IDENTIFYING A PROPER IMAGE CAPTIONING APPROACH TO FACILITATE INTERPRETING AN IMAGE IN A DOCUMENT

Project Id: 2022-024

Sanduni Madara P.G.
IT19392172

B.Sc. (Hons) Degree in Information Technology
(Specialization in Software Engineering)

Department of Computer Science and Software Engineering
Sri Lanka Institute of Information Technology
Sri Lanka

October 2022

IDENTIFYING A PROPER IMAGE CAPTIONING APPROACH TO FACILITATE INTERPRETING AN IMAGE IN A DOCUMENT

Project Id: 2022-024

Sanduni Madara P.G.
IT19392172

The dissertation was submitted in partial fulfilment of the requirements
for the B.Sc. Special Honors degree in Information Technology (Specialization in Software Engineering)

Department of Computer Science and Software Engineering
Sri Lanka Institute of Information Technology
Sri Lanka

October 2022

DECLARATION

I declare that this is my own work, and this thesis does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other university or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Name	Student ID	Signature
Sanduni Madara P.G.	IT19392172	

The above candidate is carrying out research for the undergraduate Dissertation under my supervision.

Signature of the supervisor   Date

…………………………..   ……………..…………

(Dr. Anuradha Jayakody)

ABSTRACT

The print disability prevents a person from obtaining information from printed material in the traditional manner and necessitates the use of alternative methods of accessing the information. Vision impairments, blindness, physical dexterity issues, learning disabilities, brain injuries, cognitive impairments, and literacy difficulties can all lead to print disability. According to current statistics, approximately there are more than 2.2 billion people worldwide are visually impaired or blind, which means those having print disabilities. Not only the blind or visually impaired, but also the ones with low literacy, may have difficulty reading paper documents. Hence, they have the same need as everyone else to have access to all types of information for the same reasons: leisure, education, employment, and so on, it is critical to remove impediments to their needs. In a word, they also have the feeling of being an active part of the society we live in. There have been technological advancements in scanning and reading printed materials using a variety of software and apps. Existing applications are incapable of reading equations, images, and tables as accurately as sighted people. Therefore, an innovative, effective procedure is needed to be designed to fulfill their rights. So, this study especially considers presenting an improved smart assistant which provides audio assistance to navigate through a smart assistant which functions auto focused image capturing, reading mathematical equations and table data of printed materials, classifying selected text or paragraph, images, graphs and reading aloud generated digitized text.

Keywords: Vision impairment, Print disability, Printed material, Smart assistant, Print disabled individuals

ACKNOWLEDGEMENT

First and foremost, I want to convey my heartfelt thanks to Dr. Anuradha Jayakody, my supervisor, for all of his encouragement, advice, and support during the research process, as well as the insightful remarks he provided to help the research project succeed. Aside from my supervisor, I'd like to thank Ms. Shashika Lokuliyana, the co-supervisor of this research project, for always being willing to help. This research required the perseverance of many people. Their invaluable input and time supported me in achieving ideal accomplishment over the term of my educational attempt. In addition, I'd like to thank the CDAP team for assisting us in successfully completing our project. Then, I'd like to express my heartfelt appreciation to my friends for their unfailing support during this endeavor. In addition, I'd want to express my gratitude to my parents for their constant support and encouragement. A special thanks to the other researchers that collaborated with me to make this research more effective and successful.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

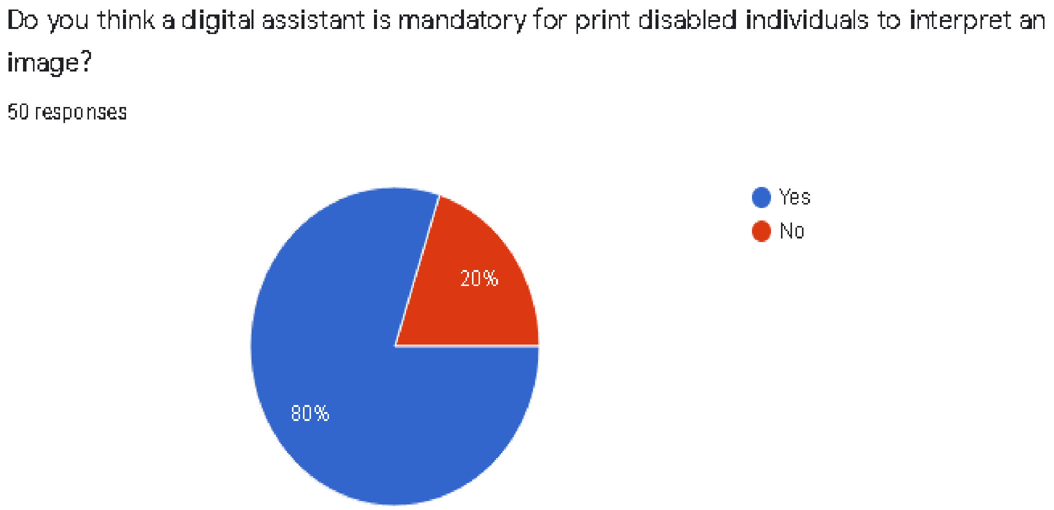
Abbreviation	**Description**
CNN	Convolutional Neural Network
YOLO	You Only Look Once
CelebA	CelebFaces Attributes
API	Application Programming Interface
OS	Operating System

1. INTRODUCTION

Vision, the most powerful of our senses, is essential in every aspect of our lives. Without that, it is difficult for us to interact with the rest of the world on a daily basis. Those who live long enough will be affected by at least one eye condition. When considering the world population, at least 2.2 billion people worldwide have a vision impairment or a blindness. From these, at least one billion cases, or about half of all these cases of vision impairment could have been avoided if they had been treated earlier [1]. To overcome this problem, even there are some accessible mediums as a solution for this, still there are some barriers. Due to copyright and other laws, most textual information can't be translated into those accessible media. This is a huge problem for the disabled people when it comes with their daily life, education and literacy. Braille is the most widely used tactile communication technique for visually impaired and sighted individuals. However, due to the high expense of braille, it is not accessible to every single person [2]. So, in some circumstances, their only option for interpreting printed materials is to depend on a third party, which will be examined more in this study. Globally, the concept of a smart assistant or digital assistant is gaining traction. One of the most important and difficult tasks in developing such applications is creating a user interface that is suitable for visually impaired users, both in terms of providing input and interpreting output feedback. Today, mobile devices have become the standards for the implementation of assistive technologies to help people with physical and cognitive disabilities. They are increasing the capacity and sensor capabilities, as well as standard possibilities for touch-based input and auditory-tactile output. But still a productive assistant has not been introduced for the vision impaired people. So, this solution will implement a smart assistant which provides audio assistance to navigate through a smart assistant which functions auto focused image capturing, reading mathematical equations and table data of printed materials, classifying selected text or paragraph, images, graphs and reading aloud generated digitized text. And under this research component, author will mainly be examined more on image captioning.

1.1. Background & Literature Survey

Image captioning can be done in many ways, but when it comes to blindness, images should be described in such a way that a person who has been blind since birth can understand. According to the findings from past studies, we can see that this component of the research has also already been researched by much expertise. Various tools have been built to interpret images in printed documents, and several assistive tools have also been implemented for print disabled people. However, majority of them are missing some critical factors that should be improved for use by print disabled people. According to the referred information, an assistant is mandatory for print disabled individuals to interact with the rest of the world on a daily basis. As shown in Figure 1.1-1 which denotes the data retrieved by the survey implemented, it is validated that most of the people said that a digital assistant is mandatory for the individuals with print disabilities to interpret an image. According to the paper [3], it depicts that they are in a level of lack of effective resources to fulfil their needs and rights to live their lives. Millions of people are harmed by this global reading imbalance. Therefore, it is mandatory to have a proper digital assistant to fulfill their needs and rights.

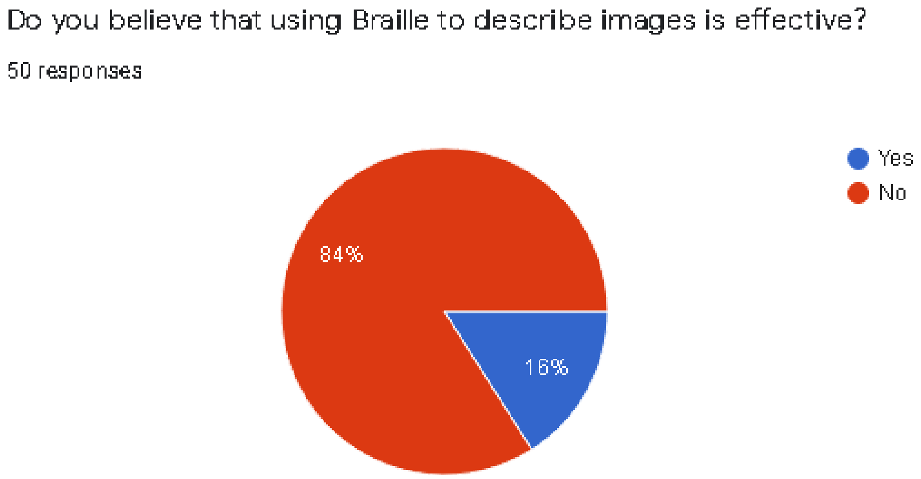


Do you think a digital assistant is mandatory for print disabled individuals to interpret an image?

50 responses

Figure 1.1.1 - Survey responses on the importance of a digital assistant

However, as mentioned earlier Braille is the way print disabled individuals are reading and writing, it is a tactile writing system. In that case those individuals cannot get more effective or accurate information on what they want to read. Especially when related to images in printed documents. As shown in the survey, Figure 1.1-2 depicts it too. Some had commented on some reasons saying that braille is not always helpful for vision impaired people to identify an image. They have said that when it comes with images it does not accurately give the sufficient information to the user. According to the survey's retrieved commentary information, it is clear that even the blind requires alternative access or accessible formats, and braille is not the best solution for them to gain information from printed materials. This alternative access can be a human or an assistive tool. But people can't be around every time for you to get help in interpreting a printed document. And when it comes to a legal or a confidential document, in that case trusting a person and involving them into those is a risk. So, it depicts that, smart assistant is very important for a print disabled person to interpret printed materials.



Do you believe that using Braille to describe images is effective?

50 responses

Figure 1.1.2 - Survey results on the effectiveness of braille for image interpretation

Accordance with the collected information by the survey, the pie chart in Figure 1.1-3 indicates that most respondents suggest that using an assistive tool is easier for a print-disabled person to interpret an image in a printed document, interact with the rest of the world on a daily basis while the lowest percentage of response is for using braille method.

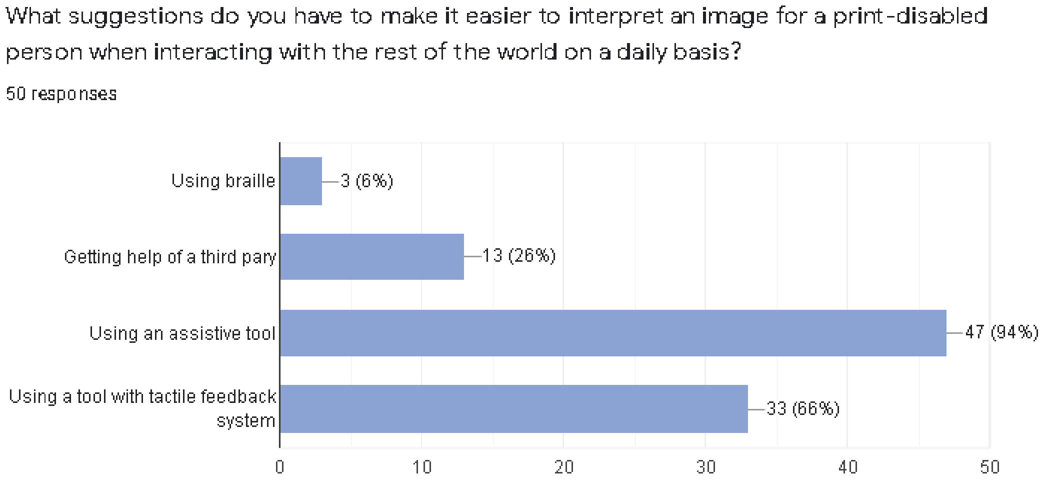


Figure 1.1.3 - Survey results on the suggestions for vision impaired people to interpret images

Furthermore, technological advancements have allowed for the scanning and reading of printed materials using a variety of software and apps. Existing applications are incapable of doing all reading equations, images, and tables as accurately as sighted people [4]. Focusing with the existing digital assistants, as shown in the Figure 1.1-4, mandatory responses are for low accuracy. Many say that the available tools are not giving an accurate interpretation of images in printed materials. The literature surveys of paper [5] highly depicts how existing similar systems function, as well as their strengths and weaknesses.

