ABSTRACT:
With the growth of Internet and advances in storage technologies, acquisition and spread of images has become imperative. However, despite unprecedented advances in technology, these images still remain inaccessible to people with visual impairment. Images available online can be accessed by people with visual impairment using screen-readers only if they have some alternate text associated with them. Despite this being crucial for the visually impaired, most of the images remain without any alternate text, thus posing great difficulty for the visually impaired in perceiving images. Here we investigate the problems and challenges faced by the visually impaired, while accessing images and do market evaluation of a tool that automatically converts images into textual description, thereby eliminating the need of alternate text for images.

APPROACH:
• To identify and study major challenges and problems faced by visually impaired people in accessing and understanding images.
• To study, demonstrate and use COCO dataset, and study its efficacy in improving the accessibility and understandability of images.

PURPOSE OF THE STUDY:
The currently available screen-readers convert the textual information available on the screen into speech format. Thus, in order for images to be accessible to people having visual impairment, some textual information needs to be associated with it. However, most of the images at present do not have any alternate text associated with them and the very few alternate texts that are present are not very helpful in understanding what is going on in the image. This makes it impossible for people having visual impairment to understand the context of the image.

Hence, the specific Objectives of this study are:
(i) To study and report the current methods used by visually impaired people for accessing images and their limitations;
(ii) To assess the resulting user anxiety, impact on personal, and professional and social well-being;
(iii) To assess how well the developed system will be able to solve the problems encountered by visually impaired people while accessing images.
CURRENT SYSTEMS AND THEIR LIMITATIONS:
Humans learn to recognize images with the help of their own real-world experiences and examples. Similarly, for a computer to understand and recognize images, it is essential that it is given adequate amount of training data. The images in the training data should be annotated. This will help the computer to recognize the images it has seen in the past. However, if the computer encounters an image that it has never seen before, it will not be able to recognize it. For instance, if a child is standing with a toothbrush in his hand and the computer has never seen a toothbrush before, it may confuse the toothbrush to something else like a baseball bat.

Facebook is currently working on an artificial intelligence-based object recognition tool to help blind users to understand the content in the photos people share on Facebook [1].

Moreover, in the past, some researchers have attempted to answer various visual questions that people with visual impairment have, with the help of crowdsourcing [2]. Two mobile phone applications called VizWiz Social and BeSpecular have been developed that allow people with visual impairment to upload pictures and ask questions about the pictures from “the crowd”.

Even though computers have been successful in identifying various images with the help of training datasets with billions of training examples provided to them, there is still a need to describe the images with more detail so as to allow people with visual impairment enjoy images as much as people with normal vision do. Therefore, in this phase of the project, we undertake the study of effectiveness of a tool that can connect the parts of images with words and phrases in the sentences. Such a tool converts images into their appropriate detailed textual description so as to make them understandable to people with visual impairment.

DATASET USED:
In order to perform various experiments, Microsoft COCO dataset was used for all purposes. Microsoft COCO is an image recognition, segmentation and captioning dataset. Microsoft COCO classifies images containing various objects into appropriate categories and provides more than one caption per image [3].

EXPERIMENTS:
The experiments were done in two phases. In the first phase, we got into association with two organizations: Finger Lakes Independence Center (FLIC) and Association for Vision Rehabilitation and Employment (AVRE), located in Ithaca and Binghamton, New York respectively. Both these organizations work for people with visual impairment. A total of 27 volunteers from both the organizations took part in a qualitative questionnaire study. Their participation in the questionnaire helped us
identify and investigate the current methods used by them for accessing images, and the problems faced by them while doing so. In the second phase, visually able people were blindfolded and provided with an image description. Microsoft COCO image captions were used for describing images to people in this phase. After giving the description, people were presented with three similar images and were asked to identify the image which according to them was described to them. These experiments were conducted in order to evaluate whether people are able to select the correct image or not when provided with COCO descriptions.

(I) UNDERSTANDING AND IDENTIFYING THE CURRENT METHODOLOGIES USED BY VISUALLY IMPAIRED PEOPLE FOR UNDERSTANDING IMAGES, AND THEIR LIMITATIONS:

This section presents an overview of the questionnaire-based interview study which is used for examining the current methods used by people with visual impairment for accessing images and their limitations.

(a) USER GROUP ENROLLMENT:

27 people associated with Finger Lakes Independence Center located in Ithaca, New York and Association for Vision Rehabilitation and Employment located in Binghamton, New York participated in a questionnaire-based interview. In order to protect the privacy of the users, the interview was conducted online using survey forms. Some of the participants were interested in discussing their challenges in greater detail. This was done on one-to-one telephonic conversation basis. Responses provided by users were recorded and analyzed in order to examine the current methods used by them, and the limitations of those methods.

