A

Mini Project Report on

Smart Keyword Extraction From Images

Submitted in the Partial Fulfillment of the
Requirements for the Award of the Degree of

BACHELOR OF TECHNOLOGY

in

Computer Science and Engineering (AI)

Submitted by

K.Jaswanth Reddy	**232P1A3146**
D.Reddy Kalyan	**232P1A3125**
I.Dileep Kumar Reddy	**232P1A3134**

Under the esteemed guidance
of
Dr.S.Bajid Vali
Professor,MTech,Phd

Department of Computer Science and Engineering (AI)

CHAITANYA BHARATHI INSTITUTE OF TECHNOLOGY
(Approved by AICTE, New Delhi & Affiliated to JNTUA, Ananthapuramu)
(Accredited by NAAC with "A" Grade and Accredited by NBA (CE, EEE, ECE, CSE))
(Recognized by UGC under section 2(f) and 12(b) of UGC Act, 1956)
VIDYA NAGAR, PALLAVOLU (V), PRODDATUR-516360, Y.S.R. (Dt.), A.P

2025-26

CHAITANYA BHARATHI INSTITUTE OF TECHNOLOGY

(Approved by AICTE, New Delhi & Affiliated to JNTUA, Ananthapuramu)
(Accredited by NAAC with “A” Grade and Accredited by NBA (CE, EEE, ECE, CSE))
(Recognized by UGC under section 2(f) and 12(b) of UGC Act, 1956) VIDYA NAGAR,
PALLAVOLU (V), PRODDATUR-516360, Y.S.R. (Dt.), A.P

Department of Computer Science and Engineering (AI)

CERTIFICATE

This is to certify that the project titled is **Smart Keyword Extraction From Images** carried out by

K.Jaswanth Reddy	**232P1A3146**
D.Reddy Kalyan	**232P1A3125**
I.Dileep Kumar Reddy	**232P1A3134**

in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in **Computer Science and Engineering (AI)** during the year 2025-26.

Signature of the Supervisor	**Signature of the HOD**
Dr.S,Bajid Vali	**Ms. Lakshmi Madhuri**
Professor,MTech,Phd	**HOD, CSE (AI)**

ACKNOWLEDGMENT

The satisfaction that accompanies the successful completion of the task would be put incomplete without the mention of the people who made it possible, whose constant guidance and encouragement crown all the efforts with success.

I wish to express my deep sense of gratitude to **Dr.S.Bajid Vali,Professor,MTech,Phd** and Project Supervisor, Department of Computer Science and Engineering, Chaitanya Bharathi Institute of Technology, for his able guidance and useful suggestions, which helped me in completing the project in time.

I am particularly thankful to **Ms. Lakshmi Madhuri**, Head of the Department, Department of Computer Science and Engineering (AI), her guidance, intense support and encouragement, which helped us to mould my project into a successful one.

I show gratitude to my honorable Principal **Dr. S. Sruthi** and Director Admin **Dr. G. Sreenivasula Reddy** for providing all facilities and support.

I avail this opportunity to express my deep sense of gratitude and heart-full thanks to **Sree V. Jaya Chandra Reddy**, Chairman and **Sree V. Lohit Reddy**, Director of CBIT, for providing a congenial atmosphere to complete this project successfully.

I also thank all the staff members of Computer Science and Engineering (AI) department for their valuable support and generous advice. Finally, thanks to all my friends and family members for their continuous support and enthusiastic help.

232P1A3146

232P1A3125

232P1A3134

ABSTRACT

This Python script leverages EasyOCR and Natural Language Toolkit (NLTK) libraries to perform automated text extraction and keyword analysis from images. Using EasyOCR, the script reads and extracts textual content from an input image. Subsequently, it processes the extracted text by tokenizing, filtering out common English stopwords and punctuation, and calculating word frequencies to identify the top keywords. The output includes the full extracted text alongside the most frequent keywords, providing a simple yet effective tool for extracting meaningful information from images containing text.

Table of Contents

CHAPTER 1 INTRODUCTION

Background:

With the increasing amount of digital data stored as images—such as scanned documents, photos of signs, or screenshots—automated extraction of textual information from images has become crucial in many applications, including document digitization, content analysis, and information retrieval. Optical Character Recognition (OCR) technology enables the conversion of text within images into machine-readable formats.However, raw OCR output often contains noise and irrelevant information, necessitating further text processing to extract meaningful insights. Natural Language Processing (NLP) techniques, such as tokenization, stopword removal, and frequency analysis, can help distill essential keywords from the extracted text, facilitating better indexing, summarization, and content understanding.

