DOCUMENT ANALYSIS USING A DEEP LEARNING APPROACH

Shireesha K B
Department of Master of Computer Application
RV College Of Engineering Bangalore
shireeshakb.mca21@rvce.edu.in

Thejas P
Department of Master of Computer Application
RV College Of Engineering Bangalore
thejasp.mca21@rvce.edu.in

Dr. S Anupama Kumar
Associate Professor
Department of Master of Computer Application
RV College Of Engineering Bangalore
anupamakumar@rvce.edu.in

Abstract: Document layout analysis is an automated process that recognizes and extracts the content and structure of various documents, including books, newspapers, and forms, using computer vision and machine learning techniques. Its main purpose is to enable accurate and efficient processing of vast amounts of digital documents in fields like digital preservation, document management, and information retrieval. Document layout analysis is widely applicable and can help with tasks such as data extraction, document classification, and information retrieval, which makes it an essential component of many modern information processing systems. This implements the layout parser and CNN architecture. The layout parser is a Python library wherein the library contains all the algorithms like R-CNN, CNN, and Faster-CNN. CNN is used to classify different components of the document based on their position in the layout, Spacy is used to extract the document and perform the natural processing tasks on the extracted text. The result is a structured representation of the layout and contents of a document that can be used for better document understanding, and information retrieval.

Keywords: ***Page Segmentation, Natural Language Processing, CNN, Layout Parser.***

I. Introduction

The process of document layout analysis automatically detects and comprehends the arrangement and layout of documents. It involves examining the visual elements and attributes of a document, such as images, text, and tables, to identify the distinct sections and elements of the document. The objective of document layout analysis is to construct an organized representation of the document that can be utilized for several purposes, such as summarizing the document, categorizing it, and retrieving specific information from it. Document layout analysis is a primary step in document processing and analysis workflows, and it can be approached through different methods. Rule-based methods rely on predefined rules and heuristics to identify document components based on their visual features. Template-based methods use pre-established templates to recognize the layout and structure of a specific document. Machine learning-based methods employ algorithms that learn from examples to detect a document's layout and structure. This requires training a model on a sizeable annotated dataset and utilizing it to predict the layout and structure of new documents. All in all, document layout analysis is a vital aspect of document analysis and processing systems, and it plays a vital role in enhancing the accessibility and usability of information.

II. Literature Survey

A deep learning method for automatically analyzing the layout of Arabic documents is presented in the publication [1]. The suggested method demonstrates its efficacy for Arabic document layout analysis by achieving high accuracy on several datasets. [1] This describes a method for segmenting newspaper elements using the deep learning architecture Mask R-CNN. The promise of automating newspaper layout analysis is shown by the suggested method's excellent accuracy in segmenting text, graphics, and advertisements on newspaper pages. [2]. Table detection method using You Only Look Once (YOLO) architecture. The proposed method outperformed state-of-the-art techniques in terms of accuracy and processing speed when tested using benchmark datasets that were made available to the general public [3]. This article thoroughly investigates the identification of compress layout

