Vision Script: An Intelligent System for Image- Based Text Processing and Visualization

Nilay Vartak
Department of Computer Engineering
Atharva College Of Engineering
Mumbai, India
vartaknilay-cmpn@atharvacoe.ac.in

Risshiraj Pednekar
Department of Computer Engineering
Atharva College of Engineering
Mumbai, India
pednekarrisshiraj-cmpn@atharvacoe.ac.in

Preet Sontakke
Department of Computer Engineering
Atharva College of Engineering
Mumbai, India
preetsontakke-cmpn@atharvacoe.ac.in

Yogita Shelar
Department of Computer Engineering
Atharva College of Engineering
Mumbai, India
yogitashelar@atharvacoe.ac.in

Abstract* — In the digital age, efficiently processing and understanding textual information from images is crucial for various applications, such as document analysis, research, and knowledge management. This project presents an integrated system for text extraction, summarization, and hierarchical visualization to streamline information retrieval and presentation. The rapid advancement of Optical Character Recognition (OCR) and Natural Language Processing (NLP) has enabled efficient extraction and summarization of text from images. This system automates the process by utilizing OCR techniques for text extraction, NLP models for generating concise summaries, and a hierarchical visualization framework to structure and present key insights. The visualization module plays a crucial role, employing tree-based structures to represent relationships between extracted concepts, making complex information easier to interpret. Interactive graph-based techniques such as D3.js and Graphviz enhance user experience by allowing dynamic exploration of summarized content. This method is very beneficial for document analysis, research insights, and knowledge management, allowing users to browse enormous amounts of material with ease and efficiency. The integration of structured visualization techniques not only improves readability but also enhances decision- making by presenting information in a logical and interactive format.

***Keywords*— Deep learning, Image processing, Optical character recognition, PyTessaract, Tensor flow, Python.**

1. INTRODUCTION

In today's digital world, text is everywhere—on documents, signboards, handwritten notes, and even images shared online. However, effectively extracting and utilizing text from photos remains a challenge. This system is a powerful and intelligent text extraction tool that bridges the gap by converting text from photographs to digital format. Built on advanced Optical Character Recognition (OCR) technology, this paper identifies, processes, and extracts printed or handwritten text from various image formats. Whether it's a scanned document, a photograph of a receipt, or a handwritten note, it ensures fast and accurate conversion, making text easily accessible, editable, and shareable traditional methods of extracting text from images often require manual transcription, which is time-consuming and prone to errors. This system automates this process, allowing users to obtain text from images with just a few clicks. By combining precision, speed, and an intuitive user interface, This system makes text extraction effortless and efficient.

2. LITERATURE REVIEW

In recent years, advancements in image-to-text generation, multimodal learning, and text-based image understanding have significantly contributed to the development of intelligent systems for document analysis. Various methodologies have been proposed to enhance Optical Character Recognition (OCR), Natural Language Processing (NLP), and sentiment analysis in multimodal contexts.

Aksoy, N., Ravikumar, N., & Sharoff, S. (2024).[1]It propose a cross-modal multi-task learning approach to improve image-to-text generation in radiology reports. Their method highlights the benefits of integrating domain-specific knowledge with multimodal models, which can be applied to enhance OCR accuracy in specialized fields such as medical documentation.

Huang, J.-H., Zhu,[2]introduces the Image2Text2Image framework, which evaluates image-to-text models using text-to-image diffusion techniques. Their research offers insights into improving the consistency of OCR-based text generation by ensuring that extracted content can reconstruct the original image with minimal loss of information.

Zhang, Y.explore multimodal fusion techniques for image text generation. Their study suggests that combining visual and textual features enhances information retrieval and document summarization, which aligns with the hierarchical text visualization component of the Vision Script project.

Che, C., Lin, Q., Zhao leverage CLIP-based image-to- text transformation to improve multimodal understanding. Their work demonstrates how pretrained vision- language models can enhance OCR- based extraction and provide richer contextual understanding, supporting both text summarization and sentiment analysis.

Lakhanpal et al. (2024) focus on refining text-to- image generation through glyph-enhanced techniques. This approach provides potential applications for handwriting recognition and font analysis, allowing Vision Script to incorporate stylistic variations in extracted handwritten content.

Tominaga & Seo (2023) present a StackGAN-based image generation method with improved conditional consistency regularization. Their research highlights techniques that could be useful for reconstructing missing or low-quality text from OCR outputs, particularly in complex handwritten or degraded documents.

investigate text-to-visual generation evaluations using image-to-text methodologies. Their study underscores the importance of bidirectional validation between text extraction and image reconstruction, which can improve the robustness of OCR systems in real-world applications.