Objective of the Project:

The primary objective of this project is to develop an automated system that extracts textual information from images using Optical Character Recognition (OCR) and processes the extracted text to identify key terms and keywords. By integrating EasyOCR for accurate text extraction and NLTK for natural language processing, the project aims to provide a streamlined solution for converting image-based text into meaningful, concise insights.

Significance Of The Project:

This project addresses the growing need to efficiently extract and analyze textual information embedded in images, which is increasingly common in various fields such as document management, digital archiving, and data analysis. By automating text extraction and keyword identification, the system reduces manual effort, saves time, and minimizes errors associated with manual transcription.

CHAPTER 2

SYSTEM REQUIREMENTS

Hardware Requirements:

- **Processor:** Intel i3 or equivalent (for basic OCR tasks)
- **RAM:** Minimum 4 GB (8 GB recommended for faster processing)
- **Storage:** At least 500 MB free disk space for libraries and temporary files • **Graphics:** GPU not mandatory but can speed up EasyOCR if available

Software Requirements:

- **Operating System:** Windows, macOS, or Linux
- **Python:** Version 3.6 or higher
- **Libraries/Packages:**
    - easyocr (for Optical Character Recognition) ○ nltk (Natural Language Toolkit for text processing) ○ numpy (dependency for EasyOCR) ○ torch (PyTorch backend required by EasyOCR)
- **Additional:**
    - Internet connection (for initial download of NLTK data packages and PyTorch, if not pre-installed

CHAPTER 3

IMPLEMENTATION

The system first initializes the EasyOCR reader and downloads necessary NLTK resources. It takes an image file as input, verifies its existence, and extracts text via OCR. The text is then tokenized and cleaned by removing stopwords and punctuation. Word frequencies are calculated to identify the top keywords, which are displayed along with the extracted text to the user. This pipeline efficiently converts image text into valuable, concise information.

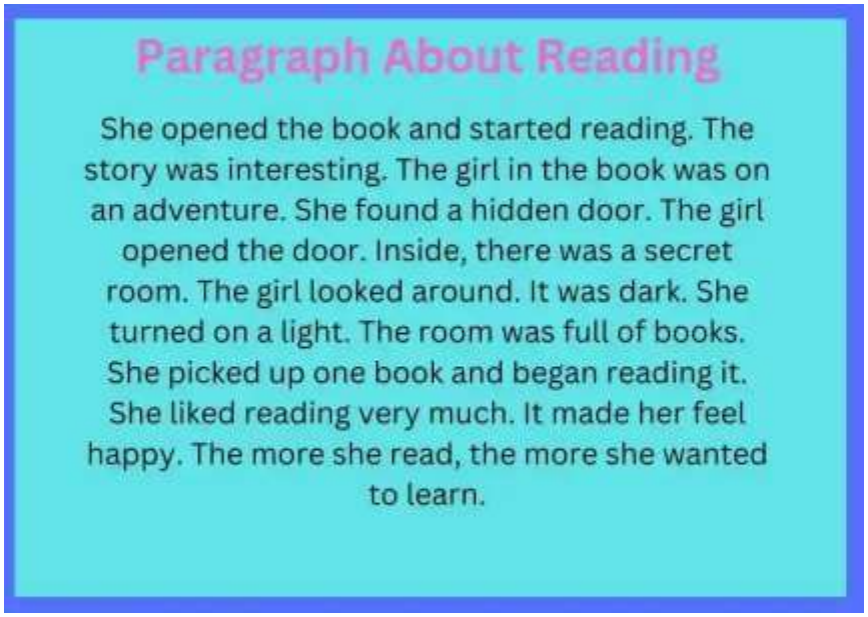
Source code:

```
import easyocr
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
import string
from collections import Counter

# Initialize OCR reader
reader = easyocr.Reader(['en'])

# Download stopwords and punkt tokenizer once
nltk.download('punkt')
nltk.download('stopwords')


stop_words = set(stopwords.words('english'))


def extract_keywords(text, top_n=10): words
    = word_tokenize(text.lower())
    filtered_words = [w for w in words if w not in stop_words and w not in string.punctuation]
    freq = Counter(filtered_words)
    return freq.most_common(top_n)


def main(image_path):
    # OCR
    image_path="C:/Users/Lenovo/Desktop/CVIP/b.webp"
    results = reader.readtext(image_path, detail=0)
    extracted_text = " ".join(results).strip()


    print("\n--- Extracted Text ---")
    print(extracted_text)
    if extracted_text:

        keywords = extract_keywords(extracted_text, top_n=10)

        print("\n--- Top Keywords ---")

        for word, freq in keywords:

            print(f"{word}: {freq}")

    else:

        print("No text detected in the image.")



if _name ___== "_main_":

    image_path = input("Enter image path: ")

    main(image_path)
```