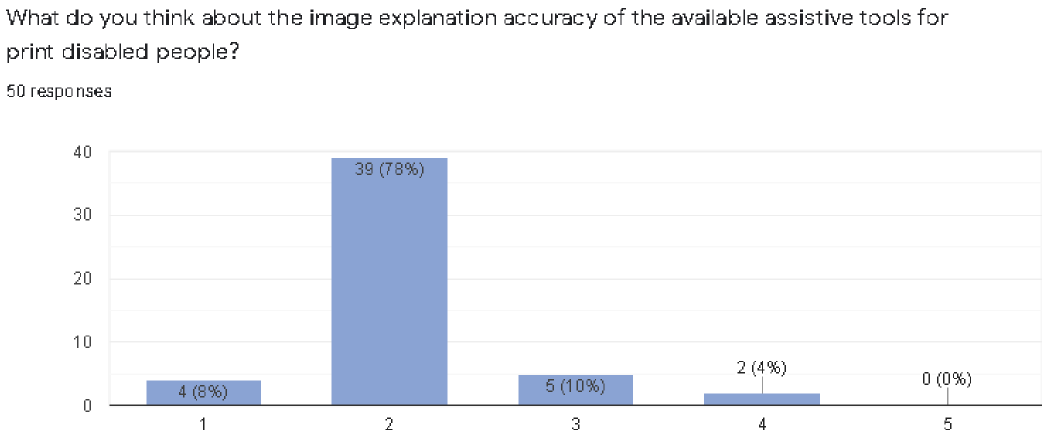


Figure 1.1.4 - Survey results on image explanation accuracy of the available tools

According to the findings of the literature surveys and the information gathered from the survey, implementing a new smart assistant is an important thing for visually impaired people. So, this study aims to present an improved smart assistant with a haptic feedback system which provides audio assistance to navigate through a smart assistant to coexist in the society without feeling dependent on others too much. It functions auto focused image capturing, reading mathematical equations and table data of printed materials, classifying selected text or paragraph, images, graphs and reading aloud generated digitized text.

1.2. Research Gap

Comparing recent research on image interpretation, this component of the system is mainly focused on filling the image interpretation gap in the overall system. Studying the literature points, various tools have been built to interpret images in printed documents, and several assistive tools have also been implemented for print disabled people. However, the majority of them are missing some critical factors that should be improved for use by print-disabled people.

The majority of existing tools, even those that are great and are available on the Play Store, do not provide an explanation for this factor [6]. Some tools don't even describe the basic colors in it [7], [8]. Since "Image captioning algorithm based on multi-branched CNN and Bi-LSTM" paper [7] is a great paper which uses an attention mechanism to get the key features of the image to describe it, is not done for print disabled people. So, it also lacks the sufficient explanation for a blind person to understand an image in a printed document.

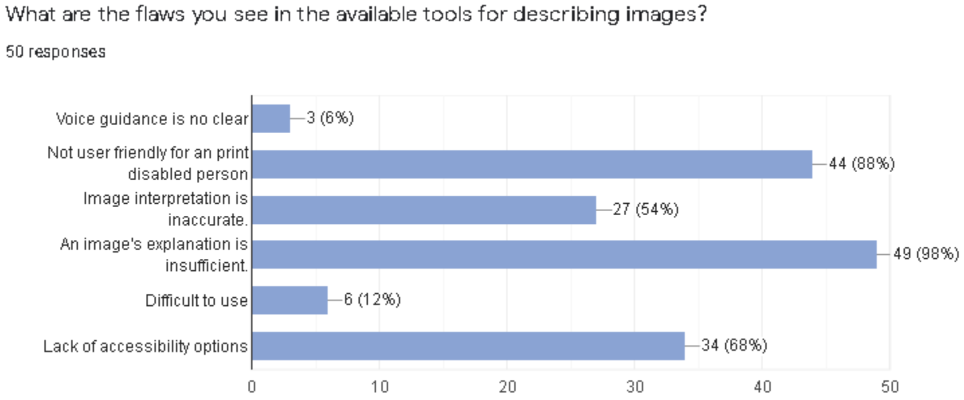
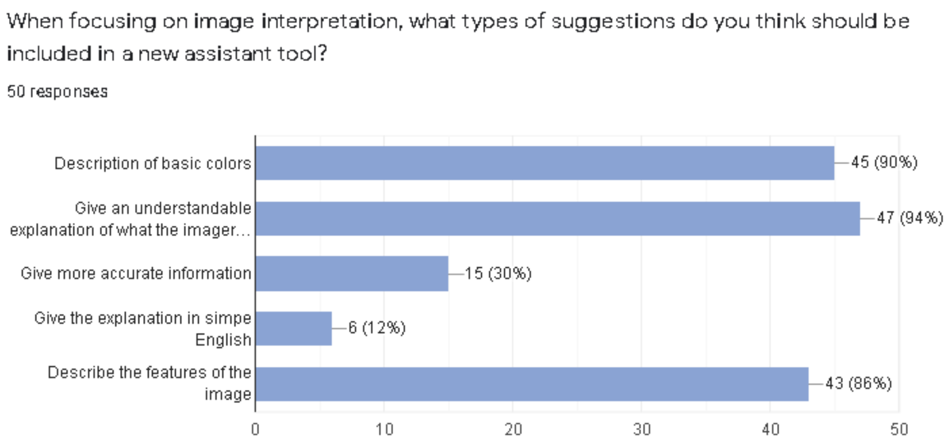


Figure 1.2.1 - Survey results on flaws in available tools

Addition to this, the majority of respondents in the survey stated that the image explanations of the available tools are insufficient. Figure 1.2-1 exemplifies this point. In consideration of Figure 1.2-2, many people suggest that feature descriptions, color descriptions, and sufficient explanations be improved in existing tools for vision impaired people to accurately interpret images.



When focusing on image interpretation, what types of suggestions do you think should be included in a new assistant tool?

50 responses

Figure 1.2.2 Survey results on the suggestions for a new assistant tool

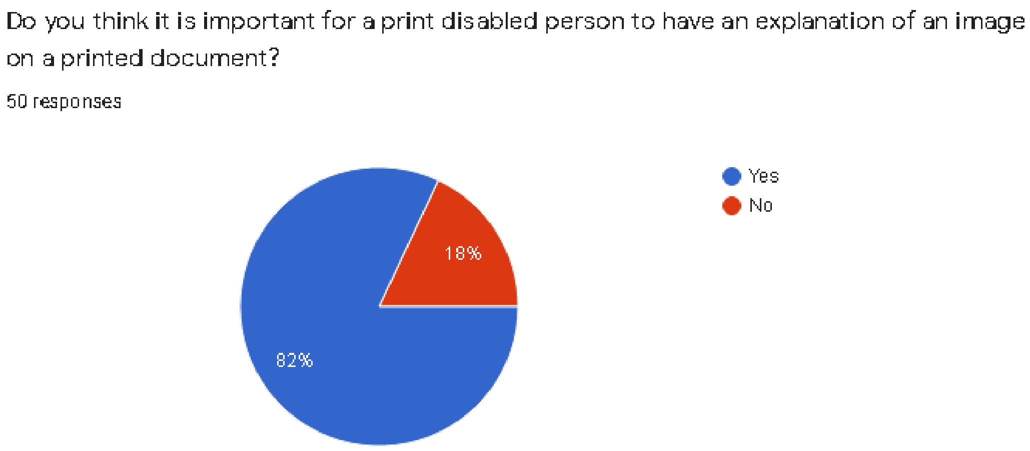
Table 1.2.1 Research gap compared to existing systems

	Research A	Research B	Research C	Proposed solution
Describing colors within the image	No	Yes	No	Yes
Describe the surrounding content other than the main subject	No	Yes	Yes	Yes
Introduce an attention mechanism to select key features	Yes	No	No	Yes
Describe the images in a way even for people who are blind since the birth can feel	No	No	No	Yes
Optimized for mobile/cloud use	No	Yes	Yes	Yes

Furthermore, Table 1.2-1 briefly compares the above-mentioned issues in the existing systems with the suggested solution. By reviewing the outcomes, it is clear that this solution is implemented with far more innovative functionalities than other currently available investigations.

1.3. Research Problem

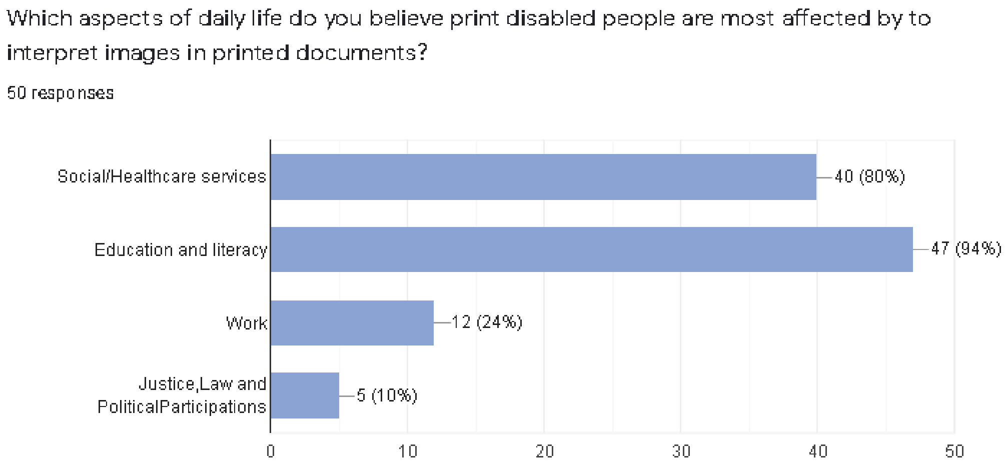
According to the survey conducted, it concludes that vision impaired individuals have many difficulties when interpreting an image in a printed document. As a result of the survey, Figure 1.3-1 depicts that many people think identifying an image is very important for print disabled people. Figure 1.3-2 says that vision impaired people are most affected to interpret an image in the educational sector and health services. Paper [9] also depicts that because every time they cannot ask help from a third party to interpret an image. Also, few say that when working with a legal or a confidential document they also face many problems. Because trusting a person in such a case is not secure. Without any help by a third party, they cannot interpret an image in a printed document [4].



Do you think it is important for a print disabled person to have an explanation of an image on a printed document?

50 responses

Figure 1.3.1 Survey results on the importance of image explanation



Which aspects of daily life do you believe print disabled people are most affected by to interpret images in printed documents?

50 responses

Figure 1.3.2 Survey results on how print disabled people affected by interpreting images

Most of the print disabled individuals use braille to make their da today work easy [10]. According to the survey commentary section and the Figure 1.1-2, it says that braille cannot give a descriptive explanation to an image in a printed document and using braille is not practical when interpreting images. Concurring to the results of the surveys, Figure 1.2-1 and literature reviews the accessibility issues with the existing tools, lack of understanding, insufficient explanation of the images and incapability to distinguish the needed information of an image in a printed document without depending on others are some other problems that are collected through the survey.

2. RESEARCH OBJECTIVES

2.1. Main Objective

The main objective of the research is to explore and identify the feasible system with a proper image interpretation model to facilitate reading a printed document. This study aims to identify the key features of an image and considering the other objects around the main features to give a proper explanation. The main target is to give the descriptive explanation of the image in simple English to understand it for a blind person who is blind since the birth.

2.2. Specific Objectives

In addition to the main objectives, there are some specific objectives related to the implementation

- To identify a proper image identification model to extract image features

First, we have to identify the existing image interpretation models and introduce a new or improved image recognition algorithm to identify the images in a document with a proper interpretation model. And also, we should identify the key features of the image to give an accurate explanation to the image. Captured images will be identified as the inputs and the new identifying model can be used to extract the features of the captured image.

