iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VII **Month of publication:** July 2022

DOI: https://doi.org/10.22214/ijraset.2022.45811

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VII July 2022- Available at www.ijraset.com

Identical Image Extraction from PDF Document Using LBP (Local Binary Patterns) and RGB (Red, Green and Blue) Color Features

Dr. Girish Kumar D[1], Vinod. K[2], G. Yogananda Reddy[3], Rakesh Kumar. M[4]
[1, 2, 3, 4]*Department of Computer Science and Engineering, Ballari Institute of Technology & Management, Visvesvaraya Technological University, 583104 Ballari, India.*

Abstract: ***Content based image retrieval (CBIR) has become a main research area in multimedia applications. In the literature, there is lot of papers focusing on the content-based image retrieval in order to extract the semantic information within the query concept. The content based image retrieval aims to find the similar images from PDF document against a query image. Generally, the similarity between the representative features of the query image and PDF images is used to rank the images for retrieval. In early days, various hand designed feature descriptors have been investigated based on the visual cues such as color, texture, shape, etc. that represent the images. In this project we have developed a method to extract the images present in PDF documents based on appropriate features extracted from query image such as color and local binary pattern features to find perfect matching image is present in PDF or not. Based on similarity analysis model we will display the query input image is present in PDF documents or not. If query matching image is present in the PDF documents then we will display the list corresponding PDF documents using suitable applications to show the presence.***

Keywords: Image extraction from PDF

I. INTRODUCTION

An image texture is a set of metrics calculated in image processing designed to quantify the perceived texture of an image. Image texture gives us information about the spatial arrangement of color or intensities in an image or selected region of an image [1]. Texture analysis is a technique for evaluating the position and intensity of signal features, that is, pixels, and their gray level intensities. The distribution of these pixels can be computed to produce mathematical parameters which characterize the texture type and thus the underlying structure of the objects shown in the image; these values are also known as texture features. Real world image textures are often not uniform due to many different variations such as illumination conditions and arbitrary spatial rotations constantly. Basically the methods of textures analysis are categories into four types [2]: (1) Structural, (2) Statistical, (3) Model based and (4) Transforms. Structural:-These methods represent texture by primitive patterns which are regular in appearance and systematically located on the surface. Statistical: - they represent the texture by nondeterministic properties of image pixels and regions which are usually natural and consist on randomly distributed surface elements

II. OBJECTIVES

1) To read and extract query image features for searching similar images in PDF documents.
2) To extract similar images present in PDF documents.
3) To display the PDF documents in which query image is present.

III. FUNCTIONALREQUIREMENTS

1) Allow the user provide query image.
2) The query image may be of .jpg or .png format.
3) Read query image properly.
4) Preprocess input image to resize, remove the noise etc. if necessary.
5) Extract image features for finding similarity images present in PDF document.
6) Display the matching images present in PDF document.
7) Display appropriate PDF document.
8) Display the appropriate message for unsuccessful search.

