ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 4, April 2024

Information Extractor: A Unified Framework for Information Extraction from Multimodal Sources

[1] Hemant Suteri, [2] Subodh Sharma, [3] Vandana Choudhary, [4] Namita Goyal

[1] [2] [3] [4] Maharaja Agrasen Institute of Technology, Delhi, India
Corresponding Author Email: [1] hemantsuteri@gmail.com, [2] sharma2000subodh@gmail.com, [3] vandanachoudhary@mait.ac.in, [4] namitagoyal@mait.ac.in

Abstract*— In today's digital world everything is either converted into or is in process to be converted into digital form in one way or the other, similar is the thing with various documents like books, newspapers and documents of vehicles etc. and you might have also realized that nowadays, sales of digital form of books be it Google scholar or audio format like kindle by amazon, you might have seen newspapers/magazine focused more on digital format nowadays because of various reasons of convenience. We have seen every document going digital and with all such things around we can confidently say that the future is going to be of digital format of documents and managing such documents needs various technologies, one such technology is Tesseract OCR by Google which is an open-source Platform. Tesseract OCR where OCR stands for optical character recognition where it uses Artificial Intelligence to search the text and identify the image from the document.

Index Terms*— Face++, Image Recognition, Tesseract OCR, Text Recognition.

I. INTRODUCTION

Highlight Nowadays usage of digital documents has increased a lot and other traditional information sources like newspapers, magazines have gone digital. Tesseract OCR where OCR stands for optical character recognition which was developed for converting any kind of image into the digital format and searching the desired text into it. Tesseract OCR is a powerful tool as it can convert various fonts types and styles where the writing style of a same character may vary like it is sometimes confusing to differentiate between an 'o' (alphabet character) and '0' (zero) in various font styles, but tesseract OCR is able to identify and differentiate them. When the image is given to the Tesseract OCR it takes some steps for it to make it suitable for the searching of text, these steps include straightened de-speckled and convert it into the black and white image and the searching in the image is done on the basis of lines and strokes matching and the one with best matching pattern is shown as output. There is one more approach of pattern recognition where we try to match the pattern at the pixel level, in this approach we see the pattern of the pixel and match that in the whole document, and a pattern similar to that pattern is searched in the file.

II. MOTIVATION

If the purpose of this project is to ease the searching of any image or text in many formats of documents. Idea of such a project was in my mind since the first year of my college as in 2020 when the whole college life was going online, I faced the issue of having a lot of notes, file work mostly in pdf made of clicked pictures, there I often used to struggle in finding the specific relevant information in the all those files, there I thought of such an app/web app where any keyword or diagram/image etc. can be searched and also I used to read newspaper in online formats due to covid so there I used to search for some specific type of information and it used to take time, from there I got the idea of making such a project.

III. OBJECTIVES

The Objective of Information Extractor is made to make working with digital forms of documents easy with the use of various technical tools like OCR i.e., optical character recognition and Face++ along with OpenCV. It can have applications in various fields like from Research based Journalism to enforcement of law and from historical studies to academic studies, and many more. The objective is to make a holistic web app to search different types of information from multiple types of formats.

IV. AUTHOR'S CONTRIBUTION

Authors have worked on the easy and accurate extraction of data of different type from the source of multiple formats. Authors have worked on using multiple technologies together to get the accurate results along with the diverse features in the project. Information extractor is the solution for all those who have data to search in the various formats of data.

V. RELATED WORK

There has been a significant amount of research on the topic of information extraction with OCR and how it works, with a focus on developing methods that use natural language processing (NLP) techniques. Some examples of related work in this area include:

The study of challenges in OCR technology gave a brief explanation about critical fundamentals to keep in mind while working on OCR such as uneven light and color

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 4, April 2024

combination, skewness (rotation), aspect ratio etc. [1]

The study of OCR accuracy and improvement include brightness equalization, brightness and contrast adjustment grayscale, unsharp masking conversion of image removing background from image. [2]

The study of tesseract OCR includes work flow of an OCR and how OCR detect text from image. [3]

The study includes the algorithm involved in implementation of OCR like word and line finding such as baseline fitting, fixed-pitch detection & chopping of word. Recognition of word done by chopping joined characters, associating broken words. [4]

The study on object and face detection includes the algorithm such as local binary pattern and local binary pattern histogram in which pixel of the image is converted to binary matrix and stored as reference parameter. [5]