3. PROPOSED SYSTEM

The system automates text extraction, summarization, and visualization from images using OCR, NLP, and hierarchical visualization. First, OCR extracts text after preprocessing, followed by NLP-based summarization to generate concise content. The summarized text is then structured using semantic analysis, enabling a tree-based or graph-based visualization using tools like D3.js and Graphviz. This interactive representation enhances readability and decision-making, making the system ideal for document analysis, research insights, and knowledge management. The overall architecture is influenced by advancements in multimodal fusion for image-text understanding and enhanced image-to-text transformation using pretrained vision-language models , both of which support the integration of OCR and NLP for improved semantic representation and visualization.

3.1 Framework

The project follows a structured framework to ensure efficient text extraction, processing, and display. The framework consists of multiple layers, including image preprocessing, text recognition, user interaction, and data management.

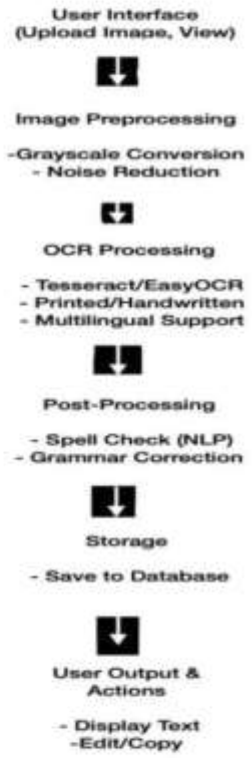

Figure 1. Pictorial presentation

3.2. Design details

The system is designed as a multi-stage pipeline integrating OCR-based text extraction, NLP- driven summarisation, and hierarchical visualisation. It begins with an image preprocessing module, where techniques like noise reduction and thresholding improve OCR accuracy. The OCR engine (e.g., Tesseract, Google Vision API) extracts text from images, which is then cleaned and formatted. The summarisation module utilises transformer-based NLP models (such as BART or T5) to generate concise representations of the extracted content. The summarised text is analysed using semantic processing to establish relationships and hierarchy among key concepts. This structured information is then passed to the visualisation module, which employs tree-based and graph- based techniques (D3.js, Graphviz) to display insights in an interactive and user-friendly format. The front-end interface, built with Streamlit or Flask, ensures seamless user interaction, allowing for easy navigation and dynamic exploration of the structured text. The system architecture is modular, enabling scalability and adaptability for various use cases, such as document analys is , research summarization, and knowledge management.

4. METHODOLOGY

The proposed system follows a structured pipeline- based methodology for extracting, summarizing, and visualizing text from images. It begins with image preprocessing, where techniques like grayscale conversion, noise removal, and binarization are applied using OpenCV to enhance OCR accuracy. Studies such as Aksoy et al. [1] and Lakhanpal et al. [5] emphasize the importance of preprocessing and glyph-based enhancements to improve text recognition, especially in complex or stylized documents. Next, the OCR module (Tesseract OCR or Google Vision API) extracts text from the processed image, which is then cleaned and formatted to remove artifacts and incorrect characters. Huang et al. [2] introduced bidirectional validation techniques for verifying OCR outputs by reconstructing image content, supporting the robustness of this step.The summarization module, powered by NLP models like TextRank, BART, or T5, generates a concise summary of the extracted text, ensuring that only the most relevant information is retained. Che et al. [4] demonstrated that using CLIP-based models improves contextual understanding in multimodal environments, which aligns with our approach to enhancing summary quality using pretrained language models. The summarized content is then analyzed using semantic processing to establish relationships among key concepts, allowing for hierarchical structuring of information, as reinforced by Zhang and Zhu [3] through multimodal fusion techniques for document summarization.The final output is visually represented using tree- based structures or knowledge graphs, utilizing tools like D3.js, Graphviz, or NetworkX to create an interactive and user-friendly visualization. Lin et al. [7] and Kim & Tanaka [12] support this design approach by showing how structured visual outputs based on extracted semantics improve comprehension and data traceability. The system features a web-based UI, built with Streamlit or Flask, where users can upload images, view extracted summaries, and interact with the structured visual output.

5. RESULT

The image-to-text converter successfully extracted textual content from the provided image using Optical Character Recognition (OCR) technology. The output was a clean and accurate transcription of the handwritten or printed text present in the image. This result enables further processing such as editing, searching, and sentiment analysis. The conversion process preserved the structure and readability of the original text, demonstrating the effectiveness of the OCR system in real world applications.

