

# **1.INTRODUCTION**

The aim of this project is to extract text from images using python. It efficiently reads text from images. We import all the required libraries (tkinter, tesseract, opencv). Python will automatically find and extract text from an image. In this, we will learn how to extract text content from images using openCV and tesseract. Tesseract is an open-source engine for optical character recognition (OCR). It efficiently reads text from images and is very easy to use. As mentioned earlier it is open source so it is free to use. For text detection and extraction project development, first we install required libraries. Provide the location of the tesseract.exe file. Tkinter provides GUI functionalities: open an image dialog box so user can upload an image . Let's jump to the extract function which takes the path of the image as a parameter .In this function, we will read the image using cv2.imread. We will also resize the image so that we can get well-formatted output for all different sizes of input images. Tesseract works on RGB images and opencv reads an image as BGR image, so we need to convert the image and then call Tesseract functions on the image. Here, the conversion is done using cv2.cvtColor(). We have stored height, width, and thickness of the input image using img.shape for later use. After the pre-processing, call image\_to\_data() function of tesseract which returns a string of extracted text from the image. Print the whole string for better understanding. The string is a multiline string, where each line contains extracted text but its first line (starting from zero) contains headings that are not useful for us, so we will skip the very first line. Now, split the string to get the extracted text and finally print the extracted text on the screen.

## **1.1 Product Scope**

This project currently extracts the text from image. We can include more options like extracting the numerals, special characters etc. The project only supports the English language and not numerals and special characters. This project can be made more accurate by extracting each character or every text in an image no matter what font it is. The project can be made efficient if it extracts texts of each font and any calligraphy the text is written.

## **2. LITERATURE SURVEY**

### **2.1 Existing System**

There are various extractors in the market day to day that provide different types of services to the user using different methods. Till date there is no such text extractor that helps users in extracting their texts from image. Applications have been developed that system automatically extract text from the uploaded image. Text extractor have been developed with similar features but with ease. This system uses opencv for image processing and pytesseract for extraction.

### **2.2. PROPOSED SYSTEM**

The aim of this project is to extract a text from image which is uploaded by the user. Some knowledge has been embedded into the machine so that it identifies the texts in the image and extracts it out. This system can be used by any user who want to extract the text from an image for entertainment.

#### **2.2.1 Advantages**

- This system will help the user by detecting the text in an image.
- The detected text is extracted by the system automatically and displayed to the user.
- Image processing using opencv provides efficiency.
- Usage of pytesseract in the system helps to extract the texts of some fonts effectively.
- In future reference we are trying to implement our system that will cover all the calligraphy styles as well as numerals and special characters.

### **3. SYSTEM REQUIREMENT**

#### **3.1 Hardware Requirements**

- Processor: Intel and above, AMD A6 and above
- Speed: 3 GHz and above
- RAM: 4Gb
- Disk Space: 50Gb
- Keyboard: Standard Windows Keyboard
- Monitor: 5VGA

#### **3.2 Software Requirements**

- Operating system: Windows 10
- Software Used: Anaconda
- Data science tool: pytesseract, numpy, cv2, tkinter, PIL
- Script: Python
- Interpreter: Python 3.8 or above
- Code Editor: Jupiter Notebook/Pycharm

## **4.LANGUAGE SPECIFICATIONS**

### **Technology/Platform/Tools used**

#### **Technology:**

- Python
- Anaconda

#### **Data Science tools:**

- Numpy
- Tensorflow
- Tkinter
- OpenCV
- pytesseract
- PIL

#### **Platform /code editor:**

- Jupiter Notebook / PyCharm

### **PYTHON**

Python is an interpreter, object-oriented, high-level programming language with dynamic semantics. It's high-level built-in data structures, combined with dynamic typing and dynamic binding; make it very attractive for Rapid Application Development. Python is simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form

without charge for all major platforms, and can be freely distribute. Often, programmers fall in love with Python because of the increased productivity it provides. Since there is no compilation step, the edit-test-debug cycle is incredibly fast. Debugging Python programs is easy: a bug or bad input will never cause a segmentation fault. Instead, when the interpreter discovers an error, it raises an exception. When the program does not catch the exception, the interpreter prints a stack trace. A source level debugger allows inspection of local and global variables, evaluation of arbitrary expressions, setting breakpoints, stepping through the code a line at a time, and so on. The debugger is written in Python itself, testifying to Python's introspective power. On the other hand, often the quickest way to debug a program is to add a few print statements to the source: the fast edit-test-debug cycle makes this simple approach very effective.

#### FEATURES IN PYTHON:

- i. **Easy to code:** Python is a high-level programming language. Python is very easy to learn the language as compared to other languages like C, C#, JavaScript, Java, etc. It is very easy to code in python language and anybody can learn python basics in a few hours or days. It is also a developer-friendly language.
- ii. **Free and Open Source:** Python language is freely available at the official website and you can download. Since it is open-source, this means that source code is also available to the public. So you can download it as, use it as well as share it.
- iii. **Object-Oriented Language:** One of the key features of python is Object- Oriented programming. Python supports object-oriented language and concepts of classes, objects encapsulation, etc.
- iv. **GUI Programming Support:** Graphical User interfaces can be made using a module such as PyQt5, PyQt4, wxPython, or Tk in python. PyQt5 is the most popular option for creating graphical apps with Python.

- v. **High-Level Language:** Python is a high-level language. When we write programs in python, we do not need to remember the system architecture, nor do we need to manage the memory.
- vi. **Extensible feature:** Python is a **Extensible** language. We can write us some Python code into C or C++ language and also we can compile that code in C/C++ language.
- vii. **Python is Portable language:** Python language is also a portable language. For example, if we have python code for windows and if we want to run this code on other platforms such as Linux, Unix, and Mac then we do not need to change it, we can run this code on any platform.
- viii. **Python is Integrated language:** Python is also an Integrated language because we can easily integrated python with other languages like C, C++, etc.
- ix. **Interpreted Language:** Python is an Interpreted Language because Python code is executed line by line at a time. Like other languages C, C++, Java, etc. there is no need to compile python code this makes it easier to debug our code. The source code of python is converted into an immediate form called byte code.
- x. **Large Standard Library:** Python has a large standard library, which provides a rich set of module and functions so you do not have to write your own code for everything. There are many libraries present in python for such as regular expressions, unit-testing, web browsers, etc
- xi. **Dynamically Typed Language:** Python is a dynamically typed language. That means the type (for example- int, double, long, etc.) for a variable is decided at run time not in advance because of this feature we don't need to specify the type of variable.