The study includes the algorithm called Maximally Stable Extremal Regions (MSER) which extracts multiple co-variant regions, called MSERs, and allows blobs to be identified from an object. An MSER is a balanced linked unit under varying threshold levels with uniform intensity level. [6]

The study includes the pre-procession working mechanism such that the image is first converted into a binary image having text representation as 1 and blank space as 0. Then individual line separated from binary text and then each alphabet then that recognize by OCR algorithm. [7]

This study gave the method of color reduction to convert image to reduced format by converting RGB color to HSV color which is easy to saturate the image color in HSV and get the region of interest. [8]

The study involves recognizing each letter and allocating it to the relevant letter group, resulting in the text being converted to a machine-readable format. For this objective, several separators based on artificial neural networks (ANN) and vector support machines (SVM). [9]

This study suggests to remove noise due to this quality of the image will increase and it will affect recognition process for better text recognition in images. [10]

This study gives the algorithm to convert color image to gray scale image by using expression $(x, y) = \Sigma((x, y)r, (x, y)g, (x, y)b) \div 3$, where x, y are pixels.[11]

This study gives the overview of challenges in face detection such as odd expressions, Face occlusion, Illuminations, Complex background, Less resolution. [12]

This study includes many algorithms used in face detection technique such as AdaBoost Face Recognition Algorithm, Face Recognition Algorithm Based on Linear Subspace, Zero Space Method. [13]

This study includes the algorithms used in face detection techniques such as Use viola Jones and KLT Algorithm to extract the region of interest in a rectangular bounding box. Convert to grayscale, apply histogram equalization and resize to 100x100 then apply Principal Component Analysis (PCA). [14]

This study contributes the 3 approaches of face recognition techniques such as Feature based approach, Holistic approach, Hybrid approach. Human face recognition can be divided into two strategies: geometrical features and template matching. [15]

VI. METHODOLOGY

The extraction of meaningful information from source images and newspapers is a crucial task in various fields, including journalism, historical research, and document analysis. However, manually extracting information from these sources is a time-consuming and labor-intensive process. To address this challenge, this research proposes a novel approach that utilizes the combined power of Optical Character Recognition (OCR), Computer Vision (CV), and Face Recognition (FR) to automate the information extraction process.

The various steps included in the process of recognition of text/image from an image/file is shown in Fig. 1. It basically can be divided into four major steps.

Data Collection: A diverse collection of source images and newspapers was assembled to ensure the robustness and generalizability of the proposed approach. The dataset consisted of images and newspapers with varying scene complexities, lighting conditions, skews, blurriness, degradations, aspect ratios, tilts, and fonts. This diversity ensured that the proposed approach could effectively handle a wide range of real-world image scenarios.

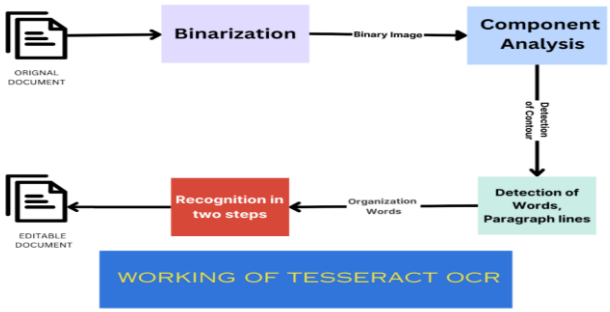

Fig1. Working of Tesseract OCR

Preprocessing: Before applying OCR, the collected images underwent a preprocessing stage to enhance their quality and prepare them for subsequent processing steps. This stage involved several techniques, including:

Noise Reduction: To eliminate noise and artifacts that could interfere with text extraction, noise reduction algorithms were applied to the images.

Contrast Enhancement: To improve the contrast between text and background, contrast enhancement techniques were employed, making the text more legible for OCR.

Skew Correction: To correct any skew or tilt in the images, skew correction algorithms were applied, ensuring that the text lines were aligned vertically.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 4, April 2024

OCR Processing: Tesseract, an open-source OCR engine, was utilized to extract text from the preprocessed images. Tesseract's capability to handle various font styles and scripts ensured accurate text extraction, even in challenging image conditions. The extracted text was then stored in a structured format for further processing.

Page Layout Analysis: OpenCV, a computer vision library, was employed to perform page layout analysis on the extracted text. This involved segmenting the text into individual lines and words, identifying text blocks, and determining their spatial relationships. Page layout analysis provided a structured representation of the text, facilitating subsequent information extraction and analysis.