- To identify a proper language model to generate a descriptive explanation

To convert the extracted image features into a natural language, we need to identify a proper language model to generate a descriptive explanation. Here, the extracted features can be output through an audio assistant in simple English. To segment and classify the regions of the document

The content in the document has to be segmented and classified in order to interpret the parts of the document using suitable methods.

- To develop a gesture-controlled process with a haptic feedback system in the mobile application

The aim of implementing a gesture-controlled process is to make the user more comfortable and used for the interfaces to navigate through them. Moreover, the gesture-controlled interaction process will be added to the system by including the tactile feedback system, which will help the disabled user to navigate through the application by feeling the vibration. It will make the user to confirm whether he selects the correct button for the navigation.

3. METHODOLOGY

3.1. Methodology

The overall research, which is under computer vision is to implement a digital assistant to aid individuals with print disabilities to interpret printed documents. The individual contribution to this project is to explore and identify models to extract the image features and convert them into a natural language. In this component, basically the user is using the mobile application while the image interpretation process will be done on cloud environment. OpenCV libraries are used in mobile devices to allow on-device machine learning and for its better performance. It is necessary to extract the images visual features with more detailed content and then generate captions which should be more catered towards vision impaired users by adding more descriptions in a way they can understand. So, this component is built with the use of various models and is mainly divided into two models like caption model and the object detection model.

This component is a combination of Neural Networks [11]. The captured image of the printed document is sent to a model to extract the features of the image. Then it identifies the main features with the respective sub objects to describe the image efficiently. So, firstly, the caption model does the caption generation after getting an input image. It uses a Seq-2Seq model to generate general captions for the image. Then, by using YOLO object detection [12], a model has been developed for object recognition. It automatically detects the important features without any human supervision. It will identify each object with the respective sub-objectives in the image separately and then divide them into two groups as human and non-human attributes. Features of the detected humans are extracted with efficientv2 implemented as a multi output model. After identifying the humans, it goes through the multi output model which is a combination of different types of models to identify the main features of the identified human to give an output of the emotions and attributes of the human subject. It takes a human count in the image and describe image by explaining each object in the image. For the face detection, it used an attribute detection model to collect the facial attributes and it uses the Kaggle's CelebFaces Attributes (CelebA) dataset [13] to recognize the facial attributes such as finding people with brown hair, are smiling, or wearing glasses likewise. This dataset is ideal for training and testing models for face detection, in particular for recognizing facial aspects of a human, as it includes a wide range of poses, background clutter, and people varieties, as well as a large number of images and descriptive annotations.

Finally, those generated outputs go through JSON objects and combine all the captions through a template model and generates a detailed description of the image in a way a blind person can understand. Through an audio assistant, the generated description of the image will output through the mobile application in a clear way even to understand it for the blind people who is blind since birth. Trough the capturing process the user will be guided by an audio assistant and the gesture-controlled process will also be guided via audio.

3.1.1. System Architecture

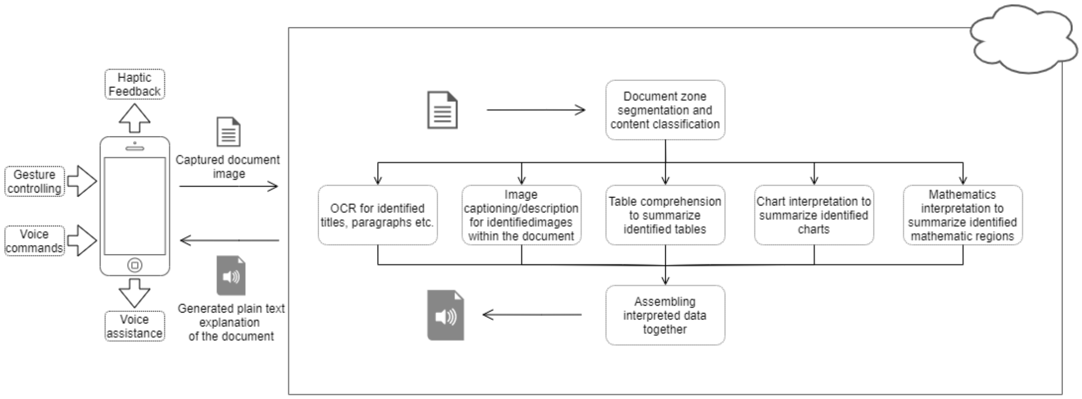

Haptic Feedback
Gesture controlling
Voice commands
Voice assistance
Captured document image
Generated plain text explanation of the document
Document zone segmentation and content classification
OCR for identified titles, paragraphs etc.
Image captioning/description for identifiedimages within the document
Table comprehension to summarize identified tables
Chart interpretation to summarize identified charts
Mathematics interpretation to summarize identified mathematic regions
Assembling interpreted data together

Figure 3.1.1 - Overall System Diagram

According to the Figure 3.1-1 it is clearly shown that all the image processing components are handled by the cloud environment which is a Django backend server and the user's interfaces, gesture controlling, and all the voice commands are handled by the mobile application. While all the client-side features are in the mobile application, Figure 3.1-2 shows a system diagram to explain how the backend of this individual research component work.

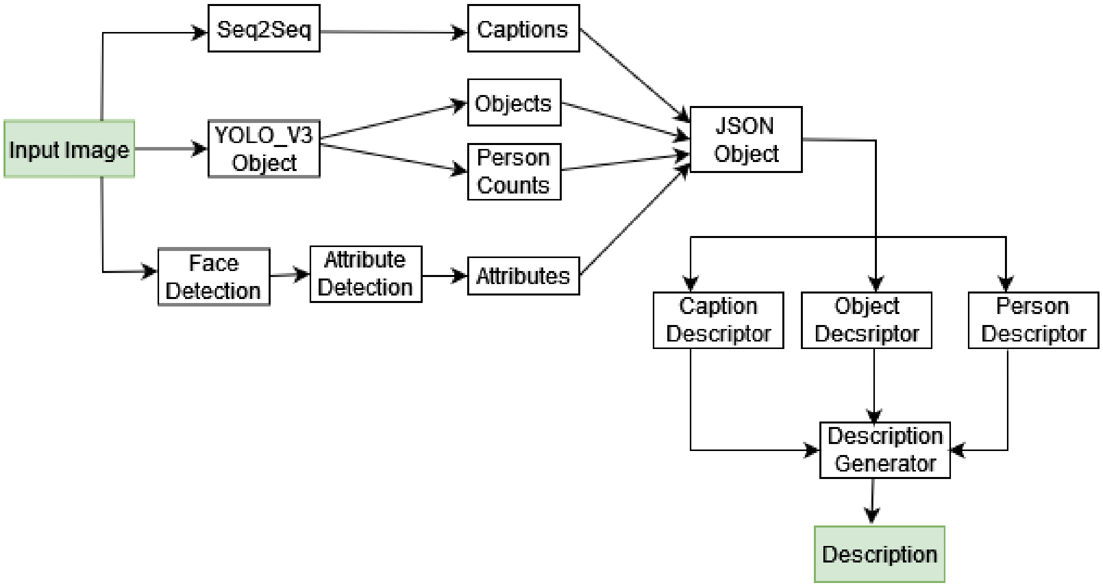


Figure 3.1.2 - System Diagram for the Individual Component

Algorithm 1 – Algorithm to generate a description doe a detected image

Algorithm 1 Algorithm to generate a description for a detected image

```
Input: InputImage
Output: Description for the InputImage
Begin:

Normalize the InputImage by pre-process input method
Extract the features using efficientv2
Generate a general caption by a Seq-2-Seq model
ObjectDetection using YOLO →
    NoOfObjects ← count the total number of objects in the image

    for Object in NoOfObjects
        if Object = human then
```

```
            Pass the detected human through different models to recognize gender,
            emotions, hair color, hair type, age

        else
            Cache the extracted non-human attributes
        end
    end
Pass through a template to generate description
end
```

3.1.2. Data collection methods

In image interpretation, since the human components have used face detection model to get human attributes, it is needed have a dataset to train the attribute detection model. There are so many datasets under image processing. So, after researching and going through the past studies we chose MSCOCO[14] and the CelebFaces Attributes (CelebA) as the best dataset to train the models. We selected the Kaggle's CelebA dataset to recognize the facial attributes of humans. It consists of 202,599 face images of various celebrities, 10,177 unique identities and 40 binary attribute annotations per image.

3.1.3. Tools and Technologies

Tools

- IntelliJ Idea
    - To implement the mobile application
- Kaggle Notebook
    - To train the machine learning models
- PyCharm
    - To implement the final Django backend

- Postman
    - To test the mobile application

Technologies

- OpenCV-python

- Dart
    - Used to implement mobile application with flutter
- TensorFlow
    - Used to implement the model
- Django
    - To implement the backend API for the mobile application

Table 3.3.1 - External tools

Description	**Tools**
Version Controlling	Gitlab
Team connectivity	Teams, WhatsApp

3.2. Commercialization aspects of the product

The commercialization of the product can be done in many ways as our final product is a mobile application. One of the best ways to commercialize the app is social media. Since this application is designed in the universal language English, this can be used not only in Sri Lanka but also all over the world. As majority of people send more time on social media, commercializing this application through Facebook, WhatsApp, Instagram and also through YouTube ads can get this app to people in a short time period. Also, as this application is used by vision impaired people, we can advertise the application through printing leaflets and aware the people by giving those to people near Eye hospitals, etc. On the other hand, by integrating the application with google adds will provide more commercialization value to the application.

Since the targeted community of this application is blind or vision impaired, and they are mainly communicating or interacting with society through hearing, we can advertise this application on radio and television. And also, in audio blogs which the vision impaired are used mostly. We can introduce this application for blind schools by conducting an event. And now a days podcasting is getting popular with people. So, it is also a great media where we can advertise this application.

3.3. Testing and Implementation

3.3.1. Implementation

The system implementation is done by focusing on the gathered information to interpret images on printed documents. Before starting the implementation, first a requirement analysis must be done. Therefore, user requirements functional requirements and non-functional requirements were gathered as below.

Functional requirements

- Extract data from the image
- Identify the objects of the image
- Describing the colors of the image
- Describe the main features of the image
- Identify humans in the image
- Generate meaningful captions
- Further description of the image using surrounded objects near main subject

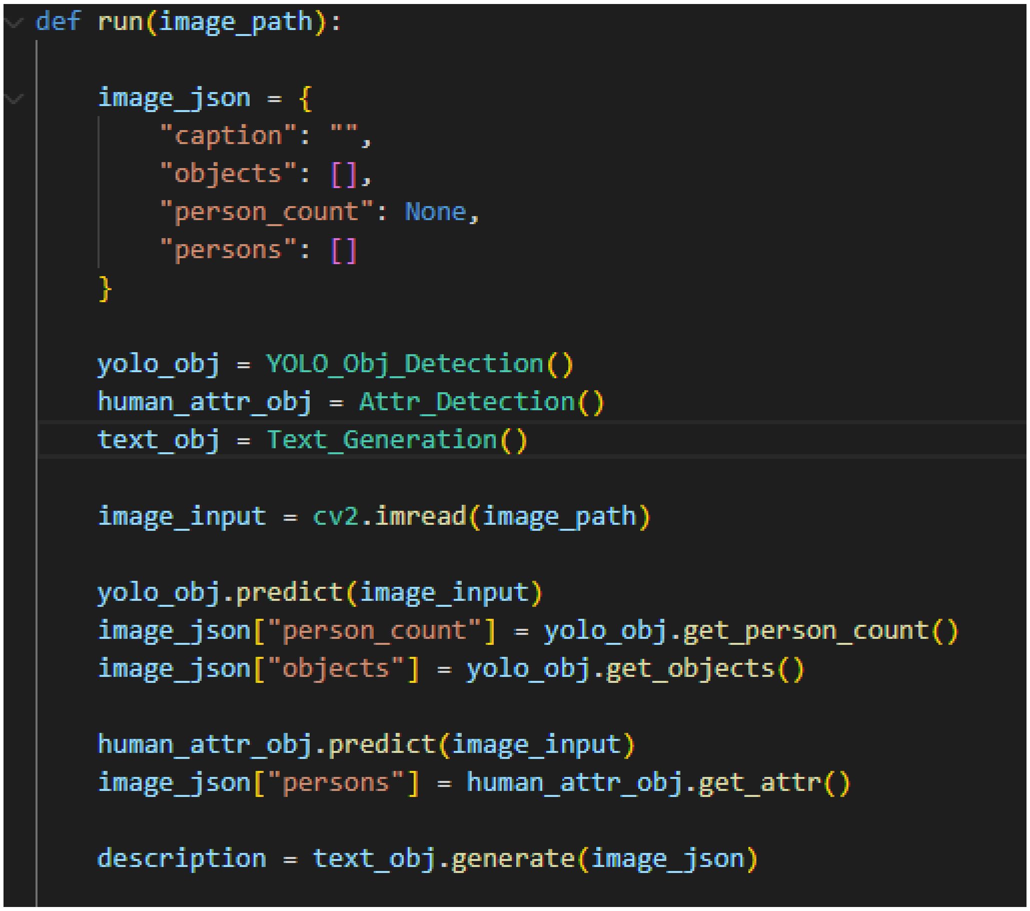
Non-functional requirements

- Usability
- Accuracy
- Availability
- Well optimized for cloud/mobile use

User requirements

- User should have a mobile phone to use the application
- User should have an English knowledge to understand the guidelines
- User should be able to hear
- User should have a simple knowledge to use a mobile application
- User should be able to touch and feel the screen to navigate through the app

