A Comprehensive Study on Text Detection and Extraction from Images and PDF Documents

Mayank Deshmukh[1], Saloni Rabde[2], Priyanka Makode[3], Sourabh Jasuja[4], Prof.Bhavesh Khasdev[5]

[1,2,3,4]B.Tech Scholar, [5]Professor
Department of Artificial Intelligence & Data Science
Shri Balaji Institute of Technology and Management
Betul, RGPV University M.P India.

Abstract:
The growing need for digitization and intelligent document processing has led to significant advancements in text detection and extraction technologies. This paper reviews methodologies and tools employed for extracting textual information from images and Portable Document Format (PDF) files. Both traditional Optical Character Recognition (OCR) techniques and modern deep learning-based approaches are discussed. Five major research contributions in this area are analyzed in detail. The paper further explores challenges in handling complex document layouts, multilingual text, and low-quality images, and highlights research gaps and future directions that emphasize the potential of artificial intelligence and multimodal learning to enhance text extraction accuracy and efficiency.

Key words: OCR, Text Extraction, Deep Learning, Layout LM, Scene Text Detection, Document Analysis.

I. INTRODUCTION

Text detection and extraction are fundamental processes in computer vision and document analysis, enabling the conversion of image-based text into machine-readable formats. With the widespread use of digital imaging devices and PDF documents, organizations increasingly rely on automated text extraction systems for document management, search indexing, and data analytics.

Traditional OCR methods, such as the Tesseract engine [5], rely on image preprocessing and pattern recognition techniques to identify text. However, these systems often fail under conditions involving complex backgrounds, distortions, or handwritten scripts. Modern deep learning methods utilizing Convolutional Neural Networks (CNNs) and Transformers have drastically improved text recognition accuracy in both natural scene images and scanned documents. Despite these advances, challenges such as multi-script recognition, layout understanding, and noise reduction persist.

This study examines existing literature on text detection and extraction, compares different methodologies, and identifies opportunities for improvement in this rapidly evolving domain.

2. LITERATURE REVIEW

This section presents an in-depth review of five significant research studies related to text detection and extraction from images and PDFs.

2.1 Stroke Width Transform for Scene Text Detection (Epshtein et al., 2010)

Epshtein et al. [3] proposed the Stroke Width Transform (SWT), a robust method for detecting text in natural scene images. SWT computes stroke widths for each pixel and groups those with similar values to identify text regions. The approach effectively detects text even in cluttered backgrounds, outperforming earlier connected-component methods. However, it struggles with curved or stylized fonts, limiting accuracy on handwritten text.

2.2 Tesseract OCR Engine (Smith, 2007)

Smith [5] developed the Tesseract OCR engine, an open-source tool for recognizing printed text from scanned documents. It uses adaptive thresholding, character segmentation, and language modeling to recognize text lines. While it provides high accuracy for clean, high-resolution images, it performs poorly on noisy, low-contrast, or skewed images and has limited ability to preserve document layouts.

2.3 Connectionist Text Proposal Network (Tian et al., 2016)

Tian et al. [6] introduced the Connectionist Text Proposal Network (CTPN), a deep learning-based framework for text detection in natural images. CTPN combines CNN and RNN layers to detect text sequences at the line level. It demonstrated significant improvements in detecting horizontally aligned text in complex backgrounds, though it was less effective for multi-oriented and curved text.

2.4 CRAFT – Character Region Awareness for Text Detection (Baek et al., 2019)

Baek et al. [1] proposed the CRAFT model to enhance the precision of text localization through pixel-level character region detection. It generates heatmaps of individual characters and their affinities using CNNs. CRAFT achieved state-of-the-art results on several text detection benchmarks, effectively handling arbitrary text shapes. However, its high computational cost limits deployment on low-power devices.

2.5 LayoutLM for Document Understanding (Xu et al., 2020)

Xu et al. [8] developed LayoutLM, a transformer-based pre-trained model that integrates textual, spatial, and visual information for document representation. It significantly improved extraction accuracy in structured PDFs and outperformed standard OCR in tasks such as form extraction. Nonetheless, it requires large annotated datasets and substantial computational resources