Keyword Extraction: To identify relevant information, keyword extraction techniques were applied to the extracted text. Given an input keyword, algorithms were employed to locate instances of that keyword within the text. This enabled the identification of specific pieces of information based on user-defined keywords.

Face Recognition: To associate extracted information with specific individuals, Face++'s face recognition API was employed. This involved identifying faces in the source images and matching them against a database of known individuals. Face recognition facilitated the association of extracted information with specific individuals, enhancing the relevance and richness of the extracted data.

The whole interface for this information extractor is developed in the Django framework.

Fig. 2 shows the flow of different steps involved in the working of Tesseract OCR.

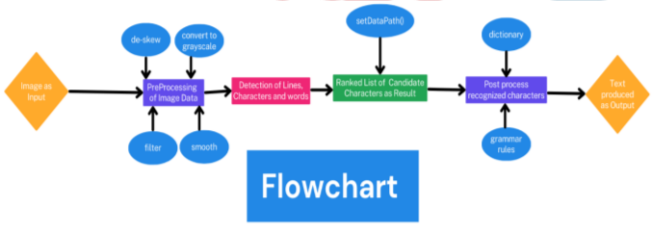

Fig. 2 Flowchart of Project

VII. RESULT

The The process of acquiring information from the source article requires accurately and effectively identifying info. This is done through the use of Tesseract and OpenCV for English and Hindi text purposes, respectively. The utilization of Face is necessary to detect and identify the mentioned individuals. Further information about these people is gathered from different sources and displayed as results through our module, presented visually in a web application built with Django. The obtained data has various uses in analyzing data, making it easier to identify trends and patterns, make comparisons between individuals, and generate reports. Face recognition outcomes are crucial in verifying the identity of people and ensuring the reliability and accuracy of the collected information. This method provides a simple and effective way to recognize and collect information on individuals discussed in the source file.

Important details about them, such as their appearance, particularly their facial characteristics, are provided along with a picture. a specific field. It is crucial to note that this system is designed to be easy to use and open to everyone, regardless of their expertise in a particular area.

Reliably identifying sources and expanding its usefulness in a wide range of scenarios, natural language processing and face recognition techniques have the potential to significantly improve journalism.

Utilize the mentioned topics in articles to their advantage.

Check the reliability of their sources and gain more knowledge about the people discussed in their articles. Moreover, scientists and academics can utilize it to easily recognize and gather data concerning individuals mentioned in scholarly articles or research papers.

Extracting text information from source:

Input Image :

Searched Information: Diplomacy

Output Image:

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 4, April 2024

Extracting image information from source:

Input Image :

Searched Image:

Output Image:

VIII. COMPARATIVE ANALYSIS

A. Successes and Challenges

Successful Integration:

Seamless integration of Tesseract OCR, OpenCV, and Face++ technologies contributed to the project's success

Accuracy Improvements:

Iterative fine-tuning of algorithms and continuous monitoring led to enhanced accuracy in text extraction and facial recognition.

Challenges in Name Detection:

Contextual precision in name detection presented challenges, with occasional false positives in non-contextual instances.

Real-Time Processing Considerations:

While real-time processing capabilities were not fully realized, optimizations were implemented to reduce processing times.

Ethical Considerations:

The project emphasized the importance of addressing potential biases in facial recognition outcomes and the ethical use of technology.

IX. CONCLUSION

In conclusion, the proposed approach of utilizing Tesseract, OpenCV, and Face++ for person identification from source data in English and Hindi offers a promising avenue for gathering pertinent information and gaining insights about individuals mentioned in articles. This implementation effectively automates the identification process and information collection by leveraging the combined capabilities of OCR, computer vision, and face recognition.

Despite its merits, it is crucial to acknowledge the potential limitations of this approach. OCR technology may encounter challenges in accurately extracting text from images, particularly when dealing with low-quality images or scripts not supported by the OCR software. Similarly, face recognition algorithms may face difficulties in identifying individuals accurately, especially when relying on poor-quality comparison images or individuals with distinctive facial features underrepresented in the training data.

Therefore, careful evaluation of the accuracy and reliability of the obtained results is essential, and this approach should be employed in conjunction with other methods and information sources. Additionally, the ethical considerations surrounding the use of this technology, particularly in terms of privacy and consent, must be carefully considered.