(b) QUALITATIVE INTERVIEW:

A Questionnaire-based interview was conducted that enquired about the current methods used by people having impairment in order to access images. The interview centered on the current challenges faced while accessing images, current software used while working on computers and need for assistance from a sighted person. Further, the participants were asked to identify features which according to them are essential for understanding the context of the image. The questions also aimed at assessing the psychological effects of inability to understand the images available on the internet.

The questionnaire consisted of the following questions:

1. What obstacles are you currently facing while accessing images on the internet?
2. What software product are you currently using while working on the computers?
3. How comfortable do you feel in asking for help in order to understand the context of a particular image?
4. If you feel uncomfortable in asking for help, please explain why.
5. What according to you are important features in an image that can help in understanding that image successfully?
6. Facebook is currently working on an artificial intelligence-based object recognition tool which will help in automatically describing any image uploaded on Facebook. Will you be interested in using such a tool?
7. Please share any other aspect, which you feel should be taken into consideration while describing images.

(II) EXPERIMENTS USING COCO DATASET:
This section presents an overview of the second-phase of experimentation which was conducted in order to study and demonstrate the efficacy of Microsoft COCO Dataset in accessing images.

(a) FUNCTIONAL OVERVIEW OF MICROSOFT COCO:
Microsoft COCO is an image recognition, segmentation and captioning dataset. It recognizes objects in the context of scene-understanding. The dataset contains more than 300,000 images with multiple objects per image. Objects are labeled using per-instance segmentations to aid in precise object localization [4]. Each image has several captions associated with it.

(b) USER GROUP ENROLLMENT:
Five people having visual abilities were asked to participate in the experiments related to the use of COCO dataset in understanding the context of the images.

(c) EXPERIMENTS CONDUCTED:
The people who took part in the experiments were blindfolded. Image captions available in COCO dataset for a particular image was then read out to them. After ensuring that they were clear with the description, their blindfold was opened and they were presented with three similar images. They were then asked to select the image, which according to them, was described to them. Five such cases were described to each participant. The effectiveness of COCO dataset captions was studied in terms of number of correct images selected by the participants.

RESULTS AND DISCUSSIONS:
This section presents the observations made during the two-phase experimentation process. The section is divided into two sub-sections. The first sub-section discusses the results of questionnaire-based survey conducted, and the second sub-section discusses the results of experiments done by blindfolding people.
I) MAJOR CHALLENGES AND PROBLEMS FACED BY PEOPLE WITH VISUAL IMPAIRMENT WHILE ACCESSING IMAGES:

The questionnaire analysis and observation yield that people having visual impairment face various difficulties while understanding the context of the images. Figure 1 describes the challenges currently faced by people having visual impairment while understanding the context of images.

![Figure 1: Problems currently faced in understanding images](image)

These problems are discussed as follows:

1. **NO ALTERNATE TEXT AVAILABLE WITH IMAGES:**
   18 participants said that many images which are available on the internet do not have any alternate text associated with them. This makes it impossible for them to understand the context of the image as the current screen readers can only read out the text available on the screen. They said because of lack of description they felt completely clueless about what the image is about.

2. **AVAILABLE ALTERNATE TEXT IS NOT GOOD:**
   11 participants said that the alternate text available with the images is not detailed enough and does not provide them with a clear idea of what is going on in the image.

3. **PRIVACY CONCERN:**
   3 participants said that they do not want to share with others the images that they are viewing and hence feel reluctant in asking for help from other people.
4. DELAY IN RECEIVING HELP FROM OTHERS:
5 participants said that they feel reluctant in asking for help from others because in the past whenever they asked for help, sighted people said that they would help but did not help immediately.

5. HIGHLY SUBJECTIVE DESCRIPTION GIVEN BY OTHERS:
3 participants said that if they ask for help from sighted people, the description they receive is highly subjective and based on the perception of the person providing them with the description. As a result, people with visual impairment felt that they were missing out on important information.