equivalence with deep learning by looking at several document picture datasets. The suggested approach provides high accuracy, outperforms current approaches in terms of detection speed and scalability, and is thus appropriate for use in practical applications [4]. Due to the complexity of the text's fonts, distribution, or background, scene text recognition is a difficult topic of computer science research. Traditional OCR technology is therefore inadequate for the job. Due to its use in fields like information retrieval, augmented reality, and autonomous driving, it has become a popular research topic [5]. An incremental improvement of the YOLO (You Only Look Once) object detection algorithm, that achieves state-of-the-art accuracy and speed in object detection tasks by using a large-scale object detection dataset and implementing various improvements in architecture and training techniques [6]. The ICDAR 2013 Table Competition aimed to compare and evaluate the performance of different table recognition algorithms. Participants were provided with a dataset of documents containing tables and were tasked with detecting, extracting, and recognizing the tables [7]. Results from the ICDAR2017 page object identification competition, which sought to locate page object bounding boxes in diverse document pictures. The competition used a dataset of 979 document images and compared various object detection methods based on accuracy and efficiency metrics [8]. This suggests employing bounding boxes of related components as a mechanism for recursive XY cut in document analysis and recognition. Multiple horizontal and vertical lines in the document image can be found and eliminated using the suggested algorithm. Results from the method's evaluation on numerous documents are encouraging [9]. This proposes a method for recognizing table structures using robust block segmentation. It uses a heuristic for detecting the presence of a table and then segments the blocks in the document based on various layout features. The method is tested on a variety of documents and achieves high accuracy [10]. The approach involves detecting cells and their spatial relationships to generate a graph representation, which is then used to classify tables. The proposed method achieves promising results on benchmark datasets [11]. The ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) in 2014 saw the best performance from a deep convolutional neural network for large-scale image recognition known as VGG. Convolutional layers with tiny 3x3 filters and a fully connected layer at the end make up the VGG network design [12]. This article introduces Faster R-CNN, a real-time object identification technique that combines a Fast R-CNN detector with a region proposal network (RPN). The Fast R-CNN detector classifies the proposals that the RPN efficiently creates after scanning the image. On the PASCAL VOC and MS COCO datasets, the suggested method produces results that are state-of-the-art [13]. This introduces a unique method for semantic segmentation tasks known as Fully Convolutional Networks (FCNs). FCNs can accept inputs of any size and execute end-to-end training, yielding outputs of the same size as the input. The method provides cutting-edge outcomes on numerous benchmarks and permits real-time inference on video streams [14]. A deep residual learning framework that allows for training extremely deep neural networks with more than 150 layers [15]. Utilizing Discrete Wavelet Transform (DWT) feature extraction and Neural Network classification, a method for online handwritten signature verification. The system successfully detects fake signatures with high accuracy and a low mistake rate, making it a promising option for safe authentication and fraud detection in online transactions. [16]. A brand-new technique based on region-bounding boxes for page segmentation and region classification. Using text region detection and grouping, the technique creates regions of interest for segmentation. [17]. Examines page segmentation methods used in document analysis, such as table detection, text line extraction, and geometric and logical layout analysis. Additionally, the author discusses difficulties and potential directions for page segmentation research. [18]. A comprehensive review of document image analysis (DIA), covering image acquisition, Preprocessing, segmentation, feature extraction, and classification. It discusses challenges such as variations in text and background, noise, skew, and distortion, and outlines possible solutions. The paper also provides a bibliography of important works in the field [19]. An overview of the ICDAR 2013, 2017, and 2019 competitions on table recognition in document images. It covers the datasets, evaluation protocols, and various approaches and models proposed by the participants. The paper highlights the challenges in table recognition and provides insights for future research [20].

III. Proposed Methodology

Convolutional Neural Network (CNN)- CNN is a deep learning technique that is used for automated vision and recognition of image jobs. CNNs are composed of numerous layers of neurons that learn to recognize features at various levels of

abstraction and are inspired by the way the human visual system is organized. Convolutional layers, pooling layers, and fully linked layers make up the fundamental components of a CNN. Convolutional layers apply a set of filters to the input image to extract features, which are then passed through pooling layers that reduce the dimensionality of the feature maps. After being flattened, the generated feature maps are then used to categorize the input picture using fully linked layers. CNNs have the benefit of being able to automatically learn features from data without the requirement for manual feature engineering. This makes them especially helpful for picture identification applications, where the required characteristics may be intricate and challenging to describe manually. Numerous applications, including object recognition, face recognition, and self-driving cars, among others, have successfully used CNNs. Additionally, they are employed in NLP tasks like sentiment analysis and translation.

CNN architecture: Convolutional Neural Networks (CNNs) are neural networks created specifically to process pictures. In the field of computer vision, they are often employed.