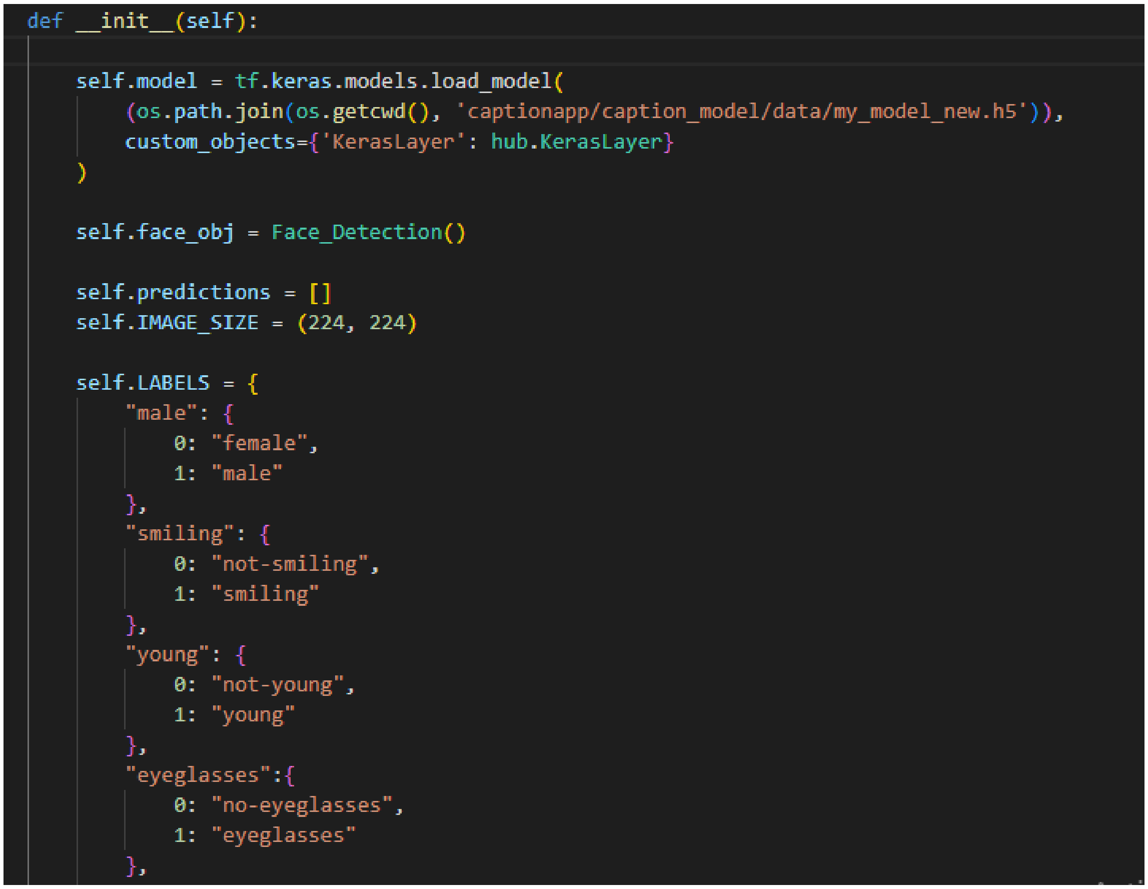
For the implementation in the image interpretation, it has both server side and client-side implementations, client-side implementations mean the mobile application and the server-side implementation means the image processing process. In the mobile application implementation, it is developed using flutter framework. It is developed with a voice assisted, haptic feedback system for make it ease the user when navigating through the pages.



```
def run(image_path):

    image_json = {
        "caption": "",
        "objects": [],
        "person_count": None,
        "persons": []
    }

    yolo_obj = YOLO_Obj_Detection()
    human_attr_obj = Attr_Detection()
    text_obj = Text_Generation()

    image_input = cv2.imread(image_path)

    yolo_obj.predict(image_input)
    image_json["person_count"] = yolo_obj.get_person_count()
    image_json["objects"] = yolo_obj.get_objects()

    human_attr_obj.predict(image_input)
    image_json["persons"] = human_attr_obj.get_attr()

    description = text_obj.generate(image_json)
```

Figure 3.3.1 - Usage of YOLO, human attributes and text detection models

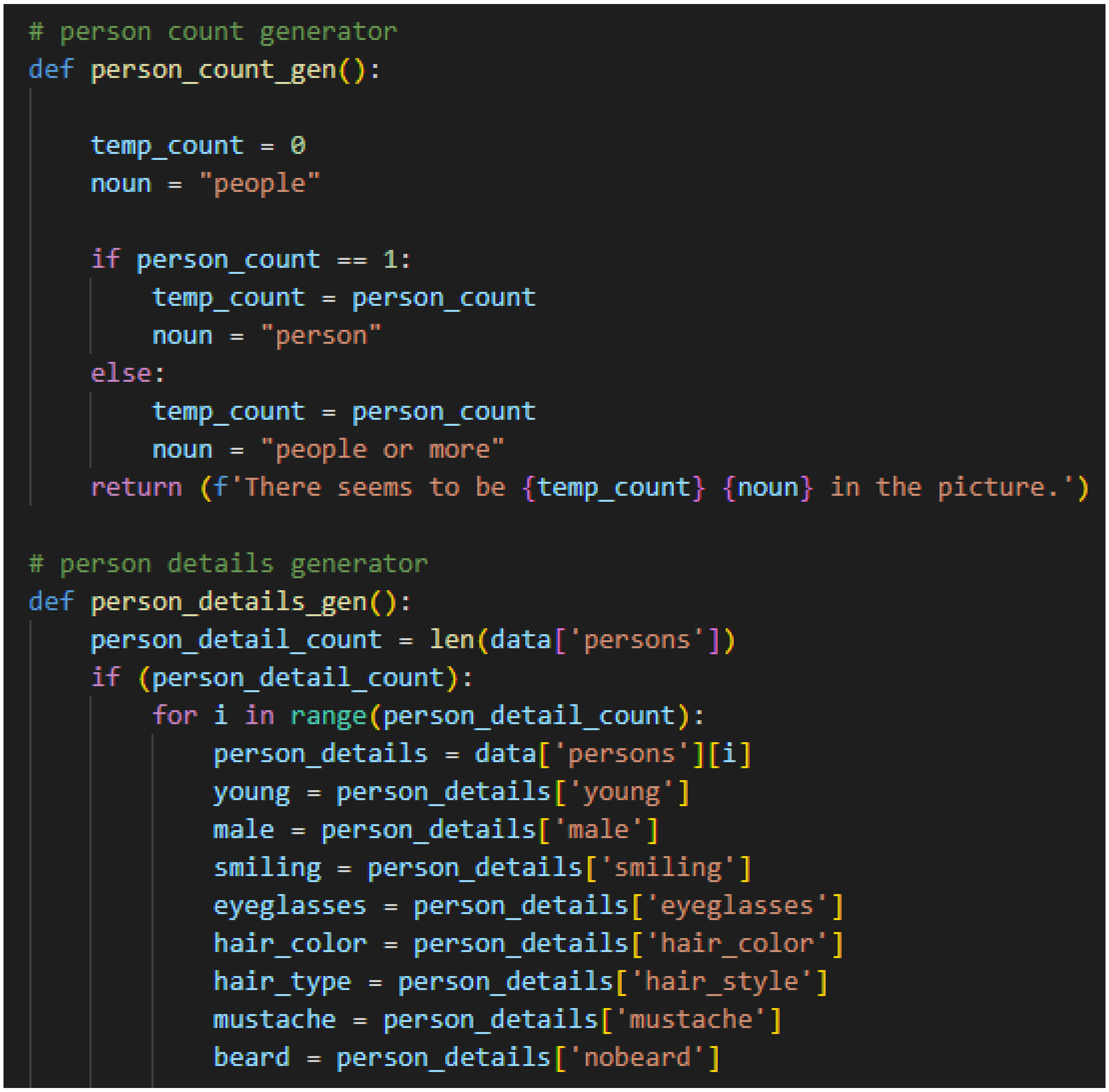
For data extraction from the captured images there are many algorithms that has been used in the implementation. First, I used a seq-2-seq model to generate a general caption for the image. Then the image is passed through YOLO object detection model to identify the objects in the image. The model separates humans and other objects and when it comes to humans, it takes the identified human count. It is capable to identify whether the human is faced to the photo or nor. Then through a face detection model, and with the use of CelebFaces Attributes (CelebA) dataset, it recognizes the human attributes like people with brown hair, smiling, wearing glasses. It is implemented to identify the hair type, hair color, whether the person is smiling, male or female, young or not whether the person have a mustache or beard or not. Features of detected humans are extracted with efficientv2 implemented as a multi output model.



```
def __init__(self):

    self.model = tf.keras.models.load_model(
        (os.path.join(os.getcwd(), 'captionapp/caption_model/data/my_model_new.h5')),
        custom_objects={'KerasLayer': hub.KerasLayer}
    )

    self.face_obj = Face_Detection()

    self.predictions = []
    self.IMAGE_SIZE = (224, 224)

    self.LABELS = {
        "male": {
            0: "female",
            1: "male"
        },
        "smiling": {
            0: "not-smiling",
            1: "smiling"
        },
        "young": {
            0: "not-young",
            1: "young"
        },
        "eyeglasses":{
            0: "no-eyeglasses",
            1: "eyeglasses"
        },
```

Figure 3.3.2 - Labeling the human attributes

To generate a descriptive description, I have used template based pre-defined templates for the description generation. Rather than implementing all the models separately, I have used a multi output model to make the implementation simple. With the use of JSON objects, it passes the separate outcomes from the models to the template model So, after the implementation images gave a successfully high accurate description and read it to the user through the mobile application in plain English sentences.



```python
# person count generator
def person_count_gen():

    temp_count = 0
    noun = "people"

    if person_count == 1:
        temp_count = person_count
        noun = "person"
    else:
        temp_count = person_count
        noun = "people or more"
    return (f'There seems to be {temp_count} {noun} in the picture.')

# person details generator
def person_details_gen():
    person_detail_count = len(data['persons'])
    if (person_detail_count):
        for i in range(person_detail_count):
            person_details = data['persons'][i]
            young = person_details['young']
            male = person_details['male']
            smiling = person_details['smiling']
            eyeglasses = person_details['eyeglasses']
            hair_color = person_details['hair_color']
            hair_type = person_details['hair_style']
            mustache = person_details['mustache']
            beard = person_details['nobeard']
```

Figure 3.3.3 - Getting the person count and person details

For the feasibility study regarding the implementation, below mentioned feasibility studies are done in the development process. Scheduled feasibility, technical feasibility and economy feasibility features are described in thee below description

Schedule Feasibility:

The suggested system should be completed within the time frame specified. To ensure a quality product, each phase should be time-bound. The Gantt chart will show the time restrictions for each task.

