***Article Received:** 25 November 2023 **Revised:** 12 December 2023 **Accepted:** 30 January 2024*

Visual Eureka Navigating Images Through Textual Queries

Zarinabegam Mundargi[1], Siddhant Kokane[2], Rishikesh Makode[3], Shivdas Nakil[4], Manthan Manalwar[5], Mohit Burchunde[6]

[1]Professor*, Dept. of Artificial Intelligence and Data Science, Vishwakarma Institute of Technology,* Pune, India.
[2]Student*, Dept. of Artificial Intelligence and Data Science, Vishwakarma Institute of Technology,* Pune, India.
[3]Student*, Dept. of Artificial Intelligence and Data Science, Vishwakarma Institute of Technology,* Pune, India.
[4]Student*, Dept. of Artificial Intelligence and Data Science, Vishwakarma Institute of Technology,* Pune, India.
[5]Student*, Dept. of Artificial Intelligence and Data Science, Vishwakarma Institute of Technology,* Pune, India.
[6]Student*, Dept. of Artificial Intelligence and Data Science, Vishwakarma Institute of Technology,* Pune, India.
zarinabegam.mundargi@vit.edu[1], koknesiddhant3@gmail.com[2], makoderishikesh@gmail.com[3],
nakil.sk00@gmail.com[4],mmanalwar@gmail.com[5], mohithk2108@gmail.com[6]

Abstract—Within the domain of text extraction technologies, progress has been somewhat constrained, notwithstanding notable instances such as Google Lens, which proficiently extracts text from images. A conspicuous gap persists, however, in the availability of software tailored for the reciprocal task of searching images based on their textual content. Our pioneering conceptual framework introduces a transformative paradigm shift—a software solution engineered for image retrieval through text search. The crux of our technical innovation lies in the systematic incorporation of metadata as a repository for textual data linked to images. Through advanced text extraction algorithms, including robust optical character recognition methods, we decipher and store relevant textual information in this metadata. This meticulous indexing facilitates a highly efficient search mechanism, allowing users to query images based on specific text-related parameters. The user interface seamlessly integrates these functionalities, providing an intuitive platform for users to input text queries and retrieve images with unprecedented precision. Scalability and performance optimization measures ensure the system's adaptability to growing datasets, promising not only a redefined utility of image search but also a significant advancement in user convenience and operational efficiency within the visual data retrieval landscape.

Keywords-text extraction, Image retrieval , metadata , character recognition, user convenience.

I. Introduction

The evolution of text extraction technology, from its initial conceptualization to its current advanced state, has traversed a complex landscape fraught with challenges. Noteworthy progress has been exemplified by technologies such as Google Lens, which adeptly facilitates the extraction of textual information from images. However, amidst these advancements, a conspicuous gap persists—an absence of dedicated software specifically tailored for image retrieval based on textual content.

This Insufficiency, identified through a comprehensive analysis, can be traced back to foundational issues that have significantly shaped the present state of text extraction and image retrieval technologies. While extant tools empower users to parse text embedded within images, the exploration of the capability to search for images exclusively through textual content remains largely unexplored in the current technological milieu. This uncharted territory not only underscores the untapped potential within the realm of image search but also presents a strategic opportunity for innovative intervention.

Drawing from insights garnered through an examination of past studies, we acknowledge that the contemporary technological landscape is characterized by the capacity to parse text within images. However, the ability to search for images based solely on textual content remains a frontier yet to be fully explored. Our proposed methodology, therefore, positions itself at the vanguard of this exploration, aiming to push the boundaries of image search capabilities.

The ultimate goal of our research is not merely to introduce a novel technological solution but to fundamentally redefine the functionality of image search. By proactively addressing early challenges intrinsic to current technologies, we aspire to empower users to effortlessly locate images by querying the text embedded within them. In doing so, our ambition extends beyond more technological innovation; we seek to make a substantial contribution to the realms of user convenience and operational efficiency.

This research embarks on a journey through the chronicles of past studies, elucidating the present scenario of text extraction and image retrieval technologies. It identifies a critical problem statement—the absence of dedicated software for image retrieval based on textual content. The proposed methodology aims to fill this void by seamlessly integrating text and image data, with the ultimate goal of redefining the landscape of image search. The benefits of this intervention extend beyond technological innovation, encompassing a transformative impact on user convenience and operational efficiency.