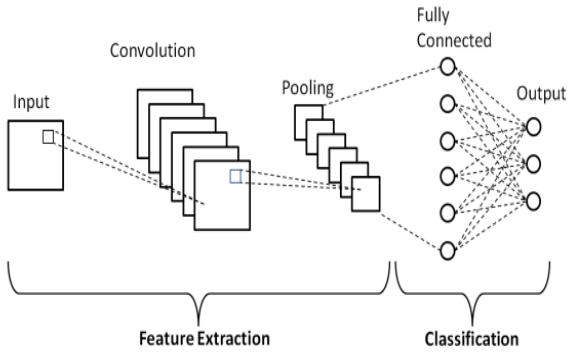


Figure: Adapted from [20]

The work has been implemented in five modules:

The input module receives the input image and passes it to the layout detection component. In this code, the input image is loaded from a file path using the OpenCV package.

Pre-processing module takes the input image and is read using OpenCV, and its color channels are reversed. The Detectron2LayoutModel object receives the picture and applies a pre-trained object identification model to identify layout components including text, titles, figures, tables, and lists. Additionally, the model excludes components with low confidence scores.

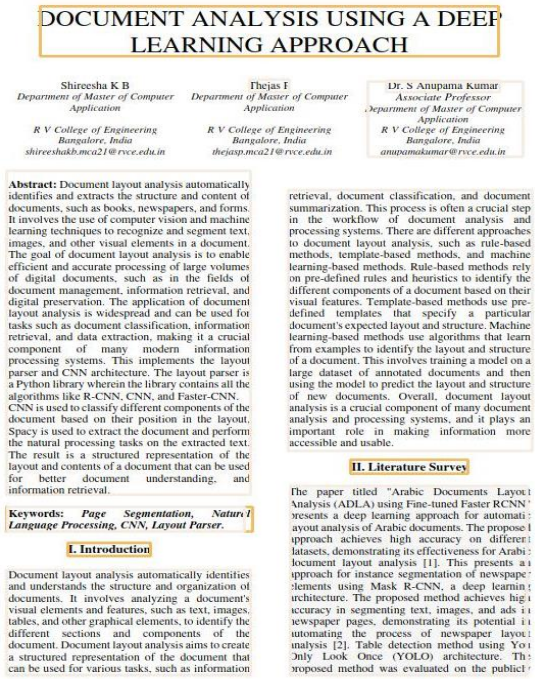
The layout analysis module gives the output of the Detectron2LayoutModel is a list of bounding boxes that represent the detected layout elements. These bounding boxes are then processed to group similar elements together. For example, the script groups all text blocks that are in the left half of the document, and all text blocks that are in the right half of the document.

Text recognition: The TesseractAgent object from the layout parser is used to perform OCR on each text block. The OCR results are then added to the corresponding text block.

Named entity recognition: The script uses the Spacy library and the en_core_web_trf model to perform named entity recognition on the text of the document. The named entities are then printed to the console.

IV. Results and Conclusion

Document layout analysis is a crucial process in modern document management systems that involves the identification and classification of different document elements such as text, images, tables, and figures. The process is necessary for tasks such as optical character recognition (OCR), information retrieval, and content analysis. A well-designed document layout analysis system must be accurate, efficient, and scalable to handle large volumes of documents. It must also meet external interface requirements such as compatibility with different file formats,


DOCUMENT ANALYSIS USING A DEEP LEARNING APPROACH

Shireesha K B
Department of Master of Computer Application
R V College of Engineering
Bangalore, India
shireeshakb.mca21@rvce.edu.in

Thejas I
Department of Master of Computer Application
R V College of Engineering
Bangalore, India
thejasp.mca21@rvce.edu.in

Dr. S Anupama Kumar
Associate Professor
Department of Master of Computer Application
R V College of Engineering
Bangalore, India
anupamakumar@rvce.edu.in

Abstract: Document layout analysis automatically identifies and extracts the structure and content of documents, such as books, newspapers, and forms. It involves the use of computer vision and machine learning techniques to recognize and segment text, images, and other visual elements in a document. The goal of document layout analysis is to enable efficient and accurate processing of large volumes of digital documents, such as in the fields of document management, information retrieval, and digital preservation. The application of document layout analysis is widespread and can be used for tasks such as document classification, information retrieval, and data extraction, making it a crucial component of many modern information processing systems. This implements the layout parser and CNN architecture. The layout parser is a Python library wherein the library contains all the algorithms like R-CNN, CNN, and Faster-CNN. CNN is used to classify different components of the document based on their position in the layout. Spacy is used to extract the document and perform the natural processing tasks on the extracted text. The result is a structured representation of the layout and contents of a document that can be used for better document understanding, and information retrieval.