## ANACONDA

Anaconda is a distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, large-scale data processing, predictive analytics, etc.), that aims to simplify package management and deployment. The distribution includes data-science packages suitable for Windows, Linux, and macOS. It is developed and maintained by Anaconda, Inc., which was founded by Peter Wang and Travis Oliphant in 2012. Package versions in Anaconda are managed by the package management system conda. This package manager was spun out as a separate open-source package as it ended up being useful on its own and for things other than Python. There is also a small, bootstrap version of Anaconda called Miniconda, which includes only conda, Python, the packages they depend on, and a small number of other packages.

### BENEFITS OF USING ANACONDA:

- It is free and open-source
- It has more than 1500 Python/R data science packages
- Anaconda simplifies package management and deployment
- It has tools to easily collect data from sources using machine learning and AI
- It creates an environment that is easily manageable for deploying any project
- Anaconda is the industry standard for developing, testing and training on a single machine
- It has good community support- you can ask your questions there
- User level install of the version of python you want
- Able to install/update packages completely independent of system libraries or admin privileges
- Conda tool installs binary packages, rather than requiring compile resources like pip - again, handy if you have limited privileges for installing necessary libraries.
- More or less eliminates the headaches of trying to figure out which version/release of package X is compatible with which version/release of package Y, both of which are required for the install of package Z

- Comes either in full-meal-deal version, with numpy, scipy, PyQt, spyder IDE, etc. or in minimal / a la carte version (miniconda) where you can install what you want, when you need it
- No risk of messing up required system libraries

## **NUMPY**

NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more. At the core of the NumPy package, is the ndarray object. This encapsulates n-dimensional arrays of homogeneous data types, with many operations being performed in compiled code for performance. There are several important differences between NumPy arrays and the standard Python sequences.

## **TENSOR FLOW**

Tensor Flow is an open source framework developed by Google researchers to run machine learning, deep learning and other statistical and predictive analytics workloads. Like similar platforms, it's designed to streamline the process of developing and executing advanced analytics applications for users such as data scientists, statisticians and predictive modelers. The Tensor Flow software handles data sets that are arrayed as computational nodes in graph form. The edges that connect the nodes in a graph can represent multidimensional vectors or matrices, creating what are known as tensors. Because Tensor Flow programs use a data flow architecture that works with generalized intermediate results of the computations, they are especially open to very large-scale parallel processing applications, with neural networks.



## **TKINTER**

Tkinter is an open source, portable graphical user interface (GUI) library designed for use in Python scripts. Tkinter relies on the Tk library, the GUI library used by Tcl/Tk and Perl, which is in turn implemented in C. Therefore, Tkinter can be said to be implemented using multiple layers. Tkinter is a Python library used for creating and developing GUI-based applications. It is completely open-source which works on Windows, Mac, Linux, and Ubuntu.

## **OPENCV**

OpenCV is the huge open-source library for the computer vision, machine learning, and image processing and now it plays a major role in real-time operation which is very important in today's systems. By using it, one can process images and videos to identify objects, faces, or even handwriting of a human. When it is integrated with various libraries, such as NumPy, Python is capable of processing the OpenCV array structure for analysis. To identify image patterns and its various features, we use vector space and perform mathematical operations on these features. An open-source machine learning and computer vision software, OpenCV is a free-to-use library. The platform offers general infrastructure to fast-track machine perception's use in commercially available creations by offering general infrastructure for computer vision applications. Besides, OpenCV operates under the BSD license, which imposes minimal restriction on library's use. This way, commercial entities can tweak the code to suit their purposes. OpenCV offers access to over 2,500 algorithms to be used for deployment of various machine learning and computer vision capabilities like object identification and facial recognition.

### **FUNCTIONALITIES OF USING OPENCV:**

- Image/video I/O, processing, display (core, imgproc, highgui)
- Object/feature detection (objdetect, features2d, nonfree)
- Geometry-based monocular or stereo computer vision (calib3d, stitching, videostab)
- Computational photography (photo, video, superres)

- Machine learning & clustering (ml, flann)
- CUDA acceleration (gpu)

## **PYTESSERACT**

Python-tesseract is an OCR library that is used to scan and transcribe any textual data in images. This library is used to recognize textual information but not to save it to any text document. The advantages of using pytesseract are:

- Information of pytesseract can be readable with high degree of accuracy. Flatbed scanners are very accurate and may produce reasonably top quality images.
- Processing of pytesseract information is fast. Large quantities of text are often input quickly.
- A paper based form are often became an electronic form which is straightforward to store or send by mail.
- It is cheaper than paying someone amount to manually enter great deal of text data. Moreover it takes less time to convert within the electronic form.
- The latest software can re-create tables also as original layout.
- This process is much faster as compared to the manual typing the information into the system.
- Advanced version can even re-create tables, columns and even produce sites.

## **PIL**

Python Imaging Library is a free and open-source additional library for the Python programming language that adds support for opening, manipulating, and saving many different image file formats. It is available for Windows, Mac OS X and Linux. Pillow offers several standard procedures for image manipulation. These include:

- per-pixel manipulations,
- masking and transparency handling,
- image filtering, such as blurring, contouring, smoothing, or edge finding,
- image enhancing, such as sharpening, adjusting brightness, contrast or color, adding text to images and much more

## **5. SYSTEM ANALYSIS**

### **5.1 PROPOSED METHODOLOGY:**

The system proposed has 3 phases:

- Data and Requirement Collection
- Image preprocessing
- Feature extractions

#### **Data and Requirement Collection:**

This phase is the collection of requirements and data which are needed for the system for the extraction of text from the image which is uploaded by the user in Tkinter GUI.