CHAPTER 4

RESULTS

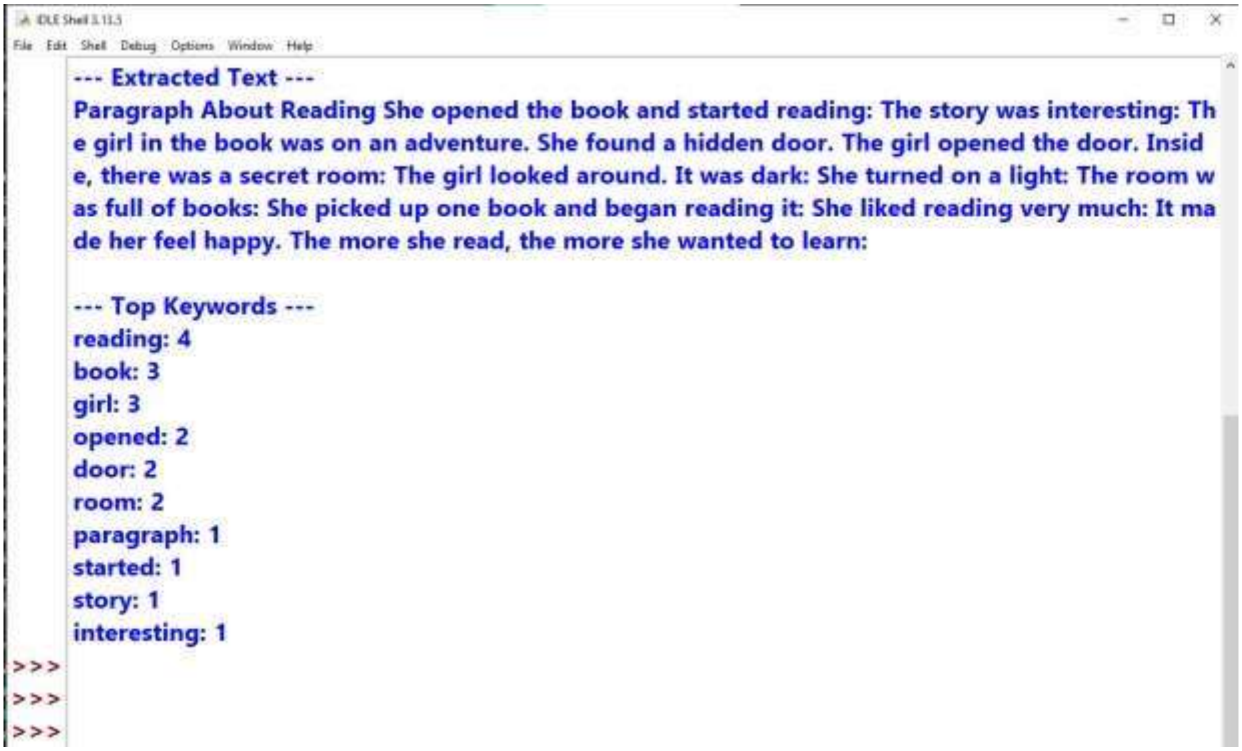
Input:



Paragraph About Reading

She opened the book and started reading. The story was interesting. The girl in the book was on an adventure. She found a hidden door. The girl opened the door. Inside, there was a secret room. The girl looked around. It was dark. She turned on a light. The room was full of books. She picked up one book and began reading it. She liked reading very much. It made her feel happy. The more she read, the more she wanted to learn.

Output:



```
--- Extracted Text ---
Paragraph About Reading She opened the book and started reading: The story was interesting: Th
e girl in the book was on an adventure. She found a hidden door. The girl opened the door. Insid
e, there was a secret room: The girl looked around. It was dark: She turned on a light: The room w
as full of books: She picked up one book and began reading it: She liked reading very much: It ma
de her feel happy. The more she read, the more she wanted to learn:

--- Top Keywords ---
reading: 4
book: 3
girl: 3
opened: 2
door: 2
room: 2
paragraph: 1
started: 1
story: 1
interesting: 1
>>>
>>>
>>>
```

REFERENCES

[1]. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. Advances in Neural Information Processing Systems (NeurIPS), 27, 2672–2680.

[2]. Li, S., & Deng, W. (2020). Deep Facial Expression Recognition: A Survey. IEEE Transactions on Affective Computing, 13(3), 1195–1215. https://doi.org/10.1109/TAFFC.2020.2981446

[3]. Karras, T., Laine, S., & Aila, T. (2019). A Style-Based Generator Architecture for Generative Adversarial Networks (StyleGAN). Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 4401–4410.

[4]. Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2017). AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. IEEE Transactions on Affective Computing, 10(1), 18–31.

[5]. Zhang, Z., Luo, P., Loy, C. C., & Tang, X. (2015). Learning Social Relation Traits from Face Images. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 3631–3639.