(a) CURRENTLY USED SOFTWARE PRODUCTS AND THEIR LIMITATIONS:
According to the results of the survey, all participants currently rely on screen-readers while working on the computers. Figure 2 shows various screen-readers which are currently used by people with visual impairment. According to the survey, 75% of the participants use JAWS, 15% use NVDA (NonVisual Desktop Access) and 10% use ZoomText. The distribution is based on the survey responses provided by the participants of our study. However, these current screen readers cannot describe images. The only way that a screen reader can convey the meaning of an image is by reading text in the document that serves as a substitute or alternative for that image. If there is no alternative text associated with an image, then the screen reader cannot convey the meaning of the image.
(b) SEEKING HELP FROM SIGHTED PEOPLE:
Most of the participants said that they do not prefer asking sighted people for help regarding understanding the context of the image available in front of them. They felt asking others for help will affect them psychologically and make them look professionally less competent. Figure 3 describes the various psychological effects that participants felt while asking others for help.

![Figure 3](image)

These psychological effects are described as follows:

1. DON’T WANT TO BOTHER OTHERS:
8 participants said that they feel uncomfortable in asking for help from others because they don’t want to constantly bother others. This thought makes them feel reluctant in approaching others for help.

2. HIGH ANXIETY:
17 participants said they feel high anxiety in asking for help from others. They constantly feel that others will think low of them and as a result they feel high anxiety in asking for help.

3. DON’T WANT TO APPEAR LESS COMPETENT:
22 participants said that they do not want to appear less competent in front of others and thus don’t want to ask for help. They feel if they will ask for help at workplace, others will feel that they are less competent and this will affect their professional life.
4. RUDE BEHAVIOR OF OTHERS:
5 participants said that in the past whenever they asked for help from others, they received rude or inappropriate responses. This made the people with visual impairment feel low about themselves.

5. DON’T WANT TO SHARE IMAGES VIEWED WITH OTHERS:
7 participants said that they felt reluctant in asking for help from others because they did not want to share the images that they are viewing with others. They felt asking others for help will allow others to judge them and intrude in their personal lives.

(c) IMPORTANT FEATURES IN AN IMAGE AND INTEREST IN USING A TOOL THAT WILL AUTOMATICALLY DESCRIBE THE IMAGES:
Many participants felt that in addition to the objects present in an image, it is essential for them to know the colors in the image. Some participants felt that the activity which is going on in the image is the most essential aspect for them in order to understand the context of the image. Figure 4 describes factors that people with visual impairment feel are important for them in order to understand the context of the image.

![Importance of Image Features](image_url)
These features are described as follows:

1. **COLORS IN THE IMAGE:**
   13 participants said that the colors of the objects is an essential feature for them, which will help them better understand the context of the image. They also said that most of the current descriptions do not include colors in them. Even though this aspect may seem less important for sighted people, it is extremely important for people with visual impairment to understand the context of the image.

2. **IMPORTANT LANDMARKS:**
   7 participants said that the description should include landmarks present in the frame of the image. This will help them understand the location of the image. It will also give them a clarity regarding whether the image was taken indoors or outdoors.

3. **ACTIVITY GOING ON IN THE IMAGE:**
   22 participants said that it is essential for them to know what activity is taking place in the image. They want to know the activities that are a part of the image, including an act of play, dance or even an act as simple as a smile. This will help them understand the context of the image in a better way.

4. **OBJECTS IN THE IMAGE:**
   12 participants said that it is important for them to know all the objects that are present in the frame of the image, both in the foreground and the background.

5. **PORTION OF INCOMPLETE OBJECTS IN THE FRAME:**
   1 participant said that it is important to know if there are any incomplete objects in the frame and if yes then what portion of the object is in the frame. For example, they want to know if there is an incomplete car in the picture frame, and if yes, what is its orientation and which part of the car is present in the frame.

6. **EXPRESSIONS OF PEOPLE:**
   7 participants said that if there are people in the frame then it is important for them to know the expressions of people. This will help them understand the overall mood of the image.

7. **POSTURES OF PEOPLE:**
   4 participants said that if there are people in the frame then it is important for them to understand their posture. It is important for them to not only understand whether a particular person is sitting or standing but also the angles at which they are bent in the picture. This helps them understand the overall body language of the people in the picture.
8. TEXT (CHARACTERS) IN THE IMAGE:

11 participants said that it is important for them to know the text present in the image. They said that it is also important for them to know the numbers (if present) in the image. It is, for instance, quite possible that an image shows a milestone reading number of miles to a particular destination, or the name of a place or a series of instructions; in such cases they want to know the text inside the images.

Moreover, all of the participants agreed that they will be interested in using a tool, similar to the one currently being developed by Facebook, which automatically converts all the images into appropriate textual descriptions.

(II) EFFECTIVENESS OF COCO DATASET:

The experiments conducted by blindfolding people demonstrated the effectiveness of captions available for images in the COCO Dataset. Table 1 summarizes the effectiveness of COCO captions in terms of number of correct images selected by the participants. It also shows the average time taken per image by the participants in selecting the image.