#### **Image Preprocessing:**

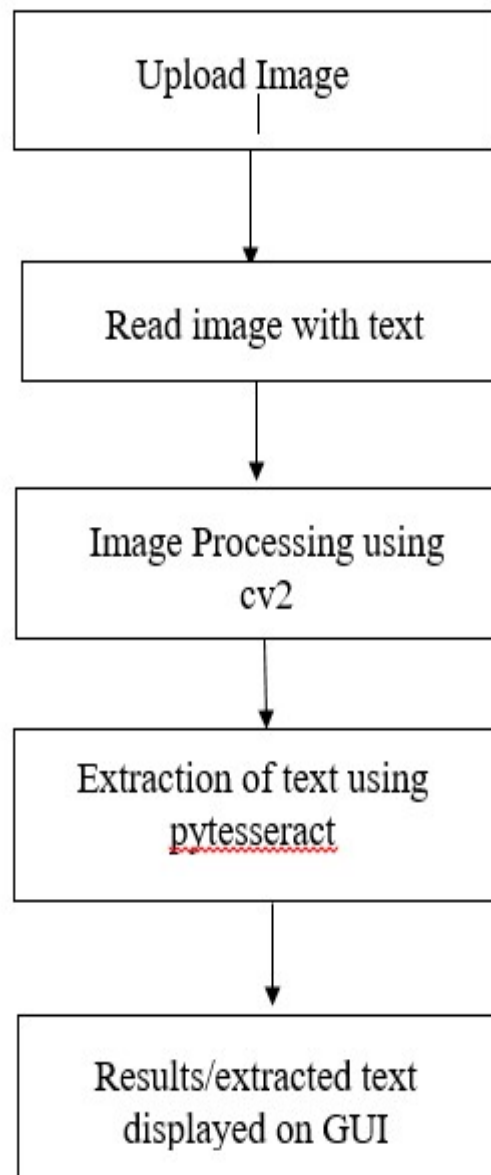
In this phase, the image which is uploaded by the user is preprocessed. Image processing is a method to perform some operations on an image, in order to get an enhanced image and or to extract some useful information from it. Image processing is basically signal processing in which input is an image and output is image or characteristics according to requirement associated with that image. Image processing basically includes the following three steps:

- Importing the image
- Analysing and manipulating the image
- Output in which result can be altered image or report that is based on image analysis

#### **Feature Extraction:**

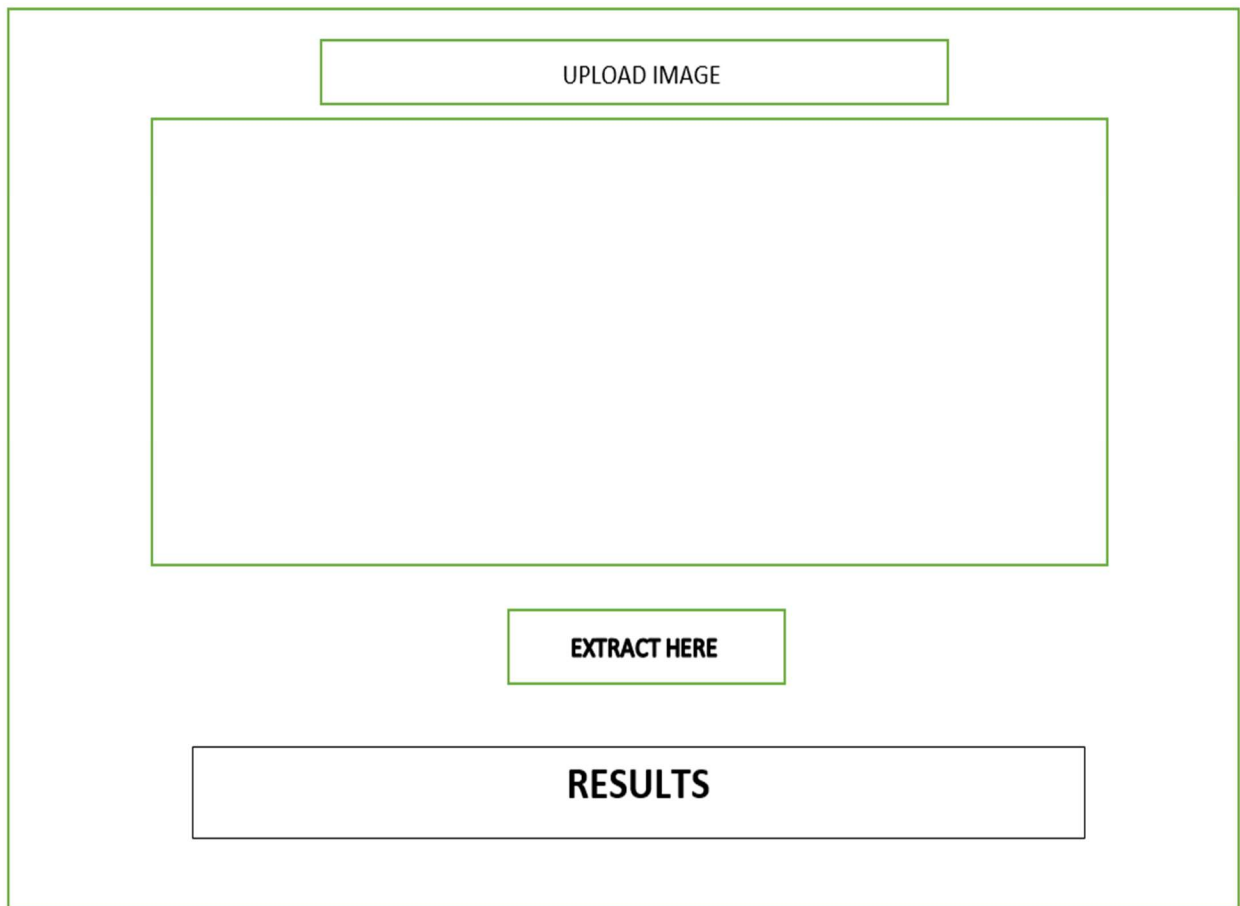
In this phase we are extracting the text from the image preprocessed. Texts are extracted from the image uploaded by user using pytesseract which is a wrapper for tesseract-ocr in python.

## **6. WORK FLOW DIAGRAM**



The working of this system is such that the user uploads an image in the tkinter GUI opened. The system reads the image using OpenCV and then it is preprocessed. Next the system detects the texts in the image uploaded and then it extracts the text using Pytesseract and displays the result in the GUI.

## **7. PROJECT OUTLOOK**



## **8. METHODOLOGY**

Agile method proposes incremental and iterative approach to software design. The agile process is broken into individual models that designers work. The customer has early and frequent opportunities to look at the product and make decision and changes to the project.

- Agile model is considered unstructured compared to the waterfall model.
- Small projects can be implemented very quickly. For large projects, it is difficult to estimate the development time.
- Error can be fixed in the middle of the project.
- Documentation attends less priority than software development.
- Every, iteration has its own testing phase. It allows implementing regression testing every time new functions or logic are released
- In agile testing when an iteration end, shippable features of the product is delivered to the customer. New features are usable right after shipment. It is useful when you have good contact with customers.
- Testers and developers work together.
- At the end of every sprint, user acceptance is performed
- It requires close communication with developers and together analyses requirements and planning.

Agile is not only about applying the set practices in developing a software. It also brings in a change in the team's mind-set, which drives them towards building a better software, working together and eventually landing them a happy customer. Agile values and principles enable the team to shift their focus and change their thought process of building a better software.

### **ROLES**

This project presents the design of an expert system for extraction of text from images.