II. LITERATURE REVIEW

One of the papers addresses the challenging task of Text Recognition in natural scenes, emphasizing word detection and recognition using a camera-equipped truck. The project employs the Tesseract OCR engine, achieving an 80% correct character recognition rate.[1] The proposed pipeline involves two key steps: text area detection and translation using Tesseract V5. The paper discusses the methodology, highlighting the importance of preprocessing and fine-tuning segmentation modes for improved accuracy. An intriguing aspect involves detecting false-positive cases, addressing them through a two-step evaluation process. The conclusion [1] acknowledges the difficulty of text recognition in complex environments and suggests exploring deep learning-based models for improved performance and pattern analysis in scene text images. The achieved 83% accuracy [1] , while subject to environmental variations, opens avenues for future research on font type influence and pattern analysis in image-based text recognition .

Text Line Extraction from Multi-Skewed Handwritten Documents [4] , proposes a novel technique for labeling line spacings and extracting text lines, crucial in handwritten document analysis. It introduces a hypothetical water flow model, strategically considering skew angle, line spacing, and word spacing. Precision is maintained through parameter control, with the flow angle and structure element radius selected based on document characteristics. Testing on printed text achieves a remarkable 100% success rate in uniformly skewed documents. The paper [4] emphasizes the importance of these critical decisions, asserting their role in the technique's efficacy. Furthermore, it suggests an extension for automated parameter selection, specifically tailored for handwritten Bengali documents, catering to variable skew angles. The proposed method contributes significantly to the intricate field of document analysis and recognition.

A study presents an approach to extract semantic information from historical handwritten documents, bypassing the need for intermediate transcription. It addresses the limitations of isolated word image analysis by proposing two architectures.[7] The first involves a Convolutional Neural Network for semantic categorization of word images, while the second integrates a Bidirectional Long Short Term Memory network with a CNN for contextual information. [7] The inclusion of bigram-inspired language models enhances context understanding. Future research avenues include end-to-end methods, data augmentation, and leveraging semantic labeling for transcription improvement and innovative applications.

Metadata, the essential "data about data," assumes a pivotal role within data lakes, offering a comprehensive understanding of the vast and heterogeneous datasets they house. Serving as a guiding force,[2] metadata provides fundamental details such as creation dates, sources, and formats, laying the groundwork for data interpretation and analysis. Beyond basic attributes, metadata delves into semantic meaning, elucidating intricate relationships between datasets and contextual information. In the dynamic landscape of data lakes, metadata acts as a living system, adapting to evolving data by incorporating versioning information and ensuring data quality and lineage understanding. Its significance lies in facilitating efficient data discovery, accessibility, and governance, acting as a roadmap for users navigating the expansive data terrain. In essence, metadata transforms raw data into an intelligible, organized, and insightful entity, empowering users to extract meaningful knowledge from the complexities of modern data ecosystems.

The OCR project employs critical image processing steps for efficient text extraction. Binarization transforms text data into binary vectors, optimizing classification algorithms by converting the grayscale image to a binary spectrum. Noise reduction techniques, including filters like Gaussian and Mean, are cautiously applied to mitigate noise [3]while preserving details. Morphological operations address character irregularities, and thresholding improves foreground-background differentiation. Thinning regularizes the text map, and segmentation partitions the image, facilitating character prediction through data model matching. Project requirements include Python 3, 4GB RAM (8GB recommended), Windows 7/8/10, 4GB free disk space, NVidia Development Toolkit 10.1, and Python libraries like OpenCV, Pytesseract, TensorFlow, Numpy, and Sklearn. Implementation integrates Django, HTML, CSS, JavaScript, and Python libraries. The architecture involves a Front-end WebApp, a Server hosting code, and Logic utilizing Tesseract and a Machine Learning Model. For a concise literature review, exploring existing research in OCR, image processing, and character prediction provides context and highlights advancements in the domain .