System Architecture Draw: -

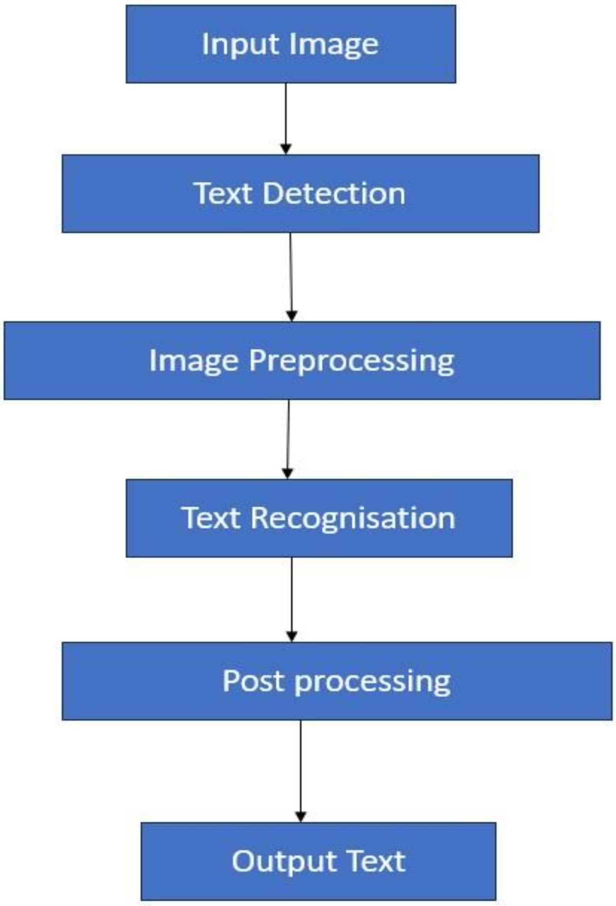

Fig 1.1

3. METHODOLOGY

3.1 Pre-processing: In this stage, input images or PDFs are converted to grayscale to simplify processing and reduce computation. To enhance text clarity, noise is removed using filters such as Gaussian or Median filters. Then, binarization is performed using Otsu’s method, which automatically separates text (foreground) from the background. After binarization, images are deskewed to correct any rotation or alignment issues, ensuring that text lines are horizontally aligned for better recognition accuracy. This step improves the quality of the data for the next phase.

3.2 Text Detection: The text detection stage identifies the regions containing text. Traditional approaches such as Edge Detection, Maximally Stable Extremal Regions (MSER), and Connected Component Analysis (CCA) are used to locate potential text areas. However, these methods may fail in complex or noisy backgrounds. To overcome such limitations, modern deep learning models like EAST (Efficient and Accurate Scene Text Detector) and CTPN (Connectionist Text Proposal Network) are used. These models efficiently detect text in different orientations and scales using convolutional neural networks.

3.3 Text Extraction (Recognition): After detecting text regions, Optical Character Recognition (OCR) is performed to extract textual content. Traditional tools like Tesseract OCR convert detected characters into machine-readable text. For improved performance, deep learning-based architectures such as CRNN (Convolutional Recurrent Neural Network) and TrOCR (Transformer-based OCR) are employed. These models handle various fonts, languages, and handwritten text effectively.

3.4 Post-processing: In the final stage, extracted text is refined using spell correction, layout analysis, and pattern matching (e.g., Regular Expressions). This ensures grammatical correctness and structural accuracy. The final, error-free text is then stored in structured formats like CSV or JSON for further processing or analysis.

4. DISCUSSION

The reviewed studies demonstrate an evolution from rule-based to deep learning-based methods. Traditional OCR engines like Tesseract remain useful for clean scanned documents, while deep learning methods like CTPN and CRAFT excel at detecting text in complex real-world scenes. No single approach, however, addresses all challenges. Scene text detectors often fail in document-based contexts, while OCR tools struggle with non-standard fonts and artistic text. Integrating layout understanding, as demonstrated by Layout LM, provides a promising direction for bridging the gap between visual and linguistic analysis. The lack of universal multilingual datasets and efficient low-resource models remains a major obstacle for comprehensive text extraction.

5. RESEARCH GAP

5.1 Multilingual and Handwritten Text Recognition: Most models focus on a limited number of languages and perform poorly on handwritten or stylized text.

5.2 Layout Preservation: Many OCR systems extract text but fail to retain document structure such as tables, columns, and forms.

5.3 Low-Quality Image Handling: Noise, blur, and illumination inconsistencies still affect model performance.

5.4 Real-Time Performance: Many deep learning models require heavy computation, hindering real-time deployment.

6. CONCLUSION

Text detection and extraction from images and PDFs play a crucial role in digitization and artificial intelligence. The evolution from traditional OCR to deep learning-based models has substantially improved extraction accuracy and flexibility. However, persistent challenges such as multi-language handling, structural preservation, and computational efficiency need further attention. Future advancements in multimodal and self-supervised learning approaches are expected to revolutionize the field, enabling context-aware and high-accuracy text extraction for diverse applications.

REFERENCES:

[1] J. Baek, B. Lee, D. Han, S. Yun, and H. Lee, “Character region awareness for text detection,” Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2019.

[2] K. Chen and J. Hull, “A fast and robust text detection approach in images and video frames,” Proc. Int. Conf. Document Analysis and Recognition (ICDAR), 2001.

[3] B. Epshtein, E. Ofek, and Y. Wexler, “Detecting text in natural scenes with stroke width transform,” Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2010.

[4] R. Sarkhel, A. Naskar, and A. Das, “A hybrid framework for extracting text from PDF documents,” Pattern Recognition Letters, vol. 125, pp. 33–40, 2019.

[5] R. Smith, “An overview of the Tesseract OCR engine,” Proc. Int. Conf. Document Analysis and Recognition (ICDAR), 2007.

[6] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, “Detecting text in natural image with connectionist text proposal network,” Proc. Eur. Conf. Computer Vision (ECCV), 2016.

[7] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, and W. He, “EAST: An efficient and accurate scene text detector,” Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2017.

[8] Y. Xu, T. Lv, L. Cui, G. Wang, and F. Wei, “Layout LM: Pre-training of text and layout for document image understanding,” Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), 2020.

[9] S. Long, J. He, and C. Yao, “Scene text detection and recognition: The deep learning era,” *International Journal on Document Analysis and Recognition (IJDAR),* 2019.

[10] P. Wang, L. Li, and C. Shen, “Towards end-to-end text spotting with convolutional recurrent neural networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI),* 2021.