## **SCRUM**

SCRUM is an agile development method which concentrates specifically on how to manage tasks within a team-based development environment. Basically, scrum is derived from activity that occurs during a rugby match. Scrum believes in empowering the development team and advocates working in small teams. The outcome is a version of scrum that is unique and specific, in order to have a process that works for us. Scrum is part of the agile movement. Agile is a response to the failure of the dominant software development project management paradigms and borrows many principles from lean manufacturing. The agile manifesto placed a new emphasis on communication and collaboration, functioning software, team self-organization, and the flexibility to adapt to emerging business realities. Scrum's early advocates were inspired by empirical inspect and adapt feedback loops to cope with complexity and risk. Scrum emphasizes decision making from real world results rather than speculation. Time is divided into short work cadences, known as sprints, typically one week or two weeks long. The product is kept in a potentially shippable state at all times. Scrum is a simple set of roles, responsibilities and meetings that never change. It consists of three roles, and their responsibilities are explained as follows:

### **SCRUM MASTER**

- Role played by Mrs. Sreeja K.
- Master is responsible for setting up the team, sprint meeting and removes obstacles to progress.
- Helps everyone involved understand and embrace the scrum values, principles, and practices.
- As a facilitator, scrum master helps the team resolve issues and make improvements  
Its use of scrum.
- He is responsible for protecting the team from outside interference.
- He takes a leadership role in removing impediments that inhibit team Productivity.
- Acts as a coach, providing development process leadership.



## **PRODUCT OWNER**

- Role played by Mrs. Jayanthi T.
- The product owner creates product backlog, prioritizes the backlog and is responsible for the delivery of the functionality at each iteration.
- He maintains and communicates to all other participants a clear vision of what the scrum team is trying to achieve.
- He is the only authority responsible for what will be developed and in what order.
- One of the most important responsibilities of product owner is to manage product backlog.

## **SCRUM TEAM**

Team manages its own work and organizes the work to complete the sprint or cycle. The team is a self-organizing and cross-functional group of people who do the hands-on work of developing and testing the product. Since the team is responsible for producing the product, it must also have the authority to make decisions about how to perform the work. The team is therefore self-organizing: team members decide how to break work into tasks, and how to allocate tasks to individuals, throughout the Sprint. The team size should be kept in the range from five to nine people, if possible (a large number make communication difficult, while a smaller number leads to low productivity and fragility) Note: A very similar term, “scrum Team,” refers to the team plus the scrum master and product owner.

## **SPRINT**

The sprint backlog is a list of tasks identified by the scrum team to be completed during the scrum sprint. During the sprint planning meeting, the team selects some number of products backlog items, usually in the form of user stories, and identifies the tasks necessary to complete each user stories. Most teams also estimate how many hours each task will take someone on the team to complete. It's critical that the team selects the items and size of the sprint backlog. Because they are the people committing to completing the tasks, they must be the people to choose what

they are committing to during the scrum sprint. The sprint backlog is commonly maintained as a spreadsheet, but it is also possible to use your defect tracking system or any of a number of software products designed specifically for scrum or agile. The sprint backlog makes visible all the work that the Development Team identifies as necessary to meet the sprint goal. To ensure continuous improvement, includes at least one high priority process improvement identified in the previous Retrospective meeting. The sprint backlog is a plan with enough detail that changes in progress can be understood in the daily scrum. The Development Team modifies the sprint backlog throughout the sprint. Only the Development Team can change its sprint backlog during a sprint. The sprint backlog is a highly visible, real time picture of the work that the Development Team.

## **9. PRODUCT BACKLOG**

<b>ID</b>	<b>USER STORIES</b>	<b>PRIORITY</b>	<b>COMMENT FROM SCRUM MASTER</b>	<b>COMMENT FROM PRODUCT OWNER</b>
1	Problem definition: <ul style="list-style-type: none"> <li>Extraction of text from images</li> </ul>	Very high		
2	Requirement Analysis: <ul style="list-style-type: none"> <li>PYTHON 3.8</li> <li>ANACONDA</li> <li>JUPITER NOTEBOOK</li> </ul>	High		
3	Create GUI based On TKINTER libraries	Very High		
4	Creating a path so as to open a file directory	High		
5	Creating an upload button in tkinter GUI	High		
6	User can choose the image from folder	High		
7	User can upload a image	High		
8	Reading image using cv2.imread()	Very High		
9	Import required libraries	High		
10	Formatting the image	High		

11	Creating an extract button for text extraction	High		
12	Detection of text from the image	Very high		
13	Extraction of text from the image	High		
14	Display the text extracted on the GUI	Very high		