By addressing these limitations and ensuring responsible implementation, this approach holds the potential to significantly enhance the process of person identification and information gathering from newspaper articles, providing valuable insights for various applications.

X. FUTURE SCOPE

The future scope of information extraction from various sources using OCR, computer vision, and data recognition is vast and promising. These technologies have the potential to revolutionize the way we collect, analyze, and understand information from the vast and ever-growing data landscape.

A. Enhanced Content Understanding

These technologies will enable machines to understand the content of unstructured data, such as text, images, and videos, with greater depth and accuracy. This will allow for more sophisticated analysis, pattern recognition, and decision-making based on extracted information.

B. Improved Search and Retrieval

With the ability to extract and understand information from diverse sources, search engines and retrieval systems will become more powerful and nuanced. Users will be able to find relevant content more easily and efficiently, regardless of the format or source of the information.

C. Personalized Experiences

Information extraction technologies will enable the creation of personalized experiences for users across various applications. For instance, personalized recommendations, targeted marketing, and tailored content delivery will become more refined and effective.

D. Augmented Reality and Virtual Reality Integration

OCR, computer vision, and data recognition will play a vital role in enhancing augmented reality (AR) and virtual reality (VR) applications. These technologies will enable real-time information extraction from the physical world, allowing for seamless integration of digital content into the user's perception of reality.

E. Healthcare and Medical Applications

In the healthcare domain, these technologies will facilitate the extraction of valuable information from medical images, patient records, and clinical trials. This will aid in diagnosis, treatment planning, and drug discovery, leading to improved patient outcome.

REFERENCES

[1] Hamad, K., & Mehmet, K. (2016). A detailed analysis of optical character recognition technology. *International Journal of Applied Mathematics Electronics and Computers*, (Special Issue-1), 244-249.

[2] Harraj, A. E., & Raissouni, N. (2015). OCR accuracy improvement on document images through a novel pre-processing approach. arXiv preprint arXiv:1509.03456.

[3] Joshi, K. (2021). Study of Tesseract OCR. GLS KALP–Journal of Multidisciplinary Studies, 1(2), 41-51.

[4] Chakraborty, P., Rakib Mia, M., Sumon, H. K., Sarker, A., Imtiaz, A., Mahbubur Rahman, M., ... & Choudhury, T. (2022). Recognize Meaningful Words and Idioms from the Images Based on OCR Tesseract Engine and NLTK. In Pattern Recognition and Data Analysis with Applications (pp. 297-310). Singapore: Springer Nature Singapore.

[5] Hasan, R. T., & Sallow, A. B. (2021). Face Detection and Recognition Using OpenCV. Journal of Soft Computing and Data Mining, 2(2), 86-97.

[6] Lim, C. K., & Flayyih, O. H. (2019). Quality Analysis of Optical Character Recognition of Hindi Language Approaches. IIRJET, 5(2).

[7] Jana, R., Chowdhury, A. R., & Islam, M. (2014). Optical character recognition from text image. International Journal of Computer Applications Technology and Research, 3(4), 240-244.

[8] Misra, C., Swain, P. K., & Mantri, J. K. (2012). Text extraction and recognition from image using neural network. International Journal of Computer Applications, 40(2), 13-19.

[9] Karthikeyan, G., Bharanidharan, G., Jeevanandham, D., & Balaji, B. G. (2022). Text Recognition Images using OCR. Int. J. Prog. Res. Sci. Eng, 3, 57-60.

[10] Manwatkar, P. M., & Singh, K. R. (2015, January). A technical review on text recognition from images. In 2015 IEEE 9th International Conference on Intelligent Systems and Control (ISCO) (pp. 1-5). IEEE.

[11] Patel, C., Patel, A., & Patel, D. (2012). Optical character recognition by open source OCR tool tesseract: A case study. International Journal of Computer Applications, 55(10), 50-56.

[12] Kumar, A., Kaur, A., & Kumar, M. (2019). Face detection techniques: a review. Artificial Intelligence Review, 52, 927-948.

[13] Sun, Y., Ren, Z., & Zheng, W. (2022). Research on face recognition algorithm based on image processing. Computational Intelligence and Neuroscience, 2022.

[14] Singh, S., & Jasmine, S. G. (2019). Face recognition system. International Journal of Engineering Research & Technology (IJERT), 8(05), 2278-0181.

[15] BHSBIET, L., Kaur, J., & Singh, H. Face detection and Recognition: A review.