IV. NON-FUNCTIONALREQUIREMENTS

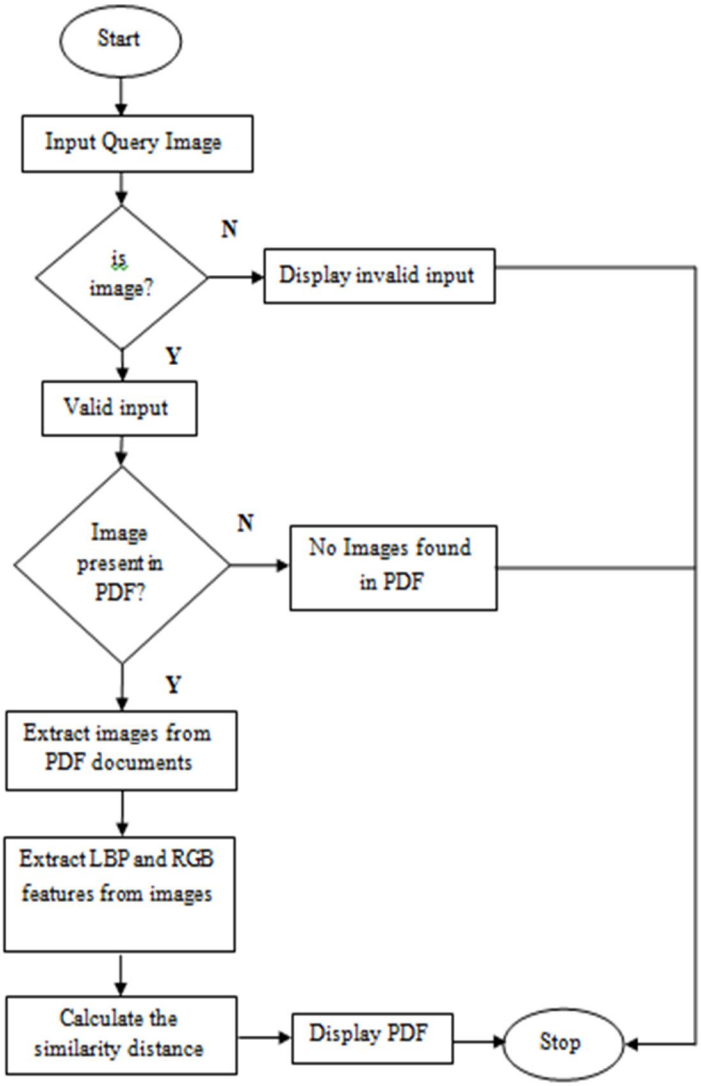
1) The read and display operations should be efficient.
2) The feature extraction module and similarity matching module may take less time.
3) The implemented code shall execute in different versions of platform.
4) The project should yield reliable results.

V. DESIGN

This section details the implementation of proposed system. The proposed system contains five models. They are given below.

1) Input module.
2) Data-preprocessing module.
3) Feature extraction.
4) Image similarity detection module
5) Output model

Compare the distance between the input query image features and extracted PDF image features
And we retrieve the corresponding PDF document and display its contents.