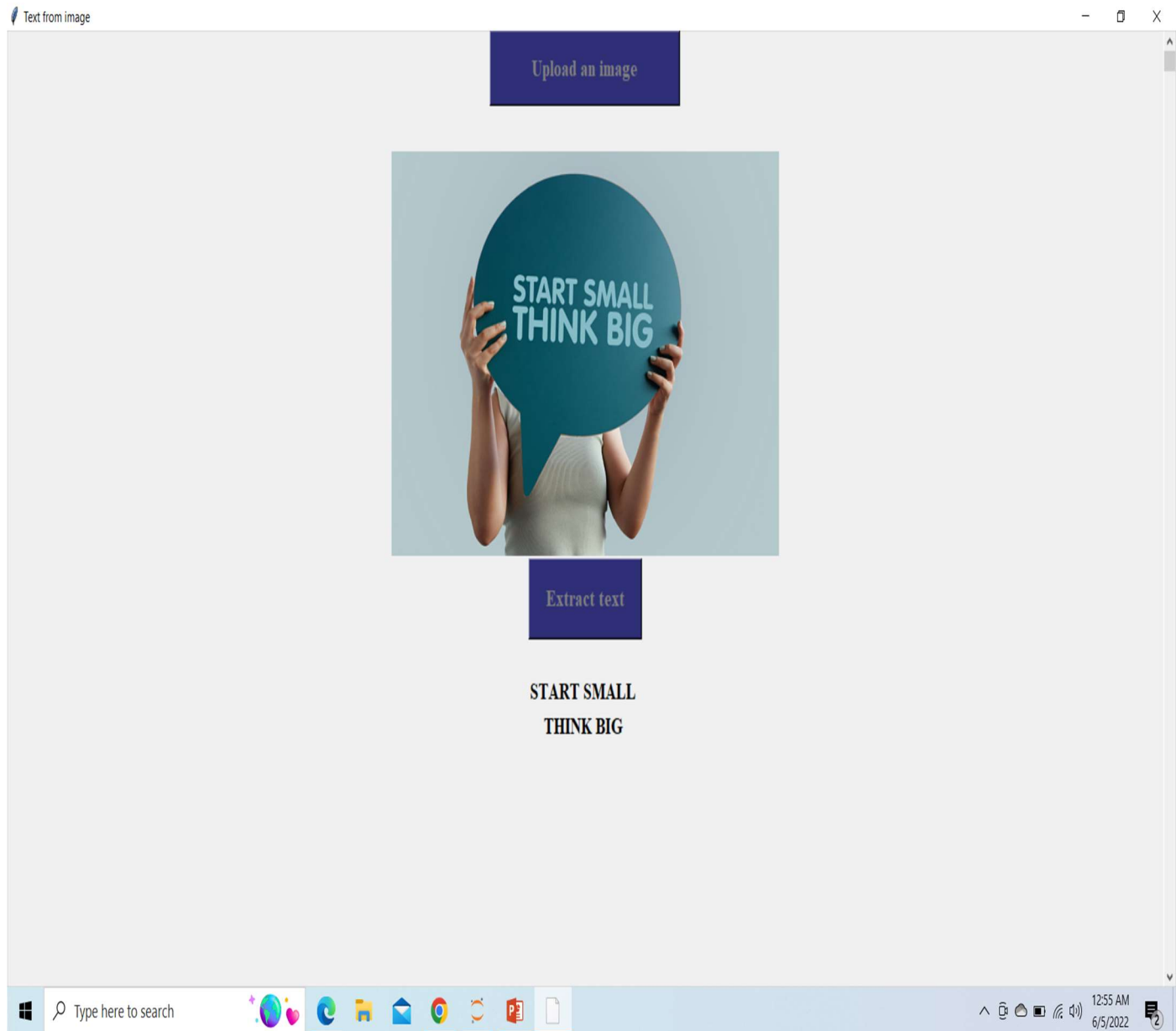
## **10. SPRINT BACKLOG**

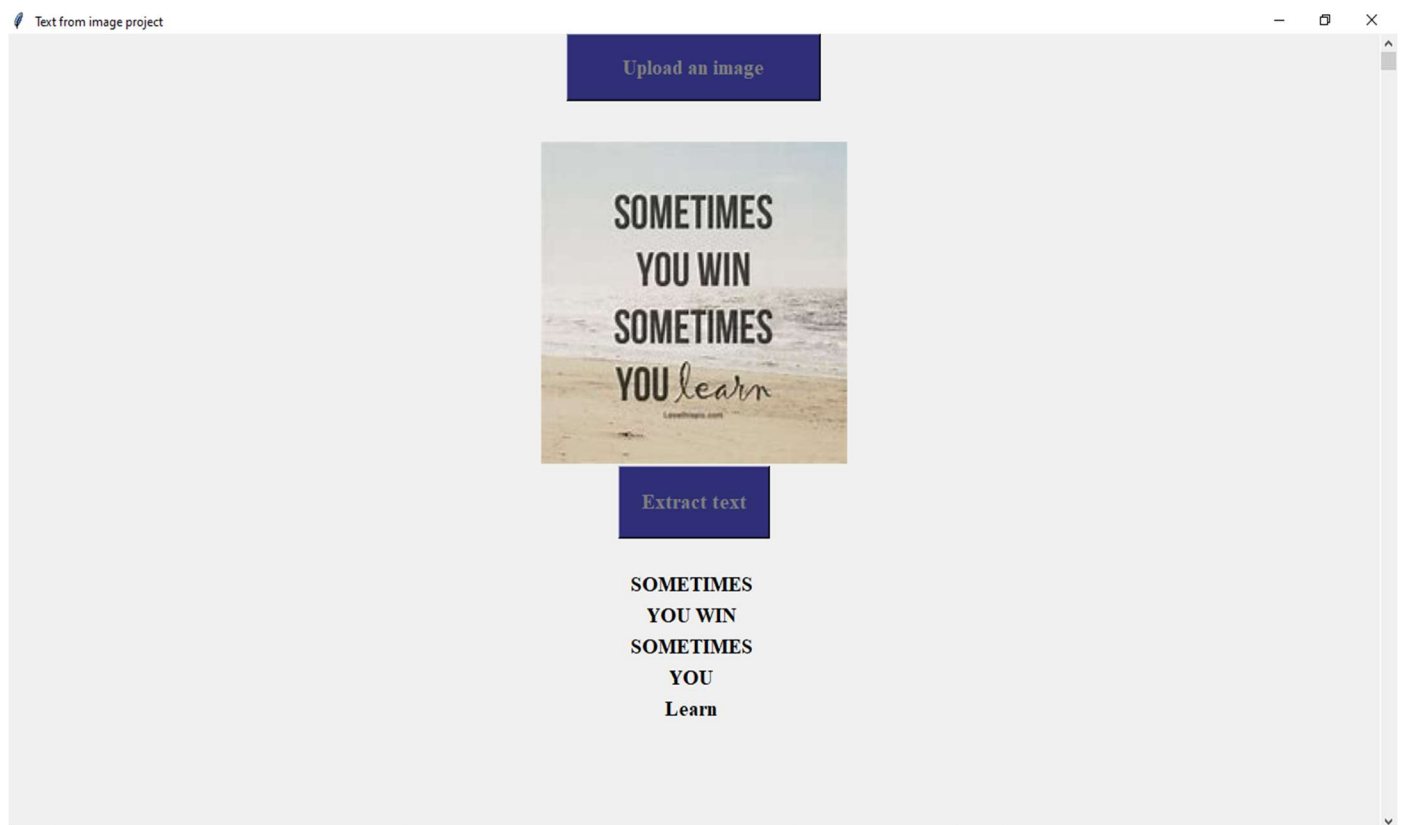
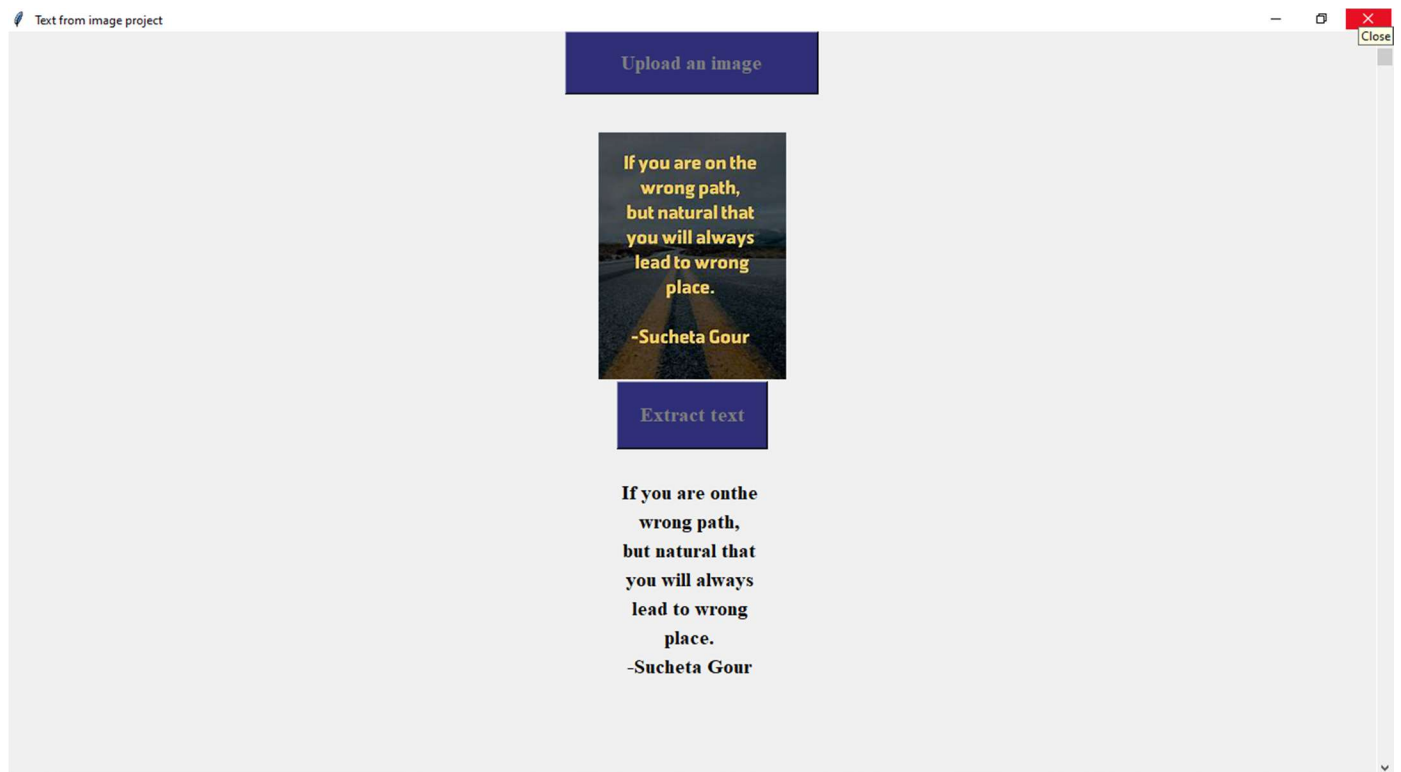
<b>Activity</b>	<b>Not Started</b>	<b>In Progress</b>	<b>Completed</b>
Requirement Collection			Completed
Image Processing			Completed
Feature Extraction			Completed
Output/Results			Completed
GUI			Completed
Documentation			Completed