Technical Feasibility:

Research team members should have some understanding of mobile application development technologies and training a model using machine learning methods. To complete the suggested application, all members of the research team should have the knowledge in computer programming languages for the implementations.

Economy feasibility:

There should be cost constraints for the product's resources. All members should be within the price range. The approach should be less expensive and more comprehensive.

3.3.2. Testing

Because appropriate testing ensures that flaws and issues are found early in the application's life cycle, the product will be tested utilizing a variety of testing methodologies such as unit testing, integration testing, and user acceptability testing.

It should be released once all of the testing is completed. If there are any problems during the testing phase, these should be addressed before the product is released. This mobile based application will be released for print disabled users after the completion and clearance of user testing.

Some of the test cases used to test the product are included below, along with screenshots.

Table 3.3.1 - Test cases for image interpretation

Test Case #	Test case	Result
001	Camera opens from the open camera button	Pass
002	All the buttons and widgets are visible	Pass
003	Navigate for pages through buttons	Pass
004	Vibration works when touch the buttons	Pass
005	Image successfully captured after opening the camera	Pass
006	Image successfully uploaded for the algorithms	Pass
007	Detect the objects	Pass
008	Detect people in the image	Pass
009	Detect the people count	Pass
010	Detect males	Pass
011	Detect females	Pass
012	Detect human attributes	Pass
013	The output consists of human features	Pass
014	The output consists of objects in the image	Pass
015	The output consists of colors of the image	Pass
016	Proper sentences are generated	Pass
017	Proper words in the sentences with correct spellings	Pass

Above table shows the test cases done for image interpretation. Also, this application was tested by using three different end users and table 3.3.2 shows the results of it. For that we contacted two people with some vision difficulties and a one who has completely huge difficulties with their vision. Those people are respectively named in the table as User 3, User 2 and User 1.

Table 3.3.2 - Test cases done by end users

Test Case #	**Test case**	**User 1**	**User 2**	**User 3**
001	Open the mobile application without any error	Opened the application without any issue	Opened the application without any issue	Opened the application without any issue
002	Camera opens from the open camera button	Able to open the camera with the guidance of voice assistant	Able to open the camera from the camera button	Able to open the camera from the camera button
003	Navigate for pages through buttons	Navigated through all the pages	Navigated through all the pages	Navigated through all the pages
004	Vibration works when navigating through buttons	Navigated through the pages with the help of vibration	Navigated through all the pages	Navigated through all the pages
005	Able to capture the image successfully	Captured the image without any issue	Captured the image without any issue	Captured the image without any issue

Finally, to complete the testing process we tested out application with different OS versions with different kind of android devices. The test cases for that is shown below in the table 3.3.3.

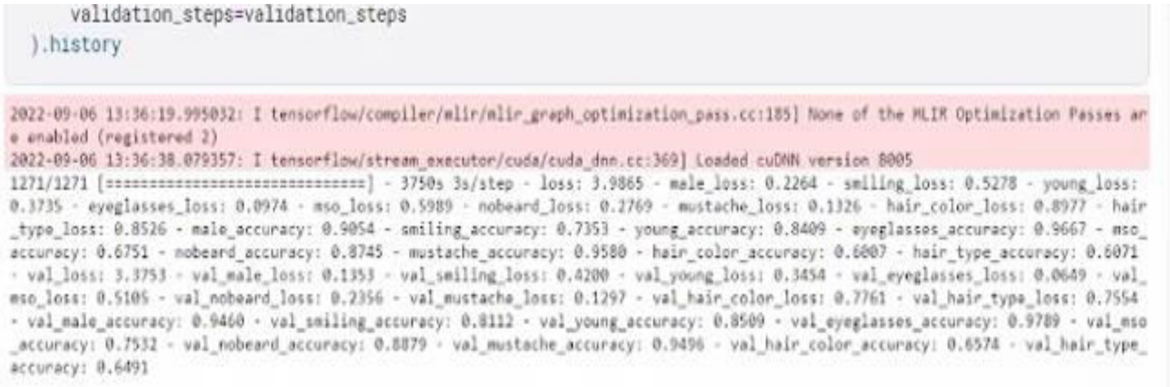
Table 3.3.3 - Test cases for devises with different OS

Test Case #	Device	OS	Version issues	Issues with the interfaces
001	Xiaomi X3Pro Poco phone	Android 13	No issues	No issues
002	Redmi 9	Android 11	No issues	No issues
003	Samsung galaxy grand prime	Android 5	No issues	No issues

4. RESULTS AND DISCUSSION

4.1. Results

The image captioning algorithm works for almost any image and generates descriptions for images that are descriptive enough for print disabled users. It identifies the main objects in the image and describe in a way that a blind personal can understand. As the features of detected humans are extracted with efficientv2 implemented as a multi output model the accuracy of the generating captions are also very high as shown in the figure 4.1.1.



```
    validation_steps=validation_steps
).history
```

```
2022-09-06 13:36:19.995032: I tensorflow/compiler/mlir/mlir_graph_optimization_pass.cc:185] None of the MLIR Optimization Passes ar
e enabled (registered 2)
2022-09-06 13:36:38.079357: I tensorflow/stream_executor/cuda/cuda_dnn.cc:369] Loaded cuDNN version 8005
1271/1271 [==============================] - 3750s 3s/step - loss: 3.9865 - male_loss: 0.2264 - smiling_loss: 0.5278 - young_loss:
0.3735 - eyeglasses_loss: 0.0974 - mso_loss: 0.5989 - nobeard_loss: 0.2769 - mustache_loss: 0.1326 - hair_color_loss: 0.8977 - hair
_type_loss: 0.8526 - male_accuracy: 0.9054 - smiling_accuracy: 0.7353 - young_accuracy: 0.8409 - eyeglasses_accuracy: 0.9667 - mso_
accuracy: 0.6751 - nobeard_accuracy: 0.8745 - mustache_accuracy: 0.9580 - hair_color_accuracy: 0.6807 - hair_type_accuracy: 0.6071
- val_loss: 3.3753 - val_male_loss: 0.1353 - val_smiling_loss: 0.4200 - val_young_loss: 0.3454 - val_eyeglasses_loss: 0.0649 - val_
mso_loss: 0.5105 - val_nobeard_loss: 0.2356 - val_mustache_loss: 0.1297 - val_hair_color_loss: 0.7761 - val_hair_type_loss: 0.7554
- val_male_accuracy: 0.9460 - val_smiling_accuracy: 0.8112 - val_young_accuracy: 0.8509 - val_eyeglasses_accuracy: 0.9789 - val_mso
_accuracy: 0.7532 - val_nobeard_accuracy: 0.8879 - val_mustache_accuracy: 0.9496 - val_hair_color_accuracy: 0.6574 - val_hair_type_
accuracy: 0.6491
```

Figure 4.1.1 - Accuracy of the generated attributes

As shown above, the accuracy for the attributes smiling, young, hair color, hair type, eyeglasses, mustache, male female, beard are respectively 87%, 85%, 75%, 78%, 97%, 94%, 97%, and 92%.

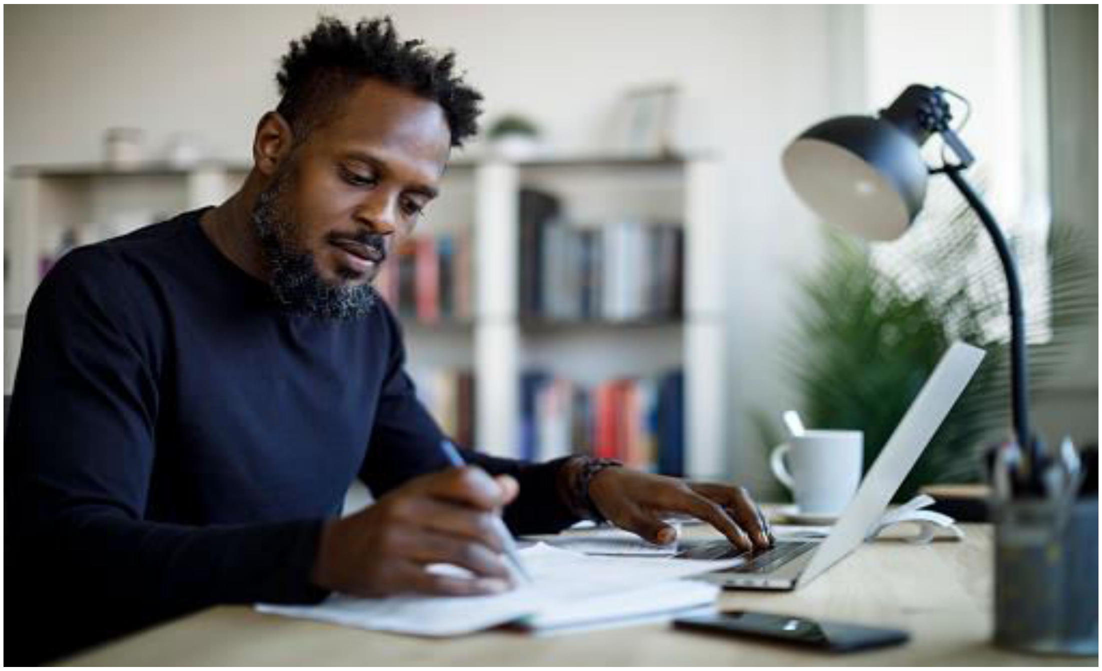
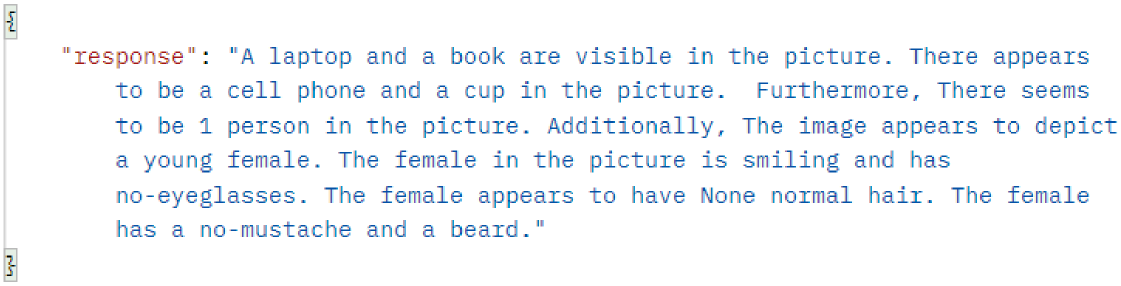


Figure 4.1.2 - Sample image tested with man

Below Figure dispalys the caption generated for the figure through the algorithm



```
{
    "response": "A laptop and a book are visible in the picture. There appears
        to be a cell phone and a cup in the picture.  Furthermore, There seems
        to be 1 person in the picture. Additionally, The image appears to depict
        a young female. The female in the picture is smiling and has
        no-eyeglasses. The female appears to have None normal hair. The female
        has a no-mustache and a beard."
}
```

Figure 4.1.3 - Results got for the sample image a man working

Figure 4.1.2 and Figure 4.1.3 shows two results that are collected from the implemented image interpretation component. As you can see the predicted descriptions are approximately equal to the real-time human generated captions. So, it is clearly shown that the implemented mobile application gives a high accurate description and is capable to identify males and female separately and describe the images by indicating the other sub objects in the image.