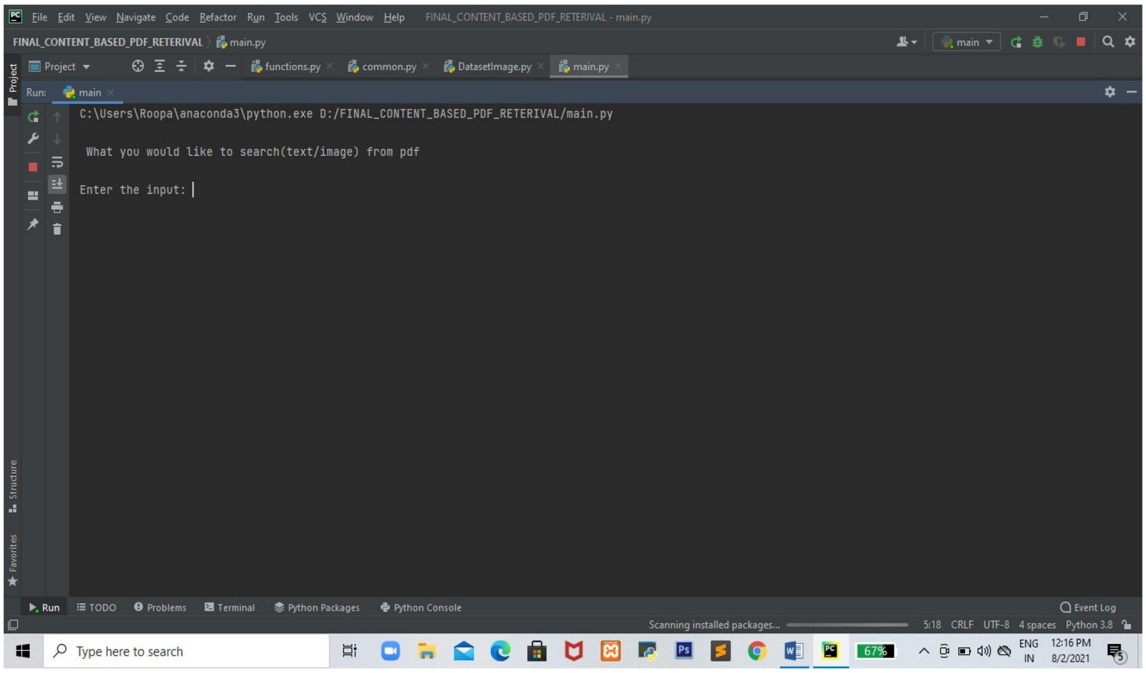


International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VII July 2022- Available at www.ijraset.com

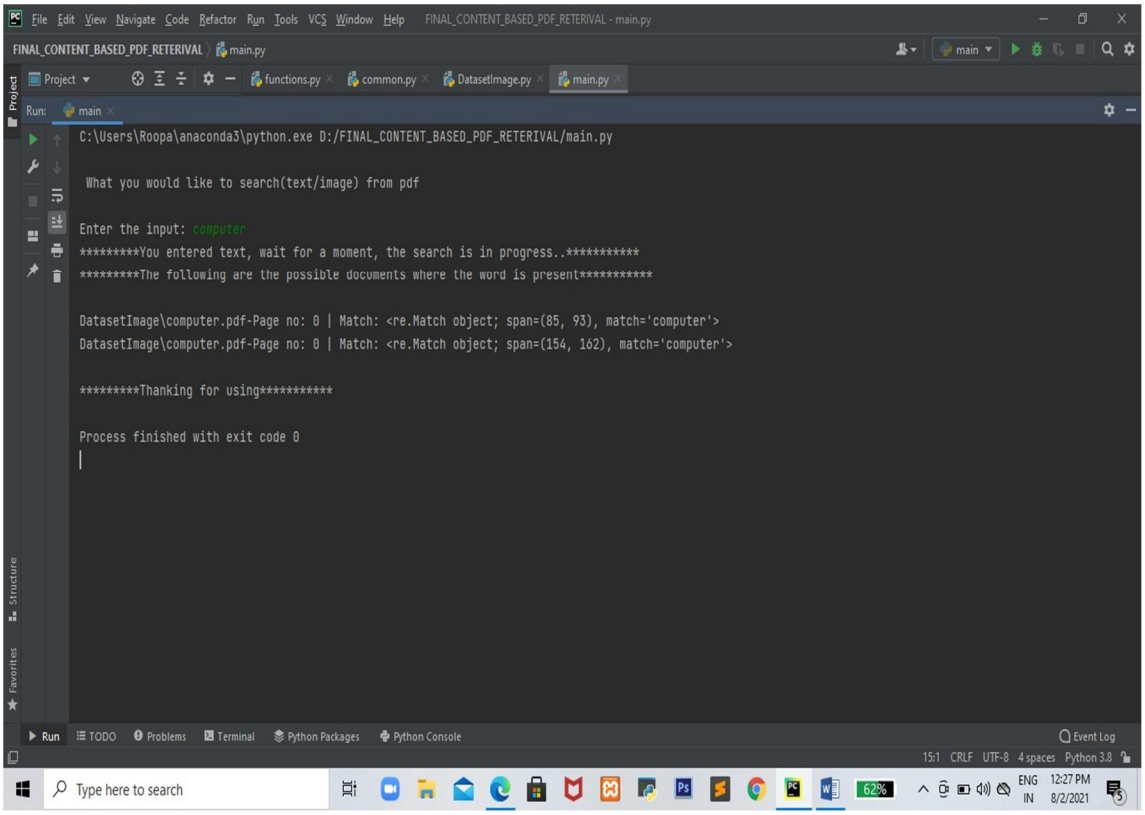
VI. RESULTS

Initially, the user is asked to enter a path of an image which is to be searched from the PDF documents. Project will dynamically identify whether the given input is of type image path/text/invalid path.