## **11. SCRUM BOARD**

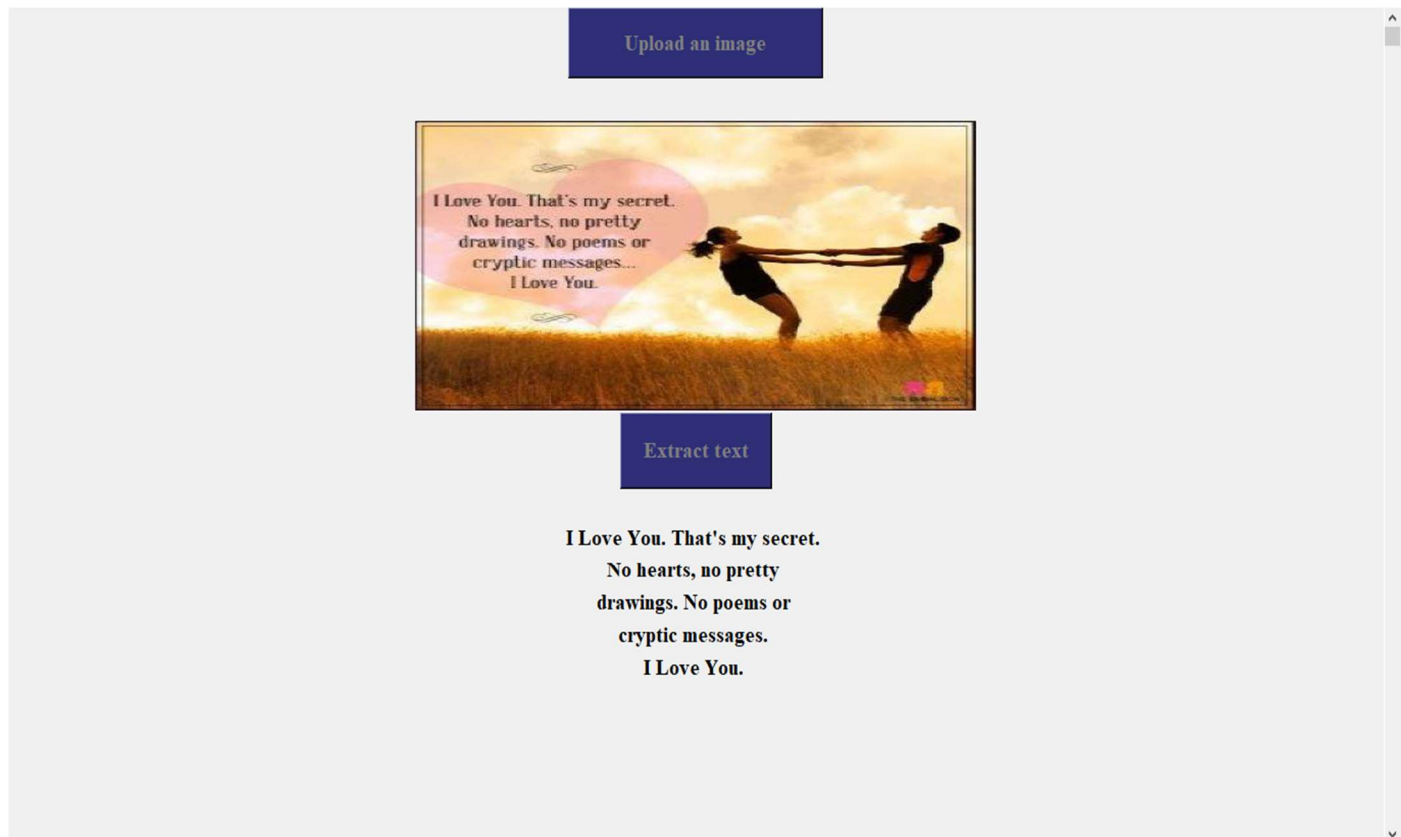
<b>SL.NO</b>	<b>TASK/SPRINT</b>	<b>START</b>	<b>FINISH</b>	<b>DURATION</b>	<b>STATUS</b>
1	SPRINT- 1 Problem identification	12/04/2022	19/04/2022	8 Days	Done
2.	SPRINT- 2 Data Collection	20/04/2022	09/05/2022	28 Days	Done
3.	SPRINT- 3 GUI	10/05/2022	17/05/2022	8Days	Done
4	SPRINT- 4 Image Processing	18/05/2022	10/06/2022	24 Days	Done
5	SPRINT-5 Feature Extractions	06/06/2022	25/06/2022	20 Days	Done
6	SPRINT- 6 Documentation	26/06/2022	30/06/2022	5 Days	Done