Keywords: *Page Segmentation, Natural Language Processing, CNN, Layout Parser.*

I. Introduction

Document layout analysis automatically identifies and understands the structure and organization of documents. It involves analyzing a document's visual elements and features, such as text, images, tables, and other graphical elements, to identify the different sections and components of the document. Document layout analysis aims to create a structured representation of the document that can be used for various tasks, such as information retrieval, document classification, and document summarization. This process is often a crucial step in the workflow of document analysis and processing systems. There are different approaches to document layout analysis, such as rule-based methods, template-based methods, and machine learning-based methods. Rule-based methods rely on pre-defined rules and heuristics to identify the different components of a document based on their visual features. Template-based methods use pre-defined templates that specify a particular document's expected layout and structure. Machine learning-based methods use algorithms that learn from examples to identify the layout and structure of a document. This involves training a model on a large dataset of annotated documents and then using the model to predict the layout and structure of new documents. Overall, document layout analysis is a crucial component of many document analysis and processing systems, and it plays an important role in making information more accessible and usable.

II. Literature Survey

The paper titled "Arabic Documents Layout Analysis (ADLA) using Fine-tuned Faster RCNN" presents a deep learning approach for automatic layout analysis of Arabic documents. The proposed approach achieves high accuracy on different datasets, demonstrating its effectiveness for Arabic document layout analysis [1]. This presents an approach for instance segmentation of newspaper elements using Mask R-CNN, a deep learning architecture. The proposed method achieves high accuracy in segmenting text, images, and ads in newspaper pages, demonstrating its potential in automating the process of newspaper layout analysis [2]. Table detection method using You Only Look Once (YOLO) architecture. The proposed method was evaluated on the publicly

integration with other systems, and ease of use. Moreover, it must comply with design constraints such as technical, environmental, regulatory, and organizational limitations.

Figure 4.1 - Output of high-resolution image

The above figure is the output of a high-resolution image. The image shows all the highlighted elements like the image, title, body, and table. It has been highlighted in different colors for better understanding.

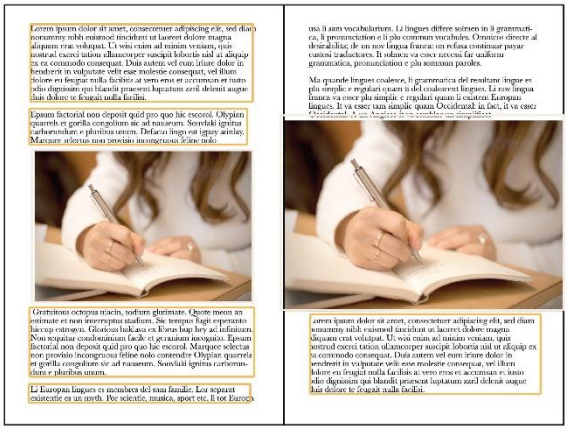


Figure 4.2 – Image with fewer elements.

The above figure is the output of an image where fewer elements are present. Here the headings, paragraphs, images, and subtitles of a document are identified.