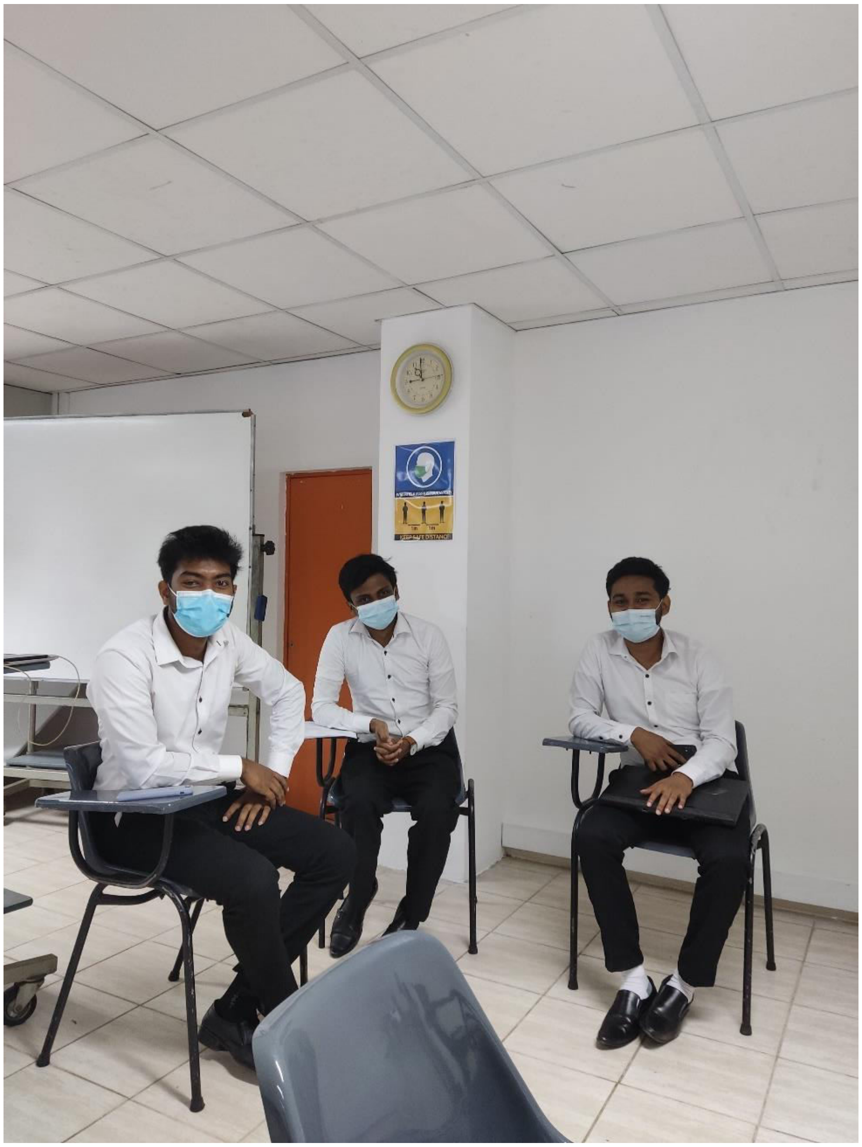
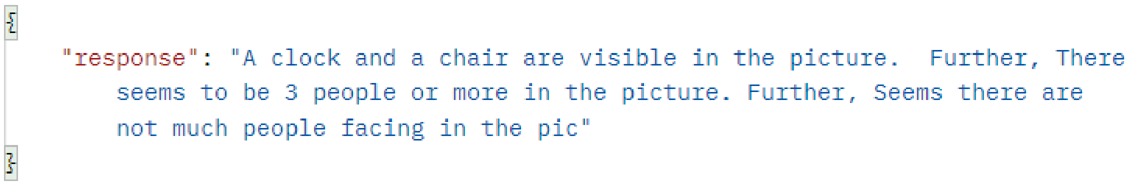


Figure 4.1.4 - Sample image tested with three people siting

For the above Figure the generated caption through the models that are trained for thr image interpretation will generate the caption as shown in the figure



```
{
    "response": "A clock and a chair are visible in the picture.  Further, There
        seems to be 3 people or more in the picture. Further, Seems there are
        not much people facing in the pic"
}
```

Figure 4.1.5 - Result got for the image of three people siting

4.2. Research Findings

This research is mainly focused on building an application to read document content for vision-impaired people to fill the lack of available resources in accessible mediums. There are many research done under this component. But still those have many obstacles that cannot be addressed using those mediums. So, the main outcome of this study is to developed mobile application for vision impaired people using many image processing techniques to fulfil their needs.

Yet, when considering the accuracy of the implemented system, it has achieved an overall accuracy more than 94%. Therefore, the system is capable of identifying the images in a high accurate rate. For the implementation of this image processing study, many available image processing, and deep learning techniques have been used and it can be seen that the currently available deep learning techniques like convolutional neural networks, object detection algorithms are really helpful for this implementation and well as in many others.

As this component is heavily based on images there are many datasets that can be used with the implementation according to the need. MSCOCO, Flickr 8k, Flickr 3k are some commonly used datasets that can be used in image processing.

So, by this implemented research study, the significant gap in reading rights and equality between print disabled people and the general population can be solved for a considerable extent.

4.3. Discussion

Several key factors are discovered throughout the development and testing phases in the study. Previously, some research was undertaken on the subject in order to acquire a better understanding of the population of visually impaired people. The literature survey included data from the worldwide vision data base collected from population-based studies of blindness and visual impairment, as well as age-specific prevalence of blindness and the number of blind people by age. After doing some preliminary study, we sought some firsthand information on the matter. That's when we decided to conduct an online poll to collect some critical research data. The survey was carried out via distributing a Google form.

Initially, the system uses a seq-2-seq algorithm to generate a general caption to the image and then using YOLO object detection model, it identifies the main objects in it. However, this model is capable of extracting and detecting humans and non-human objects from the image and get a count of it. Then the features of the detected human are extracted with efficientv2 implemented as multi output model. Finally, the final descriptive sentences are generated by pre-defined templates. As proposed in the proposal, I was able to complete all the tasks and have managed to generate a high accuracy description for a captured image from a document.

In the testing process, there are a few things that should be implemented. The quality of the images when captured from a smartphone gets low. So, it affects the accuracy of the output. If we could be able to capture the image by getting a high-quality image, the accuracy can be increased more than the current results.

However, the current capturing result of an image gives a considerable output which a print disabled user can understand. The document will be captured from the mobile device by the user initially. The input will then be received by the backend and collected for processing. In the backend, the document will be divided into numerous areas, and the regions will then be classified. The separated pieces will then be sent to the appropriate algorithms for interpretation. Following the completion of the analysis, a voice output will prompt an explanation. Furthermore, the system will guide the user during the entire process.

4.4. Summary of student contribution

Student: Sanduni Madara P.G – IT 19392172

Research component: Identifying a proper image captioning approach to facilitate interpreting an image in a document

Task: This component is aligned with to develop a system to interpret images included in the scanned material. The inputs of this component are the image portions of the scanned material, and the output will describe the images in plain English. Furthermore, the tactile feedback system of the application which will help the disabled user when navigating through the application will be implemented.

Tasks completed:

- Developed the mobile application
- Implemented the backend with Django to get the image interpretation services by the application
- Developed image captioning and describing models.
- Implemented the image interpreting algorithms to extract the image content
- Developed haptic feedback system in the proposed mobile application.

5. CONCLUSION

In this study, a new image captioning approach is implemented with many deep learning and image processing techniques. The implemented approach was tested on the MSCOCO and CelebA datasets and showed a significantly improved captioning performance over the state-of-the-art approaches. This approach has the potential to be integrated into hardware platforms such as smartphones, easing the challenges that visually impaired persons face on a daily basis. There are limitations to this method because the accuracy of the entire system is mainly dependent on the quality of the smartphone camera. To avoid this, a higher-quality external camera might be integrated into the solution. As future work, this implemented image captioning approach can be further developed by increasing the accuracy by using the above suggestion.

REFERENCES

[1] "Blindness and vision impairment," Who.int. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment. [Accessed: 11-Feb-2022].

[2] K. Smelyakov, A. Chupryna, D. Yeremenko, A. Sakhon, and V. Polezhai, "Braille character recognition based on neural networks," in *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)*, 2018, pp. 509–513.

[3] Á. Csapó, G. Wersényi, H. Nagy, and T. Stockman, "A survey of assistive technologies and applications for blind users on mobile platforms: a review and foundation for research," *J. Multimodal User Interfaces*, vol. 9, no. 4, pp. 275–286, 2015.

[4] N. D. U. Gamage, K. W. C. Jayadewa, and J. A. D. C. A. Jayakody, "Document reader for vision impaired elementary school children to identify printed images," in 2019 International Conference on Advancements in Computing (ICAC), 2019, pp. 279–284.

[5] P. Sankalpani, I. Wijesinghe, I. Jeewani, R. Anooj, M. D. J. T. H. Mahadikaara, and J. A. D. C. Anuradha Jayakody, "'smart assistant': A solution to facilitate vision impaired individuals," in *2018 National Information Technology Conference (NITC)*, 2018, pp. 1–6.

[6] B. Makav and V. Kilic, “A new image captioning approach for visually impaired people,” in *2019 11th International Conference on Electrical and Electronics Engineering (ELECO)*, 2019.

[7] S. He, Y. Lu, and S. Chen, “Image captioning algorithm based on multi-branch CNN and bi-LSTM,” *IEICE Trans. Inf. Syst.*, vol. E104.D, no. 7, pp. 941–947, 2021.

[8] P. Shah, V. Bakrola, and S. Pati, “Image captioning using deep neural architectures,” in *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, 2017, pp. 1–4.

[9] A. Jayakody, S. Lokuliyana, A. A. T. Sampath, G. T. S. Silva, S. A. L. Rajanthika, and H. M. T. B. Seneviratne, “Mobile application for vision impaired people to facilitate to learn the English language,” *Int. J. Comput. Appl.*, vol. 138, no. 12, pp. 12–17, 2016.

[10] J. John, “Recognition of Documents in Braille,” *arXiv [cs.CV]*, 2017.

[11] J. Chen, “What is a neural network?,” *Investopedia*, 12-Sep-2006. [Online]. Available: https://www.investopedia.com/terms/n/neuralnetwork.asp. [Accessed: 22-Oct-2022].

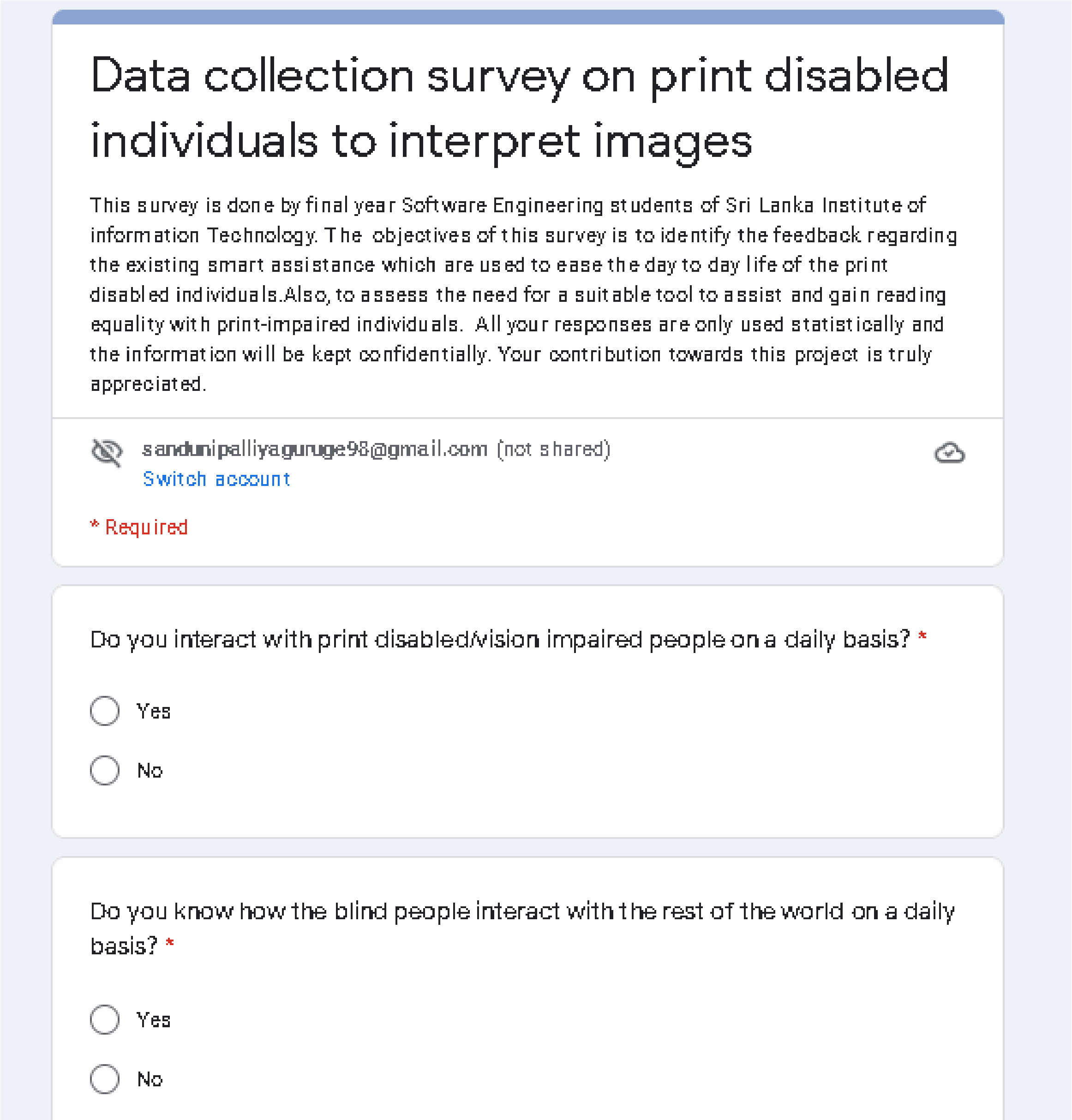
[12] G. Karimi, “Introduction to YOLO algorithm for object detection,” *Engineering Education (EngEd) Program | Section*. [Online]. Available: https://www.section.io/engineering-education/introduction-to-yolo-algorithm-for-object-detection/. [Accessed: 21-Oct-2022].