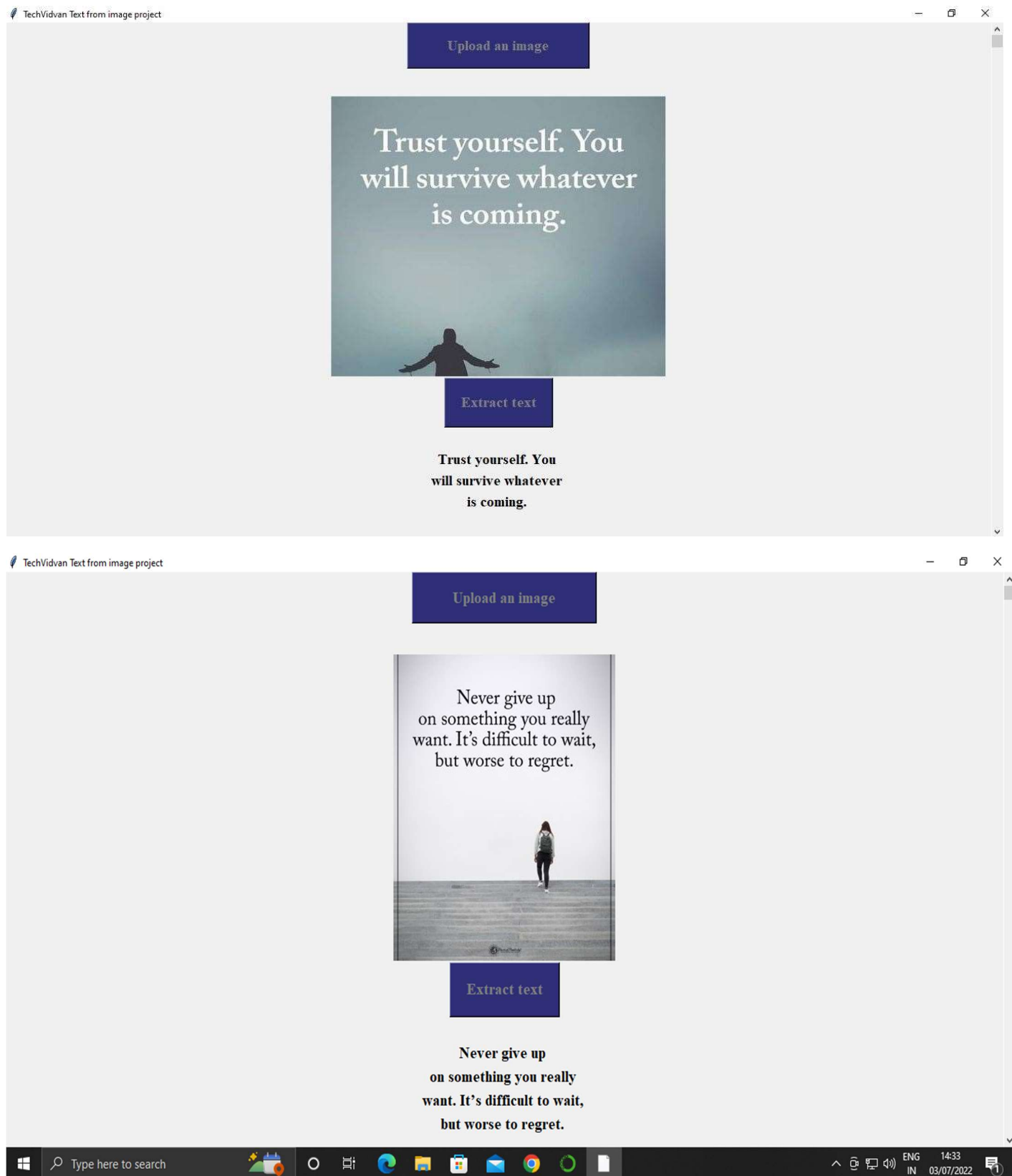
## 12. SCREENSHOTS

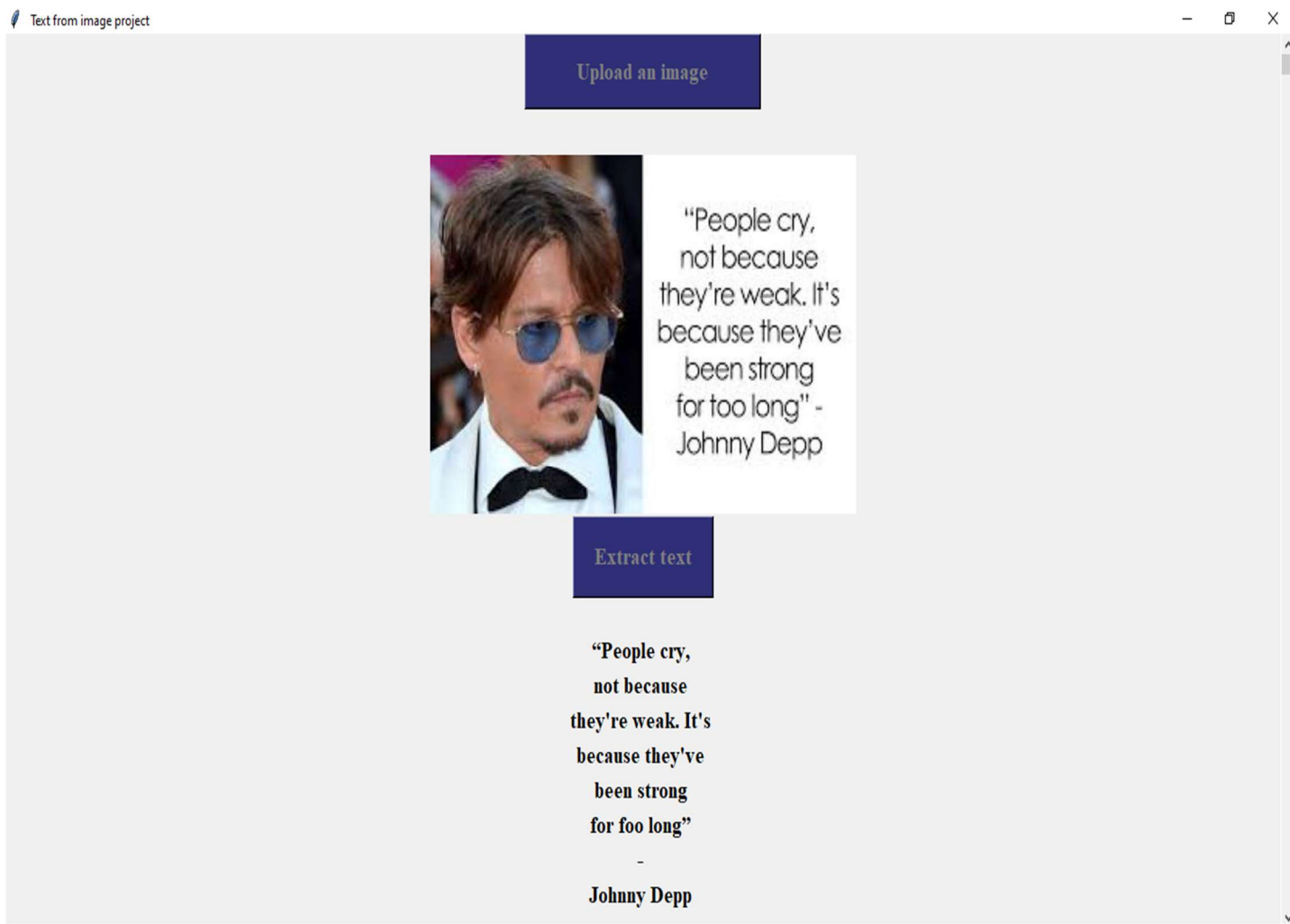












## **13. CONCLUSION**

This project is to extract text from images using python. It efficiently reads text from images. We import all the required libraries (tkinter, tesseract, opencv). Python will automatically find and extract text from an image. In this, we will learn how to extract text content from images using openCV and tesseract. Tesseract is an open-source engine for optical character recognition (OCR). It efficiently reads text from images and is very easy to use. As mentioned earlier it is open source so it is free to use. For text detection and extraction project development, first we install required libraries. Provide the location of the tesseract.exe file. Tkinter provides GUI functionalities: open an image dialog box so user can upload an image. Now, split the string to get the extracted text and finally print the extracted text on the screen. In this, we will read the image using cv2.imread. We will also resize the image so that we can get well-formatted output for all different sizes of input images. Tesseract works on RGB images and opencv reads an image as BGR image, so we need to convert the image and then call Tesseract functions on the image. Here, the conversion is done using cv2.cvtColor(). We have stored height, width, and thickness of the input image using img.shape for later use. After the pre-processing, call image\_to\_data() function of tesseract which returns a string of extracted text from the image. Print the whole string for better understanding. The string is a multiline string, where each line contains extracted text but its first line (starting from zero) contains headings that are not useful for us, so we will skip the very first line. Now, split the string to get the extracted text and finally print the extracted text on the screen.

## **14.REFERENCES**

- <https://www.ijtsrd.com/computer-science/simulation/2501/text-extraction-from-image-using-python/tgnana-prakash>
- M. Flickner, H. Sawney et al., Query byImage and Video Content: The QBIC System, IEEE Computer 28 (9) (1995) 23-32.
- J. Zhang, Y. Gong, S. W. Smoliar, and S. Y.Tan, Automatic Parsing of News Video, Proc.of IEEE Conference on Multimedia Computing and Systems, 1994, pp. 45-54.
