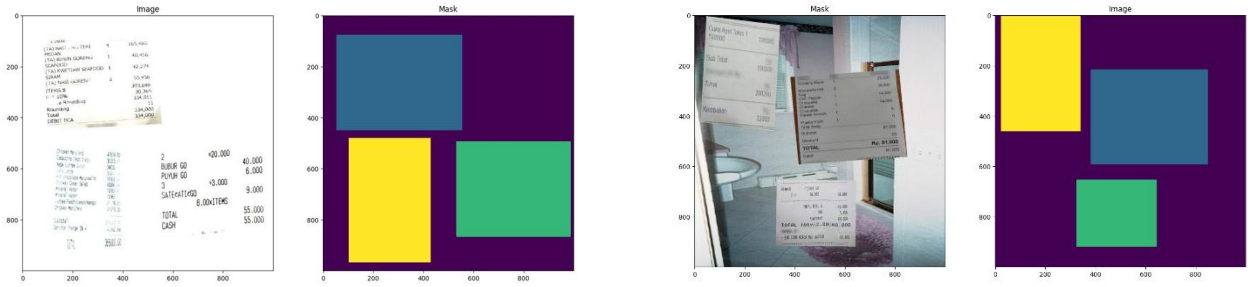

Fig. 1 Examples of synthetic receipt images with varying backgrounds and Mask

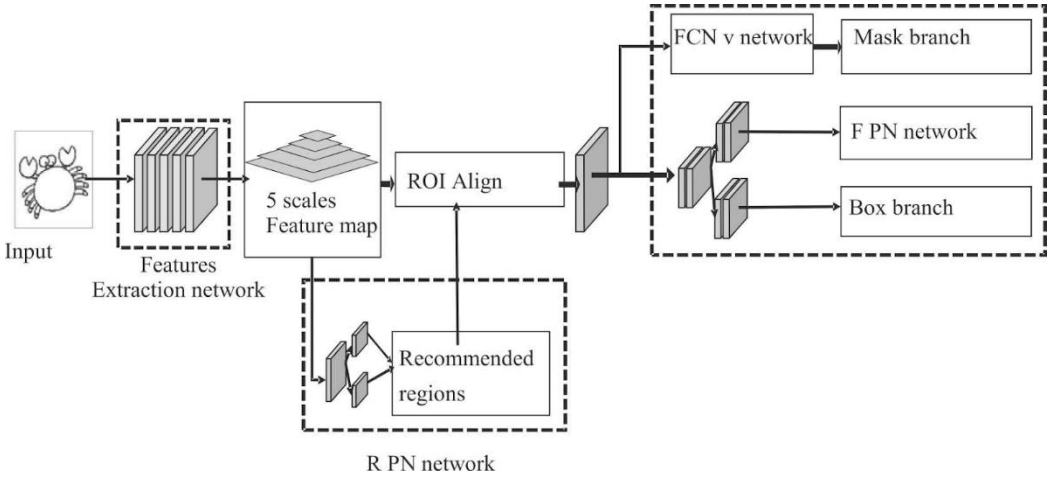

Fig. 2 Mask R-CNN Architecture

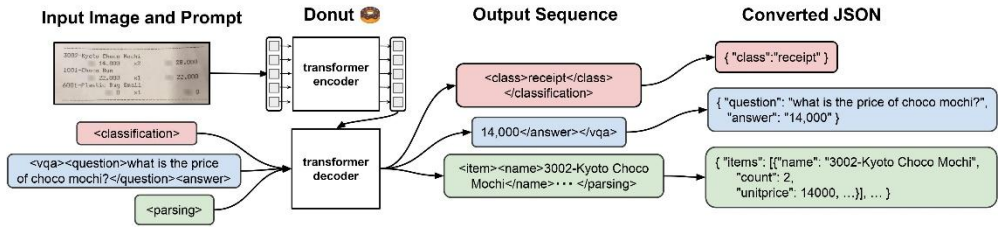

Fig. 3 Donut Model Architecture

applied. For each receipt, a corresponding mask and bounding box were generated. In the mask, background pixels are assigned a value of 0, while individual receipts are assigned unique pixel values incrementally starting from 1 (e.g., 1, 2, 3, ...). Instead of pixel-level segmentation, each receipt's mask is represented as a rectangular region to account for possible tilts caused by rotation. The dataset was further augmented to enhance model generalization. Fig. 1 illustrates examples of the synthesized images along with their corresponding masks and bounding boxes.