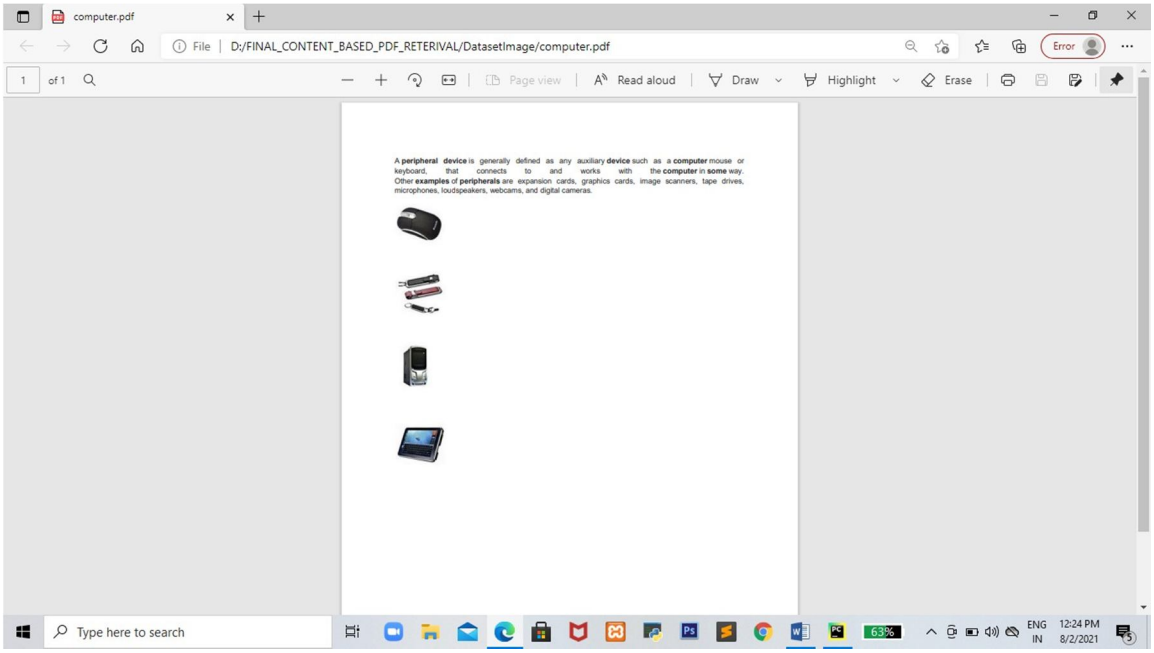
A. Enter The Query To Be Searched In Pdf File