- M. H. Yang, D. J. Kriegman, and N. Ahuja, Detecting faces in Images: A Survey, IEEE Transactions on Pattern Analysis and MachineIntelligence, 24 (1) (2002) 34-58.
- Y. Cui and Q. Huang, Character Extraction ofLicense Plates from Video, Proc. of IEEE Conference on Computer Vision and PatternRecognition, 1997, pp. 502 –507.
- C. Colombo, A. D. Bimbo, and P. Pala, Semantics in Visual Information Retrieval, IEEE Multimedia, 6 (3) (1999) 38-53.
- T. Sato, T. Kanade, E. K. Hughes, and M. A. Smith, Video OCR for Digital News Archive, Proc. of IEEE Workshop on Content basedAccess of Image and Video Databases, 1998, pp. 52-60.
- Atsuo Yoshitaka and Tadao Ichikawa, A Survey on Content-based Retrieval forMultimedia Databases, IEEE Transactions on Knowledge and Data Engineering, 11(1999) 81-93.
- W. Qi, L. Gu, H. Jiang, X. Chen, and H. Zhang, Integrating Visual, Audio, and Text Analysis for News Video, Proc. of IEEEInternational Conference on ImageProcessing, 2000, pp. 10-13.
- D. Wactlar, T. Kanade, M. A. Smith, and S. M. Stevens, Intelligent Access to Digital Video: The Informedia Project, IEEE Computer, 29(5) (1996) 46-52.
- H. Rein-Lien, M. Abdel-Mottaleb, A. K. Jain, Face Detection in Color Images, IEEE Transactions on Pattern Analysis and MachineIntelligence, 24 (5) (2002) 696-706.
- <https://nanonets.com/blog/ocr-with-tesseract/>
- <https://www.geeksforgeeks.org/opencv-overview/>

- <https://www.geeksforgeeks.org/advantages-and-disadvantages-of-optical-character-reader-ocr/>
- <https://www.geeksforgeeks.org/text-localization-detection-and-recognition-using-pytesseract/>
- <https://www.width.ai/post/the-best-ways-to-extract-text-from-images-without-tesseract-python>
- [https://opencv24pythontutorials.readthedocs.io/en/latest/py\\_tutorials/py\\_imgproc/py\\_table\\_of\\_contents\\_imgproc/py\\_table\\_of\\_contents\\_imgproc.html](https://opencv24pythontutorials.readthedocs.io/en/latest/py_tutorials/py_imgproc/py_table_of_contents_imgproc/py_table_of_contents_imgproc.html)
- [https://www.tutorialspoint.com/python/python\\_gui\\_programming.htm#:~:text=Tkinter%20is%20the%20standard%20GUI,to%20the%20Tk%20GUI%20toolkit.](https://www.tutorialspoint.com/python/python_gui_programming.htm#:~:text=Tkinter%20is%20the%20standard%20GUI,to%20the%20Tk%20GUI%20toolkit.)