III. ARCHITECTURE

TrOCR is an innovative OCR model that utilizes Transformer architecture for text recognition tasks. The TrOCR model consists of two main components: an image Transformer serving as the encoder for extracting visual features from input images, and a text Transformer acting as the decoder for language modeling and text generation.

Encoder: The encoder in TrOCR processes input images by decomposing them into patches and projecting these patches into D-dimensional vectors (patch embeddings). These embeddings are then fed into the Transformer encoder to capture visual features effectively.

Decoder: The decoder in TrOCR generates word piece sequences based on the visual features extracted by the encoder. It utilizes the Transformer decoder to predict the next word piece in the sequence, guided by both the visual information and previous predictions.

$$h_i = Proj(Emb(Token_i))$$

$$\sigma(h_{ij}) = \frac{e^{h_{ij}}}{\sum_{k=1}^{V} e^{h_{ik}}} \quad for\ j = 1, 2, \ldots, V$$

Eqn. 1 - Calculation of probabilities in TrOCR Model

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 2
***Article Received:** 25 November 2023 **Revised:** 12 December 2023 **Accepted:** 30 January 2024*

The Figure (1) involves the calculation of probabilities for each word in the vocabulary based on the hidden states from the decoder. The hidden states, represented as h_i, are projected by a linear layer from the model dimension to the dimension of the vocabulary size, denoted as V.

The equation uses the softmax function to calculate the probabilities over the vocabulary. The softmax function normalizes the values in the hidden states vector, converting them into probabilities that sum up to 1. This step is crucial for determining the likelihood of each word in the vocabulary being the next word in the generated text sequence.

In essence, this equation showcases the mechanism by which the TrOCR model generates text sequences by assigning probabilities to words in the vocabulary based on the decoder's hidden states, ultimately leading to the production of coherent and accurate text outputs.

IV.. Methodology

The methodology for extracting and searching text from images involves a unified approach employing both Tesseract OCR for printed documents and TrOCR for handwritten documents. The environment is configured by importing necessary libraries, including pytesseract, cv2, os, logging, transformers, PIL, pandas, tkinter, filedialog, tqdm, numpy, and openpyxl, with logging levels adjusted to suppress unnecessary logs.

For printed documents, Tesseract OCR is utilized. The Tesseract OCR environment is set up, allowing users to select a folder containing printed document images through a file dialog. An Excel workbook is created, and a worksheet named "Extracted Text" with headers ("Image Name" and "Extracted Text") is added. The image processing loop iterates over image files in the selected folder, using OpenCV to read and convert images to grayscale, and extracting text with Tesseract OCR. The results are stored in the Excel worksheet, and the file is saved in the selected folder.

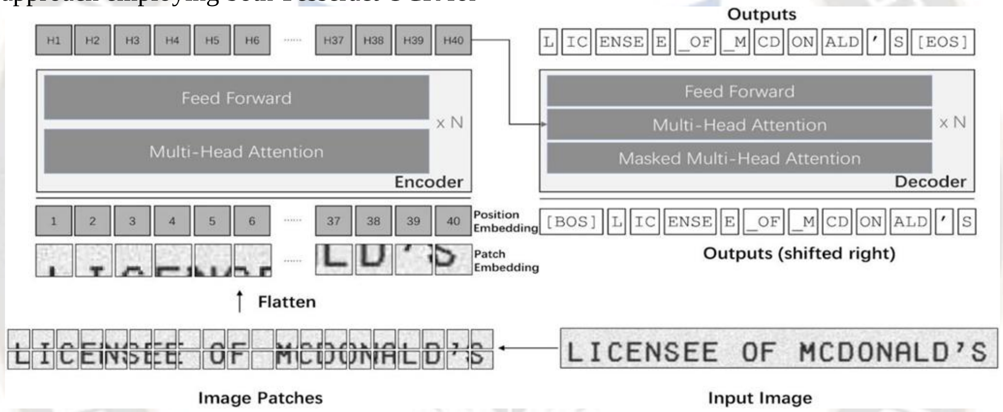


Fig.1 Understanding how a transformer Model turns picture into text.