V. Future Enhancements

The document layout analysis has the potential for further development to enhance its capabilities and provide additional benefits to users. Some possible future enhancements are discussed below: Improved OCR accuracy: One area for improvement would be to enhance the accuracy of the OCR system. This could involve using more advanced OCR techniques, such as deep learning-based approaches, or incorporating additional pre-processing steps to clean up the image before OCR. Support for additional document types: While the current system is designed to work with academic papers, it could be expanded to support other types of documents as well. For example, the system could be trained to recognize the layout of legal documents, financial reports, or medical records. Multilingual support: Currently, the system is designed to work with English language documents. However, it could be expanded to support other languages as well. This would involve training the OCR and layout analysis models on datasets in other languages, as well as incorporating language-specific processing steps. Integration with other AI technologies: Finally, the system could be enhanced by integrating it with other AI technologies, such as natural language processing or sentiment analysis. This would enable the system to extract more information from documents, such as identifying key concepts or detecting emotional tone**.**

VI. References

[1] Arabic Documents Layout Analysis (ADLA) using Fine-tuned Faster RCN Latifa Aljiffry;Hassanin Al-Barhamtoshy;Amani Jamal;Felwa Abukhodair 2022 20th International Conference on Language Engineering (ESOLEC)

[2] Instance Segmentation of Newspaper Elements Using Mask R-CNN Abdullah Almutairi;Meshal Almashan 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)

[3]A YOLO-Based Table Detection Method Yilun Huang;Qinqin Yan;Yibo Li;Yifan Chen;Xiong Wang;Liangcai Gao;Zhi Tang 2019 International Conference on Document Analysis and Recognition (ICDAR)

[4] Deep learning based Layout Equivalence Detection in Compressed Domain and analysis of different Document Image Dataset:Systematic Study Kavita V. Horadi;Jagadeesh D. Pujari;Narasimha Prasad Bhat 2022 2nd International Conference on Intelligent Technologies (CONIT)

[5] Yuxin Chen & Yunxue Shao 2019 IEEE 11th International Conference on Communication Software and Network

[6] J. Redmon and A. Farhadi, “Yolov3: An incremental improvement,” arXiv preprint arXiv:1804.02767, 2018.

[7] M. Gobel, T. Hassan, E. Oro, and G. Orsi, “Icdar 2013 table com- ¨ petition,” in Document Analysis and Recognition (ICDAR), 2013 12th International Conference on. IEEE, 2013, pp. 1449–1453.

[8] L. Gao, X. Yi, Z. Jiang, L. Hao, and Z. Tang, “Icdar2017 competition on page object detection,” in Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on, vol. 1. IEEE, 2017, pp. 1417–1422.

[9] J. Ha, R. M. Haralick, and I. T. Phillips, “Recursive xy cut using bounding boxes of connected components,” in Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on, vol. 2. IEEE, 1995, pp. 952–955.

[10] T. G. Kieninger, “Table structure recognition based on robust block segmentation,” in Document Recognition V, vol. 3305.

International Society for Optics and Photonics, 1998, pp. 22–33.
[11] E. Koci, M. Thiele, W. Lehner, and O. Romero, “Table recognition in spreadsheets via a graph representation,” in 2018 13th IAPR International Workshop on Document Analysis Systems (DAS), 2018, pp. 139– 144.
[14] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
[15] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[16]Fahmy, Maged, “Online handwritten signature verification system based on DWT features extraction and neural network classification”, Ecological Modelling - ECOL MODEL. 1. 59-70. 10.1016/j.asej.2010.09.007, 2010.[17] Thomas Lang et al., ,“Page Segmentation and Region Classification Based on Region Bounding Boxes”, Proceedings of the OAGM Workshop 2018, DOI: 10.3217/978-3-
[12] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.
[13] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: towards realtime object detection

85125-603-1-12
[18] Kise K. “Page Segmentation Techniques in Document Analysis”, In: Doermann D., Tombre K. (eds) Handbook of Document Image Processing and Recognition. Springer, London. https://doi.org/10.1007/978-0-85729-859-1_5, 2014.
[19] Kasturi, Rangachar et al. “Document image analysis: A primer.” Sadhana 27 (2002): 3-22.
[20] image reference. Md Zahangir Alom, Tarek M. Taha, Chris Yakopcic, Stefan Westberg, Vishal K. Singh
Source: ResearchGate March 2019