<table>
<thead>
<tr>
<th>PARTICIPANT NUMBER</th>
<th>AVERAGE TIME TAKEN PER IMAGE (IN SECONDS)</th>
<th>NUMBER OF CORRECT SELECTIONS (OUT OF 5)</th>
<th>SUCCESS RATE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant #1</td>
<td>4.8</td>
<td>5</td>
<td>100%</td>
</tr>
<tr>
<td>Participant #2</td>
<td>7.2</td>
<td>5</td>
<td>100%</td>
</tr>
<tr>
<td>Participant #3</td>
<td>5.6</td>
<td>4</td>
<td>80%</td>
</tr>
<tr>
<td>Participant #4</td>
<td>9.3</td>
<td>3</td>
<td>60%</td>
</tr>
<tr>
<td>Participant #5</td>
<td>3.9</td>
<td>5</td>
<td>100%</td>
</tr>
</tbody>
</table>

**TABLE 1**

According to Table 1, COCO descriptions had a success rate of 88% on average, which is very close to perfection. Moreover, the data suggests that 3 out of 5 people got a success rate of 100%. It should also be noted that the median of the data is 5 out of 5 correct selections.

Appendix I show the images used for experimentation and the related user responses.

NOTES ON THE STUDY:

1. Prior consent was taken from the representatives of both the organizations before involving users in the study.
2. No question that discloses the identity of the participant was asked. Further, all the responses were kept anonymous.

3. For each question, participants could select multiple options for an answer.

CONCLUSION:
In this study we have systematically explored and documented the current methods used by people with visual impairment for accessing and understanding images. The results depict that all participants are currently facing problems in understanding the context of the images. Majority of the participants felt that asking sighted people for help is a major source of anxiety for them. As a part of this study, we also analyzed the effectiveness of image captions available in the COCO Dataset. The participants who were provided with COCO descriptions demonstrated high success rate in identifying images.

FUTURE WORK:
People with visual impairment primarily rely on touch and auditory means in order to recognize physical objects present in their surroundings. Even though this approach works for them with physical objects, it does not work for images present on the internet or even on their touch devices such as their computers or mobile phones. However, recognition of images by people with visual impairment will be easier if the images are transported to aural form [5]. This can be potentially done with the help of Image Sonification. Figure 5 describes the transportation of images in the auditory form with the help of Sonification, which can then be understood by people with visual impairment.

![Figure 5: Visualization of the process from visually impaired to visual sonification and then to images.]

**Milestone 1:** In the next phase of the project, we intend to explore the concept of Image Sonification for images present in the COCO Dataset. Object detection and scene-understanding are important for image sonification [6]. COCO dataset already contains objects classified into different categories, thus, it will be a good choice to begin with COCO images. However, this will not allow users to upload any arbitrary image.

Further, the current user study suggests that most of the people with visual impairment feel that colors in the image are important for them to understand the context of the image. Thus, the process of image sonification can be extended to colors.

**Milestone 2:** In order to do so, the color information can be extracted from the hue, saturation and value attributes, which can then be mapped into audio attributes directly.
related to the perception of pitch, timbre and loudness [7]. This will help the users with visual impairment understand the range of colors present in the images, presence or absence of light sources, as well as the location and shape of various objects in the images.

Similar research done in the past scans the loaded image from top to bottom, and produces a sound for each row which consists of the composition of other elementary sounds. The sounds of the rows are played in sequence and not all at once. This produces simpler sounds as compared to the sounds obtained by playing all rows simultaneously. This allows users to perceive basic geometric shapes easily [7].

**Milestone 3:** We intend to extend the interface by allowing the users to select only specific objects in an image and get the audio related to the selected portion, instead of getting the auditory information related to the complete image. This tool when extended to all the smartphones will allow users to explore different objects in the image independently just by touching anywhere on the image.

**ACKNOWLEDGEMENTS:**

I am grateful to Professor Serge Belongie for giving me the opportunity to work on this fascinating project. I am also thankful to him for his valuable insights and guidance throughout the semester, which helped me proceed forward with the project. I am also thankful to all the visually impaired people who spared their valuable time and volunteered to participate in the study. I would also like to extend my gratitude to all the sighted people who came forward to take part in the second phase of the experimentation.