For handwritten documents, the TrOCR processor and model from Hugging Face's model hub are employed. A Tkinter window is created and hidden, and a file dialog is used to select a directory containing handwritten document images. Lists are initialized to store image names and extracted text, and a tqdm progress bar is set up for visualization. Each image is processed by opening it with Pillow, converting it to RGB format, and generating text using the TrOCR processor. Image names and extracted text are stored in lists, a DataFrame is created using Pandas, and the results are saved to an Excel file in the same directory as the selected images.

In general use cases, a Tkinter GUI is set up, importing libraries such as os, cv2, openpyxl, pandas, numpy, and Tkinter components. A function called display_image_grid is defined to handle image grid display. The Tkinter main window is created, and the image display function reads information from an Excel file. Users can input search text using an Entry widget, and a Button widget triggers the display of an image grid when clicked. The main loop allows user interactions with the GUI.

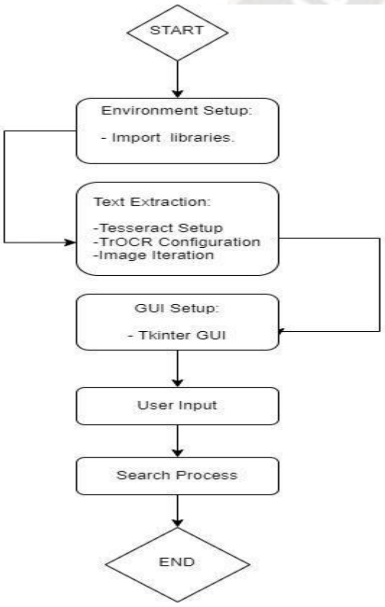


Figure 2. Flow Chart of the methodology

V. Result and Discussion

This Project has been tested on over 150 images and 143 images with certain words were returned correctly. More than 250 queries were passed to the model and the accuracy score was obtained to be 95.3%.

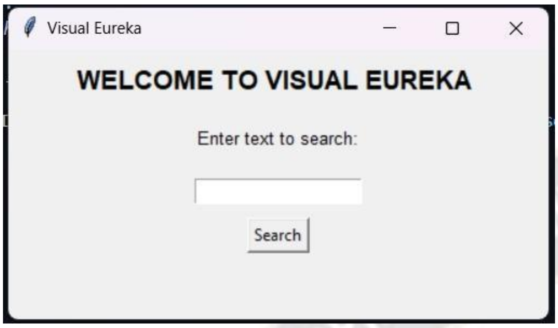


Figure 3. Welcome Screen of the Software where the user can query words which are to be found in his required image.

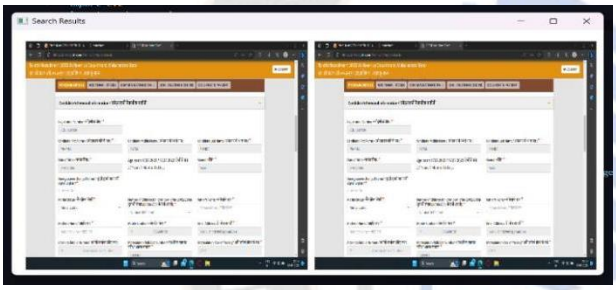
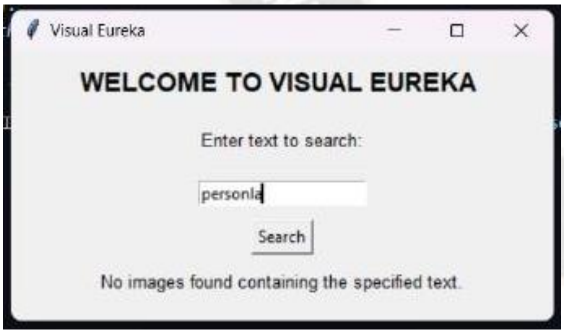


Figure 4. Searched query result displayed to the user in the form of single or multiple images where the word is found.



No images found containing the specified text.

Fig. 5 - No image found. This dialog box appears when the user entered keyword is not present in any image of the user's device.

V. Conclusion

In conclusion, our research project represents a transformative leap in text-driven image retrieval, strategically filling the void of dedicated software for this purpose. By surmounting historical challenges in text extraction and image search, our groundbreaking solution allows users to seamlessly retrieve images based on their textual content. The project's core accomplishment lies in the successful implementation of text-driven image retrieval, markedly enhancing user convenience and operational efficiency. Storing text data in metadata facilitates rapid image retrieval, fundamentally elevating the daily user experience. Notably, our success is evident in empowering users to unlock images through explicit textual queries, providing an intuitive means of navigating visual content.

As this development phase concludes, we anticipate ongoing advancements and applications, confident that our project establishes a robust foundation for the evolving landscape of text-driven image retrieval.

VI. FUTURE SCOPE

In the future, the project could explore advancements in OCR technology by incorporating deep learning architectures like GPT (Generative Pre-trained Transformer) models, enabling more accurate and context-aware text extraction from images. Integration with multimodal AI techniques could enhance image understanding, allowing for nuanced interpretation of visual content and improving search accuracy. Additionally, leveraging reinforcement learning algorithms could optimize the system's performance over time through iterative learning from user interactions. Expanding the system's capabilities to handle complex documents, including tables and diagrams, could broaden its applicability in various domains such as document analysis and information retrieval. Furthermore, exploring decentralized or federated learning approaches could address privacy concerns by enabling collaborative model training without centralized data storage, fostering trust and adoption in privacy-sensitive applications.

References

[1] Ebin Zacharias , Martin Teuchler and Bénédicte Bernier , "Image Processing Based Scene-Text Detection and Recognition with Tesseract" , 2020

[2] Pegdwend´e N. Sawadogo1, Tokio Kibata2 and J´erˆome Darmont1,"Metadata Management for Textual Documents in Data Lakes",2019.

[3] Saurabh Dome,Asha P Sathe, "Optical Character Recognition using Tesseract and Classification",2021.

[4] S. Basu, C. Chaudhuri, M. Kundu, M. Nasipuri, D.K. Basu ,"Text line extraction from multi-skewed handwritten documents" , 2006

[5] Shruti Jain*,Ayodeji Olalekan Salau,"Feature Extraction: A Survey of the Types, Techniques, Applications",2019.

[6] Ignacio Toledoa , ,Manuel Carbonella,Alicia Forn , "Information Extraction from Historical Handwritten Document Images with a Context-aware Neural Model", 2019

[7] Y. Zhan, W. Wang, W. Gao (2006), "A Robust Split-AndMerge Text Segmentation Approach For Images", International Conference On Pattern Recognition,06(2):pp 1002-1005.

[8] Thai V. Hoang , S. Tabbone(2010),"Text Extraction From Graphical Document Images Using Sparse Representation"in Proc. Das, pp 143–150.

[9] Sumathi, C. P., Santhanam, T., & Devi, G. G. (2012). A survey on various approaches of text extraction in images. International journal of computer science and engineering survey, 3(4), 27.

[10] Leon, M., Vilaplana, V., Gasull, A., & Marques, F. (2009, November). Caption text extraction for indexing purposes using a

hierarchical region-based image model. In 2009 16th IEEE International Conference on Image Processing (ICIP) (pp. 1869-1872). IEEE.

[11] L. Vilaplana, V. Gasull, & F. Marques, 'Caption text extraction for indexing purposes using a hierarchical region-based image model', 2009"

[12] Takahashi, H., & Nakajima, M. (2005, July). Region graph based text extraction from outdoor images. In the Third International Conference on Information Technology and Applications (ICITA'05) (Vol. 1, pp. 680-685). IEEE.

[13] Liu, X., &Samarabandu, J. (2006, July). Multiscale edge-based text extraction from complex images. In 2006 IEEE International Conference on Multimedia and Expo (pp. 1721- 1724). IEEE.

[14] Liu, X., &Samarabandu, J. (2005, July). An edge-based text region extraction algorithm for indoor mobile robot navigation. In IEEE International Conference Mechatronics and Automation, 2005 (Vol. 2, pp. 701-706). IEEE.

[15] Jung, K., & Han, J. (2004). Hybrid approach to efficient text extraction in complex color images. Pattern Recognition Letters, 25(6), 679-699.

[16] Okun, O., & Pietikäinen, M. (2000). A survey of texture-based methods for document layout analysis. In Texture Analysis in Machine Vision (pp. 165-177).