[13] J. Li, “CelebFaces Attributes (CelebA) Dataset.” 01-Jun-2018.

[14] “COCO - common objects in context,” *Cocodataset.org*. [Online]. Available: https://cocodataset.org/. [Accessed: 22-Oct-2022].

APPENDICES

Appendix A: Online Survey



Data collection survey on print disabled individuals to interpret images

This survey is done by final year Software Engineering students of Sri Lanka Institute of information Technology. The objectives of this survey is to identify the feedback regarding the existing smart assistance which are used to ease the day to day life of the print disabled individuals.Also, to assess the need for a suitable tool to assist and gain reading equality with print-impaired individuals. All your responses are only used statistically and the information will be kept confidentially. Your contribution towards this project is truly appreciated.

sandunipalliyaguruge98@gmail.com (not shared)

* Required

Do you interact with print disabled/vision impaired people on a daily basis? *

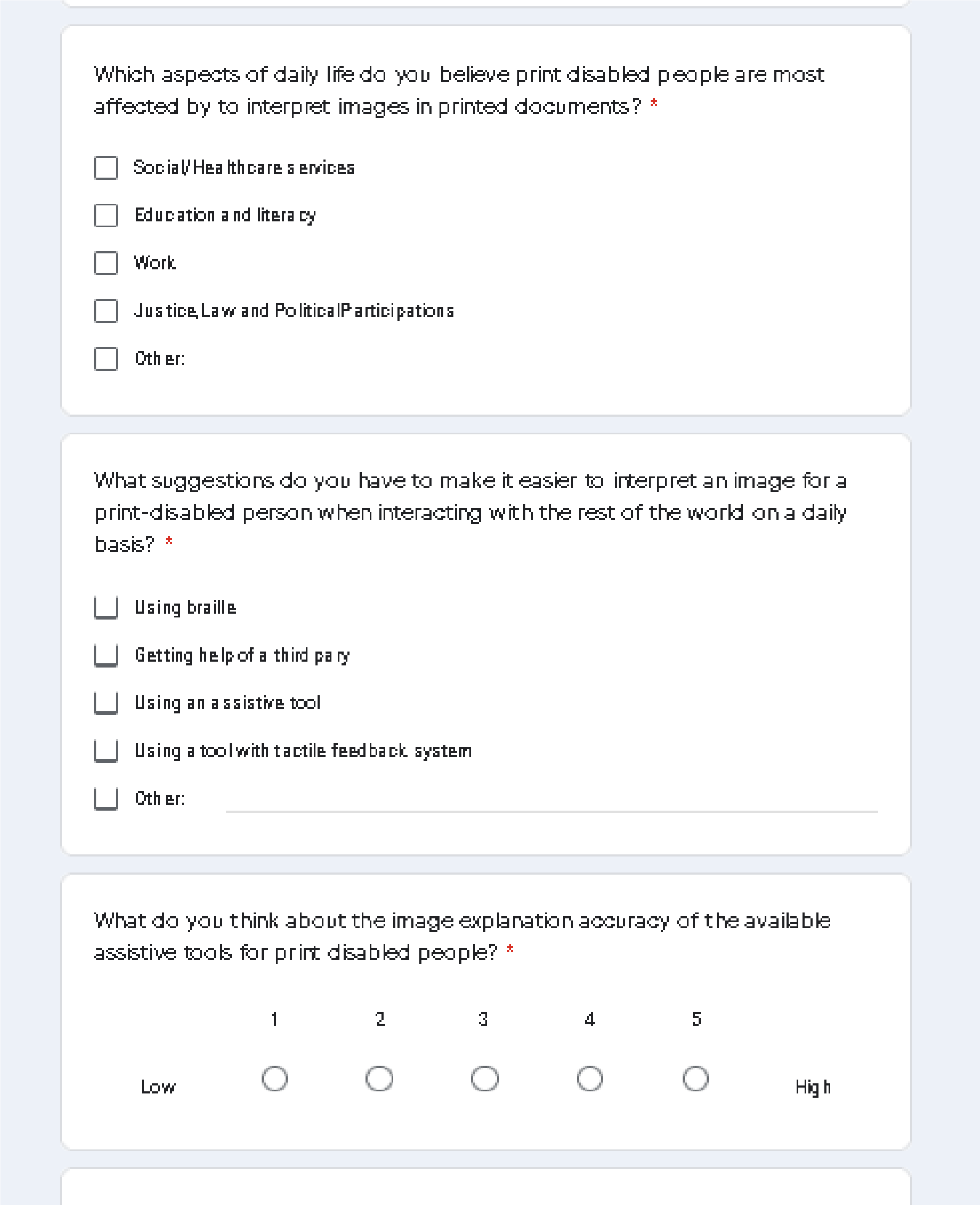
○ Yes

○ No

Do you know how the blind people interact with the rest of the world on a daily basis? *

○ Yes

○ No



Which aspects of daily life do you believe print disabled people are most affected by to interpret images in printed documents? *

□ Social/Healthcare services

□ Education and literacy

□ Work

□ Justice,Law and PoliticalParticipations

□ Other:

What suggestions do you have to make it easier to interpret an image for a print-disabled person when interacting with the rest of the world on a daily basis? *

□ Using braille

□ Getting help of a third pary

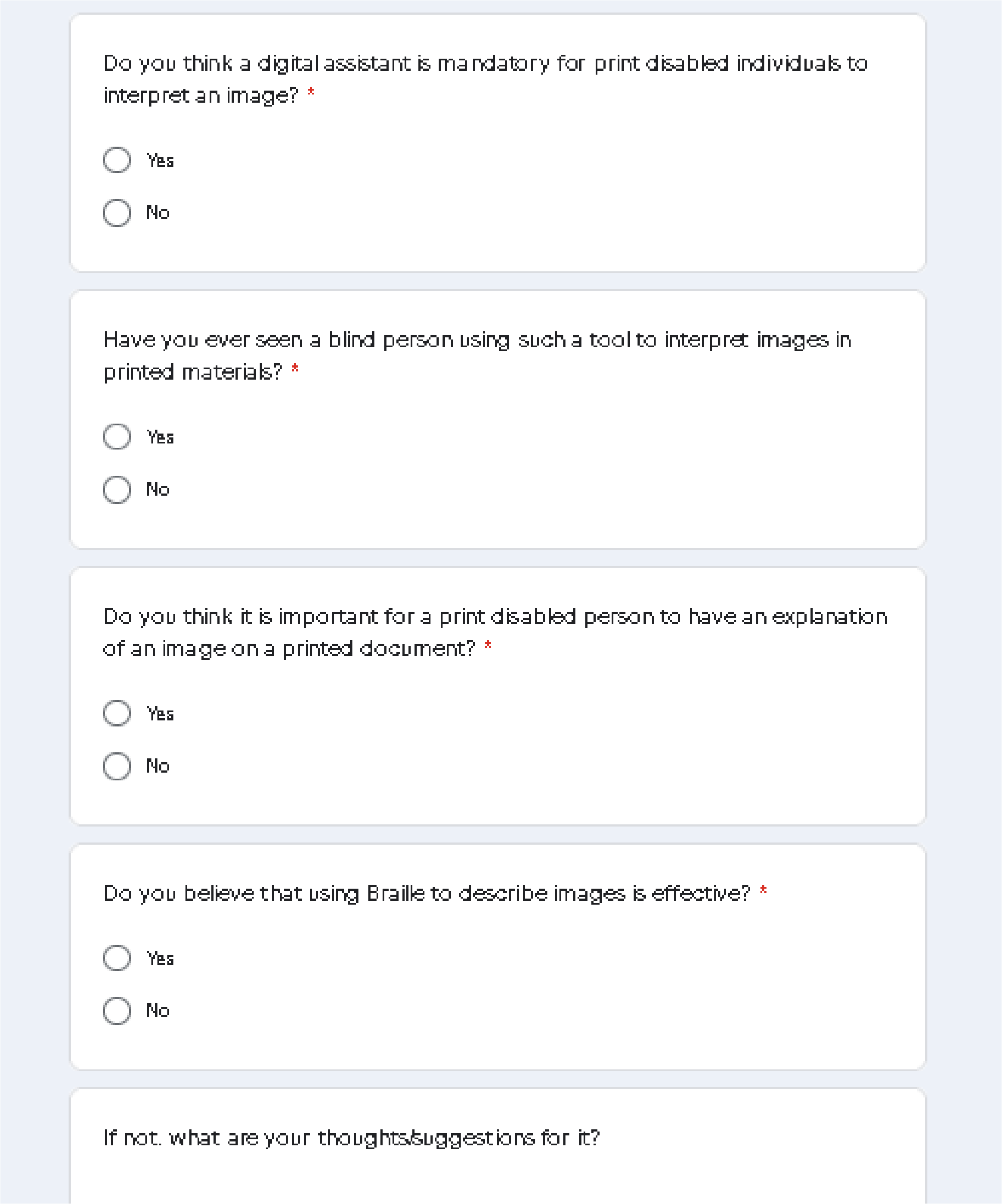
□ Using an assistive tool

□ Using a tool with tactile feedback system

□ Other:

What do you think about the image explanation accuracy of the available assistive tools for print disabled people? *

	1	2	3	4	5	
Low	○	○	○	○	○	High



Do you think a digital assistant is mandatory for print disabled individuals to interpret an image? *

○ Yes

○ No

Have you ever seen a blind person using such a tool to interpret images in printed materials? *

○ Yes

○ No

Do you think it is important for a print disabled person to have an explanation of an image on a printed document? *