- **CORD Dataset Utilization:** The CORD dataset, consisting of thousands of annotated Indonesian receipts, was used to train the Mask R-CNN model. Annotations include bounding boxes for receipts and text-based metadata.

B. Receipt Segmentation Using Mask R-CNN

The first stage of the model focuses on detecting and segmenting receipts from cluttered backgrounds. Mask R-CNN was selected for its capability to perform object detection and pixel-wise segmentation simultaneously. The model was trained on the CORD dataset. Evaluation metrics include mean Average Precision (mAP) and mean Average Recall (mAR), computed based on Intersection over Union (IoU) thresholds.

Fig. 4 End-to-End Pipeline Diagram

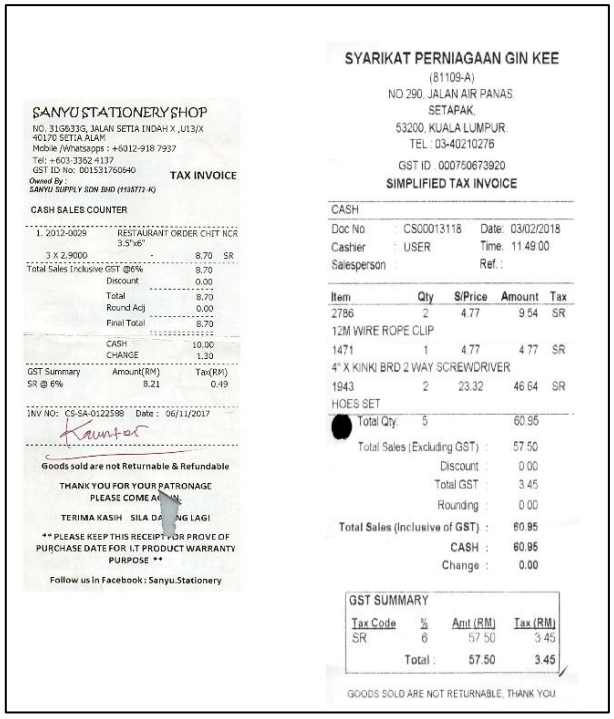

Fig. 5 Input image to the trained Mask R-CNN model

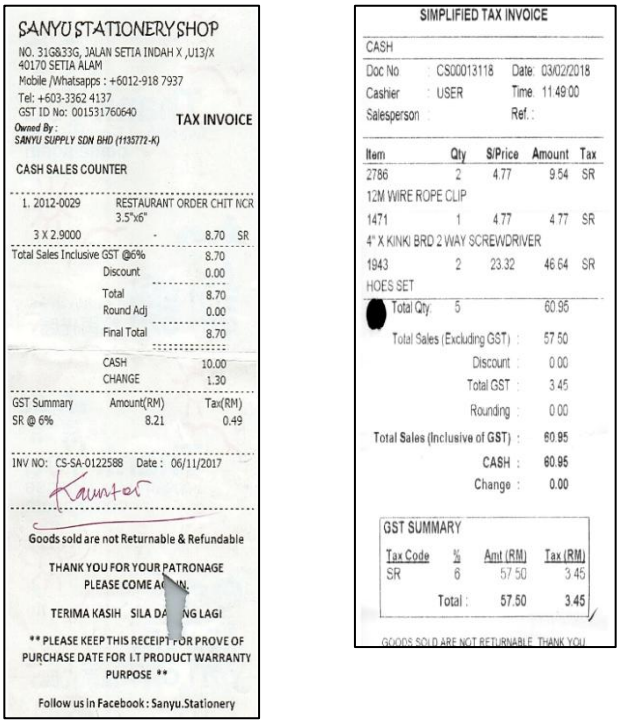

Fig. 6 Extracted Business Receipts by Mask R-CNN

Mask R-CNN was also chosen for its robustness in handling rotated or tilted images, making it suitable for future extensions involving non-axis-aligned document segmentation.

C. Structured Text Extraction Using Donut Model

Once the receipts are segmented, the Donut model is employed to extract structured information such as the company name, address, date, and total amount. Unlike traditional OCR-based approaches that struggle with complex key-value text pairs, Donut is an end-to-end Transformer-based model that directly converts visual document inputs into structured text. It utilizes a Vision Transformer (ViT) encoder to extract visual features from document images and a textual decoder to generate structured outputs in JSON format. By learning to map image features directly to structured text, Donut eliminates the dependency on OCR pipelines.

The model is trained on the SROIE dataset available on Kaggle. Evaluation is based on text extraction accuracy, which measures the similarity between the extracted information and the ground truth annotations.