**REFERENCES:**


APPENDIX I

IMAGE 1

COCO IMAGE:

a very nice small plane in a big grassy field.
a small red and white train sitting in the field
a small engine plane landing on a grass runway.
a small propeller plane is on the ground.
a small plane on a field with trees in the background

TWO OTHER RELATED IMAGES SHOWN TO USERS:
PARTICIPANT RESPONSE:

<table>
<thead>
<tr>
<th>PARTICIPANT NUMBER</th>
<th>RESPONSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant #1</td>
<td>Correct</td>
</tr>
<tr>
<td>Participant #2</td>
<td>Correct</td>
</tr>
<tr>
<td>Participant #3</td>
<td>Correct</td>
</tr>
<tr>
<td>Participant #4</td>
<td>Incorrect</td>
</tr>
<tr>
<td>Participant #5</td>
<td>Correct</td>
</tr>
</tbody>
</table>

IMAGE 2

COCO IMAGE:

![Image of a cat sitting on the keyboard of a computer on a table.](image2.jpg)

COCO DESCRIPTION:

- a cat sitting on the keyboard of a computer on a table.
- a cat laying on an open laptop on a desk.
- the cat is laying on the laptop on the desk.
- a gray and white cat laying on top of a laptop computer.
- a white and gray cat is laying on a laptop computer.
TWO OTHER RELATED IMAGES SHOWN TO USERS:

![Participant Response Table]

<table>
<thead>
<tr>
<th>PARTICIPANT NUMBER</th>
<th>RESPONSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant #1</td>
<td>Correct</td>
</tr>
<tr>
<td>Participant #2</td>
<td>Correct</td>
</tr>
<tr>
<td>Participant #3</td>
<td>Correct</td>
</tr>
<tr>
<td>Participant #4</td>
<td>Correct</td>
</tr>
<tr>
<td>Participant #5</td>
<td>Correct</td>
</tr>
</tbody>
</table>

IMAGE 3

COCO IMAGE:
COCO DESCRIPTION:
this is fruit sitting in a wood basket
there is a variety of fruit in the bowl
bananas, an apple, an orange and pears in a wooden bowl.
some different fruits are laying in a bowl
a pile of fruit sits in a clean bowl

TWO OTHER RELATED IMAGES SHOWN TO USERS:

PARTICIPANT RESPONSE:

<table>
<thead>
<tr>
<th>PARTICIPANT NUMBER</th>
<th>RESPONSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant #1</td>
<td>Correct</td>
</tr>
<tr>
<td>Participant #2</td>
<td>Correct</td>
</tr>
<tr>
<td>Participant #3</td>
<td>Incorrect</td>
</tr>
<tr>
<td>Participant #4</td>
<td>Incorrect</td>
</tr>
<tr>
<td>Participant #5</td>
<td>Correct</td>
</tr>
</tbody>
</table>

COCO IMAGE:
COCO DESCRIPTION:

a cat sits on a chair in the window.
a cat rests on a chair in latticed sunlight.
a cat laying down on a cushioned chair.
a cat is laying on the cushion of the chair.
a brown and white cat lounges in a wooden chair.

TWO OTHER RELATED IMAGES SHOWN TO USERS:

![Image 1](image1.jpg)  ![Image 2](image2.jpg)

PARTICIPANT RESPONSE:

<table>
<thead>
<tr>
<th>PARTICIPANT NUMBER</th>
<th>RESPONSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant #1</td>
<td>Correct</td>
</tr>
<tr>
<td>Participant #2</td>
<td>Correct</td>
</tr>
<tr>
<td>Participant #3</td>
<td>Correct</td>
</tr>
<tr>
<td>Participant #4</td>
<td>Correct</td>
</tr>
<tr>
<td>Participant #5</td>
<td>Correct</td>
</tr>
</tbody>
</table>
COCO IMAGE:

![Image of two elephants at a zoo eating some vegetation.](image)

COCO DESCRIPTION:

two elephants at a zoo eating some vegetation.
two elephants trying to eat some sparsely there leaves
two elephants stand on a dirt area eating leaves and twigs.
two elephants eating leaves in front of a pool.
two elephants eating trees inside of their habitat.

TWO OTHER RELATED IMAGES SHOWN TO USERS:

![Image of two elephants in a natural setting.](image)
<table>
<thead>
<tr>
<th>PARTICIPANT NUMBER</th>
<th>RESPONSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant #1</td>
<td>Correct</td>
</tr>
<tr>
<td>Participant #2</td>
<td>Correct</td>
</tr>
<tr>
<td>Participant #3</td>
<td>Correct</td>
</tr>
<tr>
<td>Participant #4</td>
<td>Correct</td>
</tr>
<tr>
<td>Participant #5</td>
<td>Correct</td>
</tr>
</tbody>
</table>