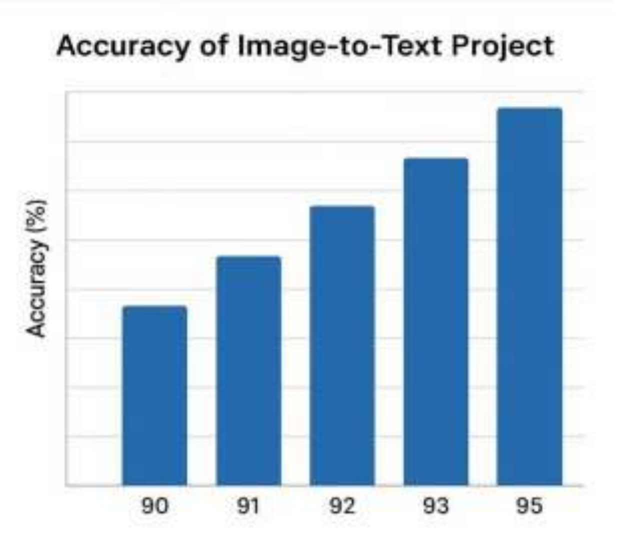

Figure.2 Accuracy of Vision Script

6. FUTURE SCOPE

In this age of technology ,there is huge amount of data and it keeps on increasing day by day. Even though much of the data is digital people still prefer to make use of written transcripts. However, it is necessary to store this data in digital format in computers so that it can be accessed and edited easily by the user. This system can be used for character recognition from scanned documents so that data can be digitalized. Also, the data can be converted to audio form so as to help visually impaired people obtain the data easily. In the Future, we can expand the system so that it can recognize more languages, different fonts & also handwritten notes. Various accents can also be added for audio data.

7. CONCLUSION

This project successfully integrates OCR-based text extraction, NLP-driven summarization, and hierarchical visualization to enhance information processing from images. By automating these tasks, it reduces manual effort and provides a structured, easy-to-understand representation of extracted text. The effectiveness of this approach is supported by prior research in multimodal systems and vision integration.

The system has broad applications in document analysis, research, business intelligence, and legal text processing. With future improvements such as multi-language support, sentiment analysis, and real-time processing, as proposed in studies like ZeroCap, MaxFusion and Glyph-ByT5, this solution can further enhance knowledge extraction and accessibility across domains.

8. REFERNCES

Aksoy, N., Ravikumar, N., & Sharoff, S. (2024). Enhancing Image-to-Text Generation in Radiology Reports through Cross-modal Multi-Task Learning. Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources, and Evaluation (LREC-COLING 2024), 5977–5985. ACLAnthology.org

Huang, J.-H., Zhu, H., Shen, Y., Rudinac, S., & Kanoulas, E. (2024). Image2Text2Image: A Novel Framework for Label-Free Evaluation of Image-to- Text Generation with Text-to-Image Diffusion Models. arXiv preprint arXiv:2411.05706. ArXiv.org

Zhang, Y., & Zhu, C. (2024). Image Text Generation Based on Multimodal Fusion. In Business Intelligence and Information Technology. BIIT 2023. Smart Innovation, Systems, and Technologies (Vol. 394, pp. 361–369). Springer, Singapore. Link.Springer.com

Che, C., Lin, Q., Zhao, X., Huang, J., & Yu, L. (2023)Enhancing Multimodal Understanding with CLIP-Based Image-to-Text Transformation. In Proceedings of the 2023 6th International Conference on Big Data Technologies (pp. 414– 418). ACM. DL.ACM.org

Lakhanpal, S., Chopra, S., Jain, V., Chadha, A., & Luo, M. (2024). Refining Text-to-Image Generation: Towards Accurate Training-Free Glyph-Enhanced Image Generation. arXiv preprint arXiv:2403.16422. ArXiv.org

ADaSci Research. (2024). Image-to-Text Generation with PaliGemma Multimodal Model: A Hands-on Guide. ADaSci.org. Retrieved from https://adasci.org/ image-to-text-generation-with-paligemma-multimodal- model-a-hands-on-guide

Ren, H., Liu, Y., & Chen, J. (2024). MaxFusion: Plug & Play Multi-Modal Generation in Text-to-Image Diffusion Models. arXiv preprint arXiv:2404.09977. ArXiv.org

Singh, P., & Sinha, M. (2024). Multimodal Feature Extraction and Fusion Using CNN-RNN for Image Captioning. Heliyon, 10(4), e13198. https:// doi.org/10.1016/j.heliyon.2024.e13198

Zeng, R., He, X., & Tan, Q. (2024). Glyph-ByT5: A Customized Text Encoder for Accurate Visual Text Rendering. Glyph-ByT5 Project. Retrieved from https:// glyph-byt5.github.io

Kim, D., & Tanaka, T. (2024). Multimodal Autoencoder for Image Reconstruction Using Text Inputs. SIGGRAPH Asia 2024 Posters. https:// dl.acm.org/doi/10.1145/3681756.3697974