D. Integrated Pipeline for Automated Document Processing

To achieve a fully automated document processing system, the Mask R-CNN and Donut models were integrated into a single workflow:

- **Step 1**: Mask R-CNN detects and segments receipts from complex images.
- **Step 2**: The segmented receipt is passed to the Donut model for text extraction.
- **Step 3**: The final output is a structured JSON representation of the extracted data.

	Address	Company	Date	Total
Cropped Image 1	'NO. 31G&33G, JALAN SETIAINDAH X ,U13/X 40170 SETIA ALAM'	'SANYU STATIONERY SHOP'	'06/11/2017'	'8.70'
Cropped Image 2	'CASHER USER TMN 1149.00 SELANGOR'	'IMPLIFIED TAX INVOICE'	'03/02/2018'	'60.95'

Table 1 Information extracted from cropped images obtained as the final output of the pipeline

III. RESULTS

A. Mask R-CNN Performance

The synthesized dataset is passed to the Mask R-CNN model, as shown in Fig. 5, to extract individual receipts from the input image (see Fig. 6). These segmented receipts are then passed to the next phase of the pipeline.

The performance of the trained Mask R-CNN model is evaluated using the Intersection over Union (IoU) metric:

- Average Precision @ IoU = 0.50:0.95 | area = all | maxDets = 100: 0.971
- Average Precision @ IoU = 0.50 | area = all | maxDets = 100: 0.989
- Average Precision @ IoU = 0.75 | area = all | maxDets = 100: 0.989
- Average Recall @ IoU = 0.50:0.95 | area = all | maxDets = 100: 0.980
- Average Recall @ IoU = 0.50 | area = all | maxDets = 100: 0.989
- Average Recall @ IoU = 0.75 | area = all |
- maxDets = 100: 0.989

B. Donut Model Performance

The output images from the Mask R-CNN model are forwarded to the Donut model, which extracts text and meaningful information from them. It successfully retrieves details such as company name, address, date, and total amount from the images. The final output example is depicted in Table. 1.

The structured text extraction accuracy achieved by the Donut model is 81%

IV. CONCLUSION AND DISCUSSION

This study presents a deep learning-based approach for intelligent document processing by integrating Mask R-CNN for receipt segmentation and the Donut model for structured text extraction. Unlike traditional OCR-based systems, this pipeline directly extracts meaningful information in a structured format, thereby enhancing efficiency in financial and business applications.

The experimental results demonstrate that Mask R-CNN effectively isolates receipts from complex backgrounds, while the Donut model accurately extracts key details such as company name, address, date, and total amount. Together, they achieve high accuracy in structured document understanding, showcasing the potential of this integrated approach in real-world automation workflows.

V. ACKNOWLEDGMENT

We would like to express our heartfelt gratitude to Dr. Vivek Singh Verma, Associate Professor, Department of Computer Science and Engineering, Harcourt Butler Technical University, Kanpur, for his invaluable guidance, continuous encouragement, and unwavering support throughout the duration of this research. His vast knowledge, expert supervision, and insightful suggestions played a crucial role in shaping the direction of our work and helped us overcome numerous challenges.

Dr. Verma's dedication to academic excellence and research, along with his patience and motivation, inspired us at every stage of the project. His detailed feedback and thoughtful advice significantly contributed to the successful completion of this study.

VI. REFERENCES

[1] S. Ren, K. He, R. Girshick, and J. Sun, ``Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

https://arxiv.org/abs/1506.01497

[2] K. He, G. Gkioxari, P. Dollár, and R. Girshick, ``Mask R-CNN," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017. https://arxiv.org/abs/1703.06870

[3] G. Kim, J. Kim, J. Ha, and G. Kim, ``OCR-Free Document Understanding Transformer," in *Advances in Neural Information Processing Systems (NeurIPS)*}, 2021. https://arxiv.org/abs/2111.15664

[4} J. Baek, B. Lee, D. Han, S. Yun, and H. Lee, ``Character Region Awareness for Text Detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*}, 2019. https://arxiv.org/abs/1904.01941

[5] S. Park, S. Shin, B. Lee, Y. Kim, and S. Lee, ``CORD: A Consolidated Receipt Dataset for Post-OCR Parsing," NAVER AI Lab, 2019. https://github.com/clovaai/cord

[6] Matterport Inc., ``Mask R-CNN for Object Detection and Instance Segmentation on Keras and TensorFlow," 2018. https://github.com/matterport/Mask_RCNN

[7] Hugging Face, ``Donut Model Documentation," https://huggingface.co/docs/transformers/model_doc/donut