○ Yes

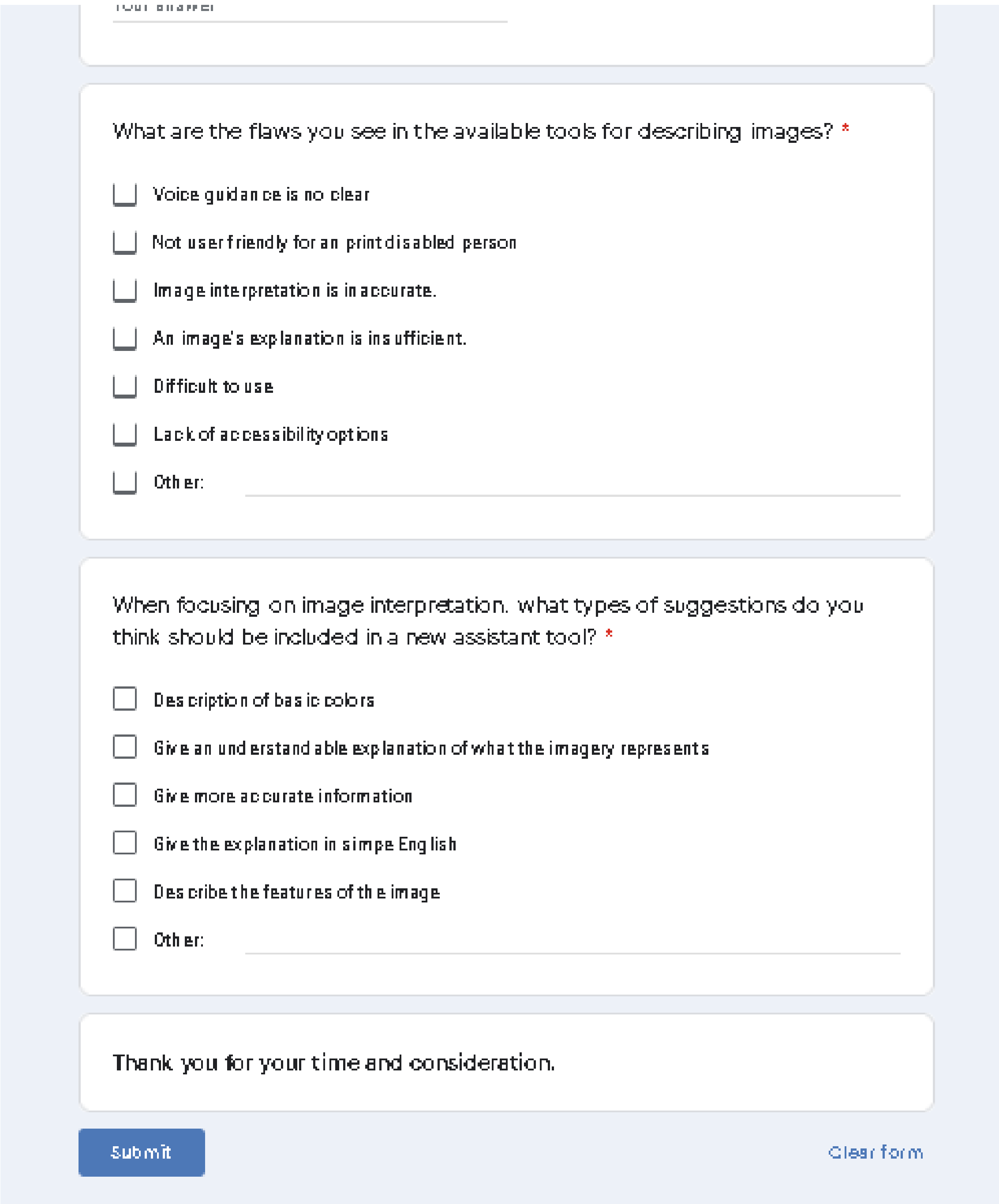
○ No

Do you believe that using Braille to describe images is effective? *

○ Yes

○ No

If not, what are your thoughts/suggestions for it?



Your answer

What are the flaws you see in the available tools for describing images? *

□ Voice guidance is no clear

□ Not user friendly for an print disabled person

□ Image interpretation is inaccurate.

□ An image's explanation is insufficient.

□ Difficult to use

□ Lack of accessibility options

□ Other:

When focusing on image interpretation. what types of suggestions do you think should be included in a new assistant tool? *

□ Description of basic colors

□ Give an understandable explanation of what the imagery represents

□ Give more accurate information

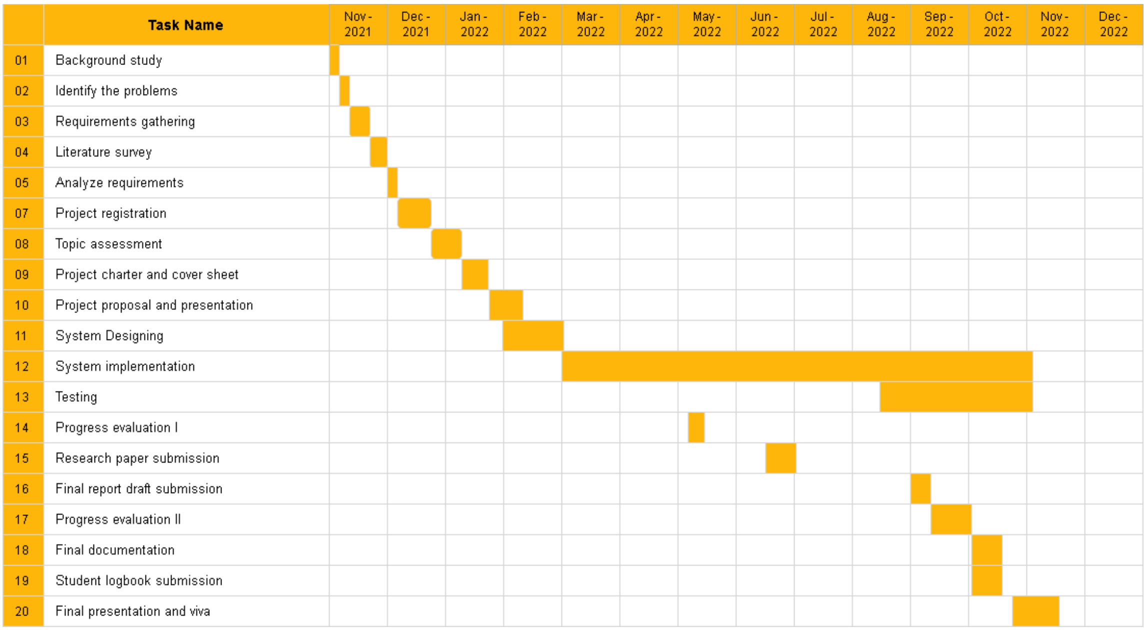
□ Give the explanation in simpe English

□ Describe the features of the image

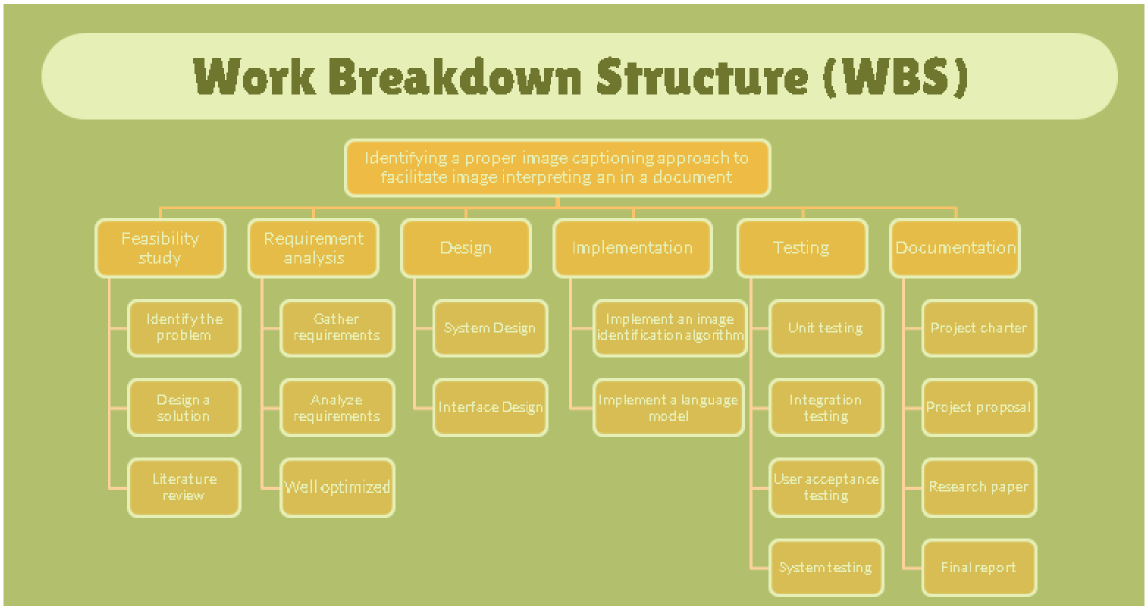
□ Other:

Thank you for your time and consideration.

Appendix B: Gantt chart



Appendix C: Work Breakdown Structure



Work Breakdown Structure (WBS)

Identifying a proper image captioning approach to facilitate image interpreting an in a document

Feasibility study
- Identify the problem
- Design a solution
- Literature review

Requirement analysis
- Gather requirements
- Analyze requirements
- Well optimized

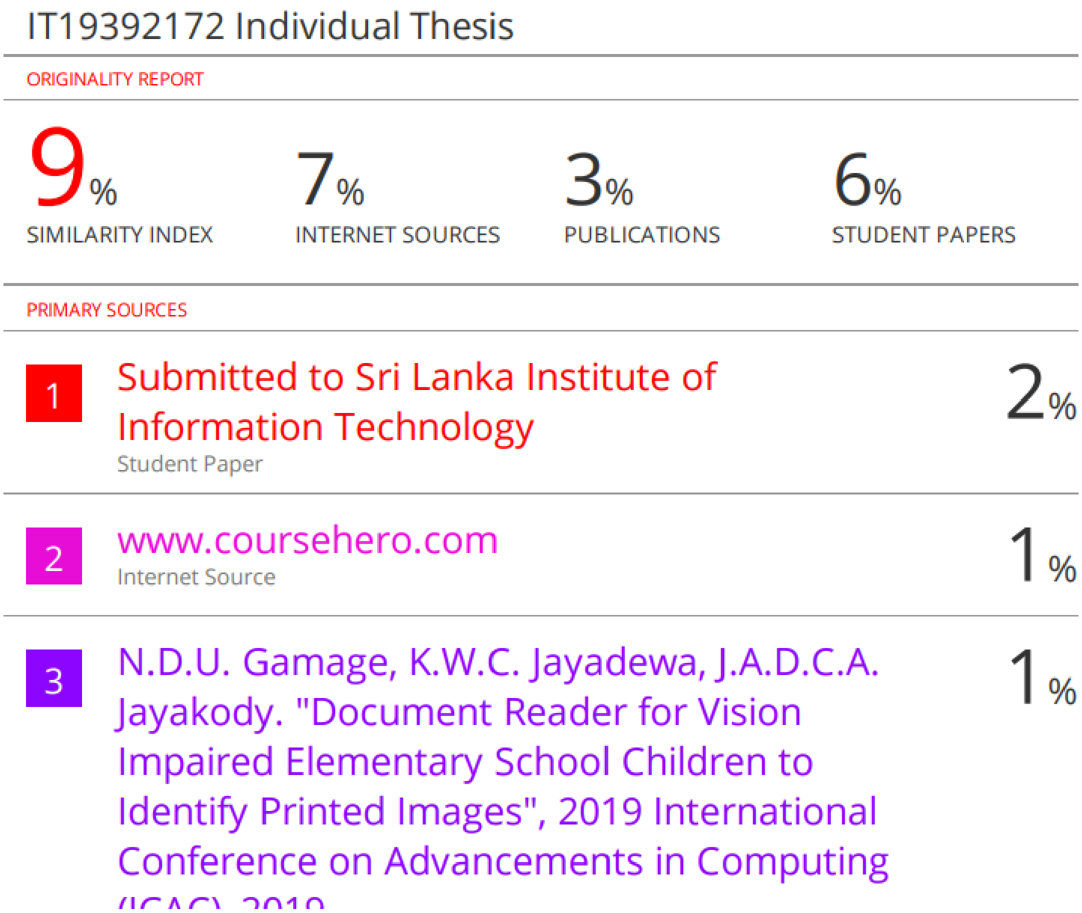
Design
- System Design
- Interface Design

Implementation
- Implement an image identification algorithm
- Implement a language model

Testing
- Unit testing
- Integration testing
- User acceptance testing
- System testing

Documentation
- Project charter
- Project proposal
- Research paper
- Final report

Appendix D: Plagiarism report



IT19392172 Individual Thesis

ORIGINALITY REPORT

9%	7%	3%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Sri Lanka Institute of Information Technology Student Paper	2%
2	www.coursehero.com Internet Source	1%
3	N.D.U. Gamage, K.W.C. Jayadewa, J.A.D.C.A. Jayakody. "Document Reader for Vision Impaired Elementary School Children to Identify Printed Images", 2019 International Conference on Advancements in Computing (ICAC), 2019	1%