B. Processing Query Image



C. Displaying The Output

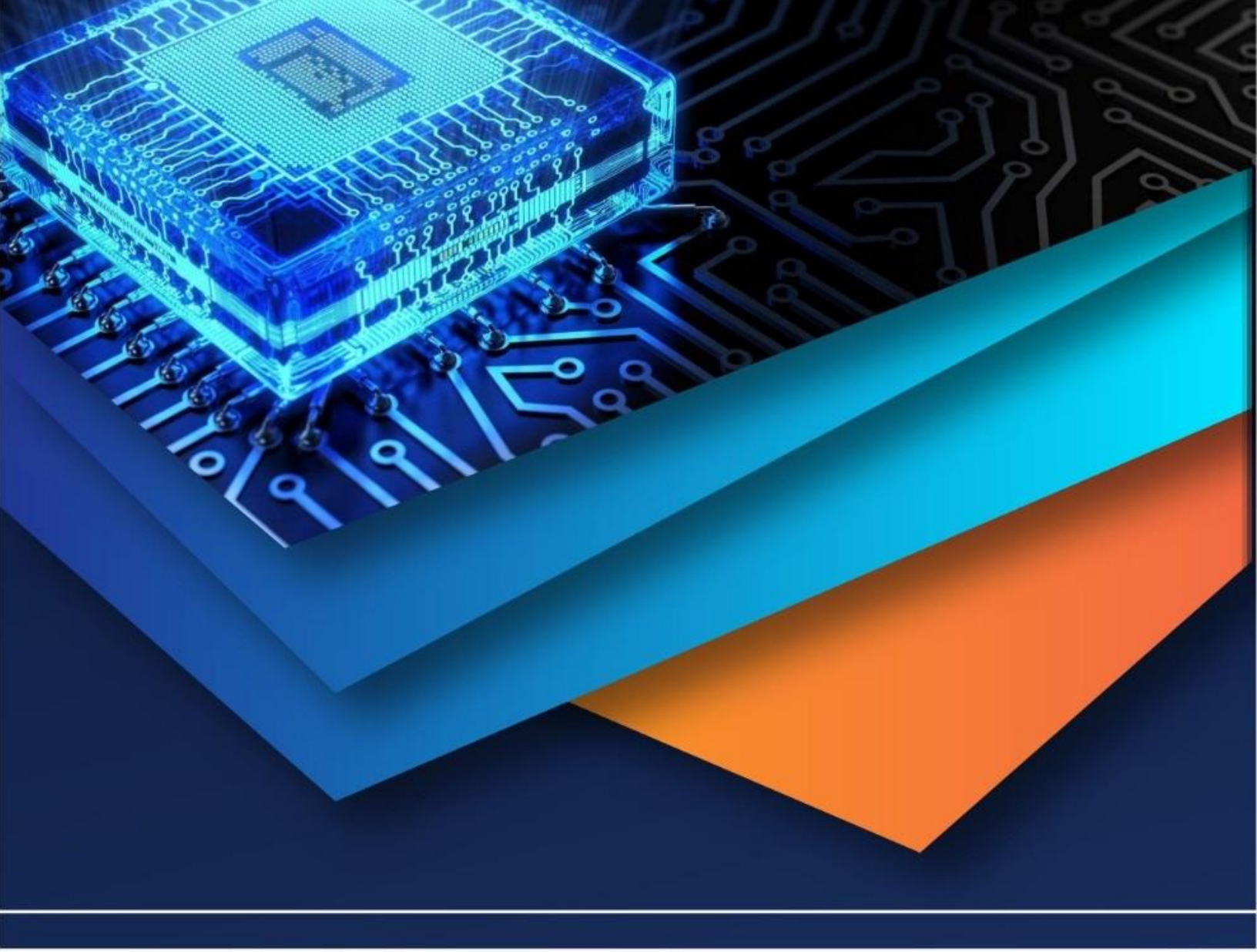


VII. CONCLUSION

The main purpose of this project is to provide an efficient and truly realizable approach for retrieving PDF document based on similar images present in PDF. We first performed the query image preprocessing to remove the noise and enhance its features. Then, employed LBP and Color feature extraction techniques to extract significant features of query image and formed a feature descriptor. We used Euclidean distance computation technique to match feature descriptor of input query image and extracted images from PDF document in order to find the similarity between the images to retrieve the documents accurately. As per the testing results the proposed system retrieve the correct PDF documents based on input query image supplied as an input. However, the proposed system does not detect if the input images are blurred and distorted. So in future we would like to enhance this system to address the issue and also make the system to for multiple images.

REFERENCES

[1] D. Zhang, M. M. Islam, and G. Lu, "A review on automatic image annotation techniques," Pattern Recognition, vol. 45, no. 1, pp. 346–362, 2012.
[2] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma, "A survey of content-based image retrieval with high-level semantics," Pattern Recognition, vol. 40, no. 1, pp. 262–282, 2007.
[3] T. Khalil, M. U. Akram, H. Raja, A. Jameel, and I. Basit, "Detection of glaucoma using cup to disc ratio from spectral domain optical coherence tomography images," IEEE Access, vol. 6, pp. 4560–4576, 2018.
[4] S. Yang, L. Li, S. Wang, W. Zhang, Q. Huang, and Q. Tian, "SkeletonNet: a hybrid network with a skeleton-embedding process for multi-view image representation learning," IEEE Transactions on Multimedia, vol. 1, no. 1, 2019.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY