Increasing Web3D Accessibility with Audio Captioning

Nicholas F. Polys
npolys@vt.edu
Virginia Tech
Blacksburg, Virginia, USA

Sheeban Wasi
swmohd@vt.edu
Virginia Tech
Blacksburg, Virginia, USA

ABSTRACT

Situational awareness plays a critical role in daily life, enabling individuals to comprehend their surroundings, make informed decisions, and navigate safely. However, individuals with low vision or visual impairments face difficulties in perceiving their real or virtual environment. In order to address this challenge, we propose a 3D computer vision-based accessibility solution, empowered by object-detection and text-to-speech technology. Our application describes the visual content of a Web3D scene from the user's perspective through auditory channels, thereby enhancing situational awareness for individuals with visual impairments in virtual and physical environments. We conducted a user study of 44 participants to compare a set of algorithms for specific tasks, such as Search or Summarize, and assessed the effectiveness of our captioning algorithms based on user ratings of naturalness, correctness, and satisfaction. Our study results indicate positive subjective results in accessibility for both normal and visually-impaired subjects and also distinguish significant effects between the task and the captioning algorithm.

CCS CONCEPTS

• **3D on Web → X3DOM**; • **Machine Learning → Computer Vision**; • **Metaverse**; • **Statistical Analysis**;

KEYWORDS

user study, neural network, YOLO, narration

ACM Reference Format:
Nicholas F. Polys and Sheeban Wasi. 2023. Increasing Web3D Accessibility with Audio Captioning. In *The 28th International ACM Conference on 3D Web Technology (Web3D '23), October 09–11, 2023, San Sebastian, Spain.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3611314.3615902

1 INTRODUCTION

It is estimated that over 250 million people worldwide are visually impaired, with over 25 million experiencing complete blindness. However, if we include those with near or distant vision impairment, this number rises to 2.2 billion [WHO 2022]. Deploying 3D environments with deep learning capabilities and narration of scene contents has the potential to significantly improve the situational awareness of visually impaired individuals. For example, such multimodal design solutions can be useful in several Metaverse applications, such as architectural reviews [Polys et al. 2017], virtual field trips [Polys et al. 2021], and environmental education destinations [Polys et al. 2018]. This accessibility is analogous to the physical accessibility improvements made to sidewalks, which not only benefit people who use wheelchairs but also parents with strollers or individuals carrying heavy suitcases [Elmqvist 2023].

1.1 Design Challenge

Identifying the types, numbers, or dimensions of objects in a room, image, or virtual environment can pose a significant challenge for individuals who are visually impaired or blind. Nearsightedness and farsightedness are two broad categories of visual impairment. Other factors that may contribute to vision impairment include uncorrected refractive errors, age-related eye issues, glaucoma, cataracts, diabetic retinopathy, trachoma, corneal opacity, or untreated presbyopia. It is worth noting that approximately 80% of individuals who are visually impaired or blind reside in low- and middle-income countries, where they may not have access to expensive assistive technology.

There are crucial design aspects to be considered for a scene description or audio captioning interface. First is the activity and task of the user. In early interviews, it became clear that vision-impaired people rely on their listening accuity to compensate for their visual accuity, especially when navigating space. Therefore, instead of notification pushes (either continuous or periodic), we focused on solutions that were on-demand. Specifically, at a user-defined time (for safety, comfort, or need), the user initiates a screenshot of the 3D image plane and receives an audio description of its contents.

In order to enhance accessibility for the visually impaired, it is necessary to provide a means of portraying the environment that relies on a strong and comprehensive spatial representation rather than only visual cues. This can be achieved through sensory substitution, or the transformation of visual representations into other linguistic encodings, and other presentation modalities. In terms of accommodating the blind, the two most practical options for sensory substitution are the modalities of touch and sound [Chundury et al. 2021]. A study involving 10 Orientation and Mobility (O&M) experts has demonstrated that spatial structure is a useful representation, even for the blind. Thus, the visual structure of the scene is the unified reference representation [Elmqvist 2023].

1.2 Contribution

Studies suggest that blind people possess the same level of visual thinking skills as sighted individuals. Interestingly, visual reasoning does not require vision [Elmqvist 2023]. In an interview with Orientation & Mobility (O&M) professionals, individuals who teach blind individuals to navigate in the real world, it was suggested that

Web3D '23, October 09–11, 2023, San Sebastian, Spain

ACM ISBN 979-8-4007-0324-9/23/10…$15.00
https://doi.org/10.1145/3611314.3615902

all of their students are virtually familiar with visualization from either their personal or professional lives, or before they became blind [Chundury et al. 2021].

The current limitations and recent advancements in accessibility research motivated us to consider solutions that can specifically aid in increasing users' situational awareness in Metaverse virtual environments (and potentially the physical world as an augmented environment). In order to enhance the comprehensiveness of the captioned scene information, we integrated four distinct algorithms (in two distinct categories) on top of Computer Vision object detection services. Furthermore, the algorithms generate grammatical natural language descriptions of the scene, which is presented to the user in an audio description.

2 BACKGROUND AND RELATED WORK

Prior to commencing the development of our proposed solution, we conducted a thorough review of relevant literature and analyzed interviews [Anderson 2022] with individuals with visual impairments to gain a better understanding of the challenges they face with respect to virtual reality (VR) technology. The literature indicates that VR poses significant accessibility challenges for blind and low-vision users, with the in-headset user interface being the most difficult aspect of the technology for such individuals to access.

2.1 Legibility

Although there are commercially available accessibility tools and software, such as screen magnifiers, high-contrast VR dashboards, screen readers, or game interfaces, these tools are not specifically designed to meet the needs of individuals with visual impairments in VR environments. For example, the SeeingVR project, a Microsoft research initiative aimed at improving the accessibility of VR technology for individuals with visual impairments, has not yet progressed beyond the research stage.

In order to improve the accessibility of virtual reality (VR) for individuals with visual impairments, it is essential to provide a range of text size options, preferably controlled via a slider controller. This feature not only benefits individuals with significant vision loss but also those who wear glasses. Magnification and menu narration are additional critical features that are essential for low-vision and blind users. These features offer a promising starting point for enhancing the accessibility of VR for individuals with visual impairments. However, continued research and development are necessary to ensure that the needs of this population are fully met in the context of VR technology.

2.2 Sonifying the Environment

Within the broader field of sensory substitution, there exist several works. For instance, Daniel Kish, who is blind, has effectively utilized "mouth clicks" for autonomous navigation activities, such as riding and trekking, through the employment of accurate echolocation abilities. Similarly, Neil Harbisson, a colorblind artist, has constructed a device capable of converting color information into sound frequencies. Voice technology [Vines et al. 2019] has made significant attempts to transfer visual perception to sound through a system that analyzes each camera snapshot from left to right, linking height with pitch and brightness with loudness. However, the employment of these sensory substitution techniques can present a high learning curve. In contrast, our study employs visual recognition algorithms with deterministic grammars, which facilitate more straightforward and direct approaches to comprehend objects in a visual scene.

In recent years, researchers have investigated sonification methods to enhance access visual information for individuals with visual impairments. Such methods have been employed to improve accessibility to graphs [Walker and Mauney 2010][Constantinescu et al. 2020], explore maps and graphics [Geronazzo et al. 2016][Ahmetovic et al. 2019], and provide rotating instructions for ease of navigation in various environments [Aziz et al. 2019] by furnishing explicit details on specific features and points of interest. Ferati et al. have proposed Audemes, a prospective solution for the provision of educational spoken writings to visually impaired students, employing auditory icons [Ferati et al. 2011].

Aziz et al. [Aziz et al. 2019] conducted a study on the use of earcons and auditory icons as a sonification strategy for auditory route overviews via text-to-speech. The research found that aural icons were appropriate for conveying information about points of interest. In a similar vein, Tislar et al. [Urbanietz et al. 2019] analyzed the effects of sonification via different mediums, such as music, spearcons, and lyricons, to explore sound-relatedness, meaning attribution, and intuitiveness. Notably, the study failed to evaluate the efficacy of the sonification mediums on individuals with visual impairments, thereby limiting the generalizability of the findings. The varied preferences and cognitive demands of sighted and visually impaired individuals for user interfaces [Walker and Mauney 2010] further underscore the need for more focused research in this domain.

2.3 Audible Interfaces: Captioning Scenes

Dingler et al. [Dingler et al. 2008] conducted a study that bears resemblance to our research in terms of sonification. The study compared the efficacy of using earcons, auditory icons, spearcons, and speech for object representation and learning. However, the study failed to incorporate temporal constraints for the duration of sounds, which is crucial when presenting multiple items within a limited period. Some methods employ computer vision techniques to detect indoor and outdoor scenes and notify users via voice [Hub et al. 2003][Zeng et al. 2017] or vibrotactile feedback [Saha et al. 2019].

Katz et al. [Mascetti et al. 2016] and Presti et al. [Katz et al. 2012] described a novel technique for supporting wayfinding. Katz et al. worked to detect outdoor environments using computer vision technologies which are sonified using spatialized 3D sound. One crucial issue is that the technique only localises a single class of items at a time. Computer vision innovations, such as YOLO, have been used to assist users in an augmented, annotated, physical world [Gupta et al. 2022; Kumar and Jain 2022; Mahendru and Dubey 2021].

3 NARRATION PLATFORM DESIGN

3.1 Object Detections from Screenhots

The solution's essential components are the object detection model and the 3D scene rendering on the Web. YOLO was utilized to

address the object recognition issue as it has shown promising outcomes in image recognition. Our four algorithms required us to extract particular image parameters, including the confidence score, anchor point (x, y), height and width of the object, class name, and area, to design them. Extracting these critical details post-recognition provided us with the flexibility to execute our algorithms. Additionally, the object detection solution and narration algorithms were combined with the 3D scene on the Web environment rendered using X3DOM technology. Our approach utilizes X3DOM, a Web technology that enables the rendering of 3D environments on a browser, and integrates object-recognition algorithms to generate a natural language description of the scene.

3.2 Architecture

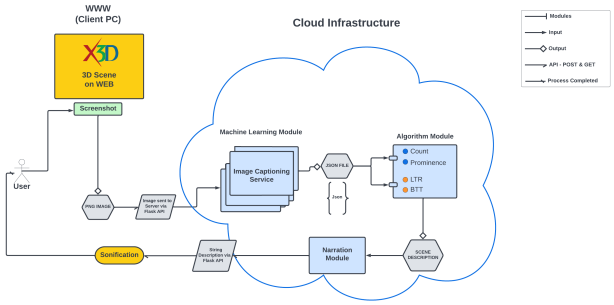

Figure 1: Architecture Diagram of the end-to-end solution.

The architecture design of our solution comprises several components that contribute to its functional end-to-end service. At the top left of the architecture diagram is X3DOM, which is a standard that aided us in rendering a 3D scene on the Web browser. Users can easily navigate the 3D scene and capture a screenshot using the green-colored button. The captured PNG image is then transmitted to the cloud infrastructure via APIs, which includes the machine learning module, algorithm module, and narration system. The machine learning module is responsible for extracting the required data from the input image and producing a JSON file. This JSON file is then served as input to the chosen algorithm type, which generates a corresponding description.

Subsequently, this description is conveyed to the narration module via APIs for the narration system to convert it into speech and present it to the user. To summarize, the narration platform's cloud infrastructure includes a machine learning module responsible for image captioning services, as well as an algorithm module. Upon processing the JSON file, the output generated by these modules is transferred to the Narration Module. This module is responsible for parsing the text of the scene description and using grammatical rules to form an accurate depiction. The resulting output is then passed to the speech or sonification plugin, which generates the audio output for the user.

3.3 X3DOM

X3DOM (XML3D in DOM) [Behr et al. 2009][Behr et al. 2011][Behr et al. 2010] is an open-source JavaScript framework that provides a declarative way to create and display 3D content on the web. It allows developers to create 3D graphics and animations by defining objects in XML and using HTML5 and JavaScript to manipulate them. X3DOM is built on top of the Document Object Model (DOM), which is used to represent HTML and XML documents as objects. This makes it easy to integrate 3D content with other web technologies and create interactive web applications. X3DOM is based on the Extensible 3D (X3D) standard, which is an open standard for representing 3D graphics on the web. X3D provides a rich set of features for creating and manipulating 3D content, including geometry, materials, textures, lighting, animation, and scripting. X3DOM extends the capabilities of X3D by providing a way to use X3D content in a webpage using standard HTML and JavaScript.

X3DOM is a powerful and flexible technology for creating and displaying 3D content on the web. It provides a declarative way to define 3D objects and animations using XML, and makes it easy to integrate 3D content with other web technologies using JavaScript and HTML. With X3DOM, developers can create interactive and immersive 3D applications that run in any modern web browser. In our application at a Web address in their browser, users navigate a 3D scene with an interactive perspective camera; at any point, they can click a button to take a screenshot of the scene. Then, the Web page sends those screenshots from the scene to the server where the captioning service and model reside and the narration is returned.

3.4 Object Detection Model - YOLO

The YOLO algorithm, 'You Only Look Once', detects objects in images through their bounding boxes. To achieve this, the algorithm employs R-CNN and other region proposal techniques to generate the bounding boxes on the images before executing the classifier on the proposed boxes. Using a single convolution neural network, the YOLO algorithm is able to predict multiple bounding boxes and class probabilities for those boxes simultaneously. This approach simplifies the object detection process, making it more efficient to deploy. This is because the detection problem is defined as a regression problem, eliminating the need for a complex pipeline. The YOLO algorithm is an efficient and effective method for detecting objects in images. Its use of a single convolution neural network to predict bounding boxes and class probabilities makes it a simple yet powerful tool for object detection.

Convolutional Neural Network of the YOLO Model. YOLO is implemented as a convolutional neural network and often evaluated by the PASCAL VOC detection dataset [Everingham et al. 2015]. The network comprises early convolutional layers that extract visual features and fully connected layers that predict output probabilities and coordinates. The model's architecture is based on the GoogLeNet image categorization model [Szegedy et al. 2014]. The YOLO network consists of 24 convolutional layers, which are followed by two fully connected layers. Unlike the GoogLeNet inception modules, the authors employed 1x1 reduction layers followed by 3x3 convolutional layers [Lin et al. 2013]. The ImageNet 1000-class competition dataset was employed to train the convolution layer of the YOLO model [Russakovsky et al. 2014]. The authors utilized the first twenty convolution layers for pre-training, followed by an average-pooling layer and a fully-connected layer. The

model was trained over several weeks and achieved a top 5% accuracy when evaluated on the ImageNet 2012 validation set, which is comparable to the performance of GoogleNet models. The Darknet framework was used for both inference and training.

Limitations of YOLO. The YOLO algorithm is limited spatially due to each cell's predictions of only two boxes and having a single class [Redmon et al. 2015]. This limitation restricts the number of objects that the algorithm can predict and poses challenges when dealing with group objects. Moreover, the model's generalizability is not particularly robust when presented with new or uncommon aspect ratios. In addition during training, the loss function used to estimate detection performance treats errors in small and large bounding boxes equally. However, a minor error in a large box is typically inconsequential while a small error in a small box can significantly impact the Intersection over Union (IOU) metric. Incorrect localizations are the primary source of errors in the model. YOLO results are returned as a JSON object that includes the objects detected (name), their bounding box in the image, and the confidence rating of each identification.

3.5 Text-To-Speech

In our implementation, we utilized the text-to-speech (TTS) method on the client side (browser) for the narration of the scene in our X3D web application. As JavaScript does not possess a built-in TTS application programming interface (API), we accessed the TTS API provided by most modern web browsers through JavaScript. The Web Speech API proved particularly beneficial in providing speech synthesis functionality for TTS, which enabled us to incorporate TTS into our web application.

We were able to access this API through any modern web browser, including Google Chrome, Mozilla Firefox, Microsoft Edge, and Apple Safari. To leverage the Web Speech API, we instantiated the SpeechSynthesis object and established the various properties of the SpeechSynthesisUtterance object, such as the voice and speech rate. Then, we pass our algorithmically-generated text description to the Utterance objects, which prompts the browser to articulate the text description using the specified voice and speech rate.

4 CAPTION GRAMMARS

From the user's screenshot of the 3D scene, the YOLO service returns its object detection results in a JSON package. The data includes the name, bounding box, and confidence score for each detection. From this result array, we want to generate a narration string that describes the scene. This string can then be given to a Text-to-Speech application to be read out loud to the user (rendered sonically). In this section, we describe the design process and rationale for each algorithm.

The primary goal of these narration algorithms is to provide situational awareness to the user: to transcode the visual signals of the scene into an audio description. Through multiple ideation sessions, we brainstormed scenarios and considered how the YOLO detection on the screenshot could be formed into an English caption (a string) that could be rendered with text-to-speech. First, we considered the question **"What is in the scene?"**. This led a set of algorithm ideas that centered on the object data, such as their name, count, and size in the image plane. Algorithm 1 counts the objects in a scene and then lists them in order of frequency: highest to lowest. Algorithm 2 lists the objects in order of size from largest to smallest.

Second, we considered how narration algorithms could support awareness of **"Where are things in the scene?"**. Algorithm 3 uses the spatial metaphor of reading European languages from left to right and lists objects in this order. Without explicit depth information in our screenshot of the image plane, we also considered the case where objects closer to the user could be more relevant. Thus, Algorithm 4 reads the object detections from bottom to top, which presumes that objects lower in the frame are closer to the user and therefore more relevant.

It is notable that the captioning algorithm designs break down roughly along the lines of results from cognitive neuroscience. Namely that the ventral pathway of processing visual information is generally concerned with objects and recognition. In contrast, the dorsal pathway processes action and attention-based aspects of the visual scene [Zhan et al. 2018]. In the following section, we provide details for each of the 4 grammars tested.

- **"Object-Centric"**
  Algorithm One - (Count): This algorithm gives us the exact Count of the different objects in the view. This count algorithm is a part of the "Object Centric" algorithms category.

  Algorithm Two - (Prominence): The Prominence algorithm gives us the results based on the largest to the smallest area of the objects in the view. This count algorithm is a part of the "Object Centric" algorithms category.

- **"Environment-Centric"**
  Algorithm Three - (LTR): This algorithm is called Left-To-Right. The algorithm is a part of the spatial environment-centric category. It utilises the coordinate geometry rules to decide the positions of the objects in the scene from the left spatial point to the right spatial point.

  Algorithm Four - (BTT): This algorithm is called a Bottom-To-Top. The algorithm is a part of the spatial environment-centric category. It utilises the coordinate geometry rules to decide the positions of the objects in the scene from the bottom (closest in the scene) spatial point to the topmost spatial point (farthest).

After constructing our narration string from the YOLO detection, we pass this string to the Text-To-Speech API, which can be customized for voices of different genders and nationalities. At this point, the narration plays through the system's audio outputs.

5 EVALUATION

Our study aimed to compare four different narration algorithms through user ratings across randomly selected 3D worlds on the Web. The data collected through this study was analyzed using comprehensive statistical techniques to answer our research questions and test our hypothesis. Adult participants were presented with several 3D worlds and asked to rate them based on their subjective experience. The study was conducted on the Web, and the

selection of worlds presented to the participants was randomized to prevent any ordering bias. We used four different algorithms to generate the world narrations, and we compared the ratings given by participants. These algorithms are presented a pseudocode above.

Based on the literature and our algorithm design, we hypothesize that:

- **Hypothesis One**: "Object-centric" algorithms will be better for Search tasks.
- **Hypothesis Two**: "Environment-centric" algorithms will be better for Summarize tasks.

5.1 User Study Design

The user study included the object detection algorithm, X3DOM's [Behr et al. 2009] implementation for supporting 3D scenes on the Web, and APIs connecting the system from end to end. The study application employed X3DOM's [Behr et al. 2011] implementation to run the 3D worlds on desktop/laptop web browsers, and it was compatible with browsers such as Chrome, Safari, Firefox, and Opera. Therefore, the user study design involved subjects interacting with the scene via their desktop or laptop web browsers, capturing a postcard image (screenshot), hearing narrations, and rating those narrations.

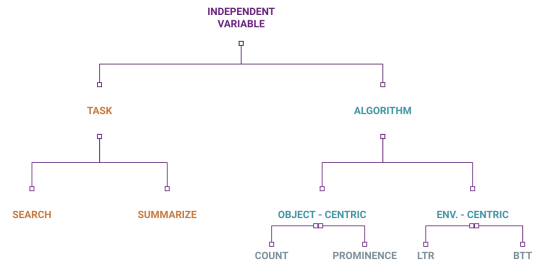

Figure 2: ANOVA Variable Tree

The first independent variable is the user's task when presented with the virtual environment. The task types are:

(1) **Search**: The user is asked a question to find relevant objects in the scene. The relevant objects can be any object present in the environment. For example: *"Find the chair/chairs in the scene and take a picture."*
(2) **Summarize**: This is the second task where the user is asked to *"Get a wider view of the scene and take a picture."* The user zooms out and finds a viewpoint that can be considered as a wider view where they can see multiple objects clearly.

The second independent variable was algorithm by type (Figure 3). A survey system was designed to conduct the study. The survey system presented users with random scenes generated by random algorithms and randomly assigned task types. It was ensured that each user had a different combination of tasks, worlds, and algorithms, ensuring complete impartiality in the study. The random assignments of questions begin by randomizing the task type. Each user participated in 24 trials, and each trial involved a randomly selected scene from a pool of eight 3D worlds. Additionally, each trial featured a randomly selected algorithm for narration. In Figure 4, there is a representation of one of the worlds. A total of eight worlds were included, and users performed 24 trials in total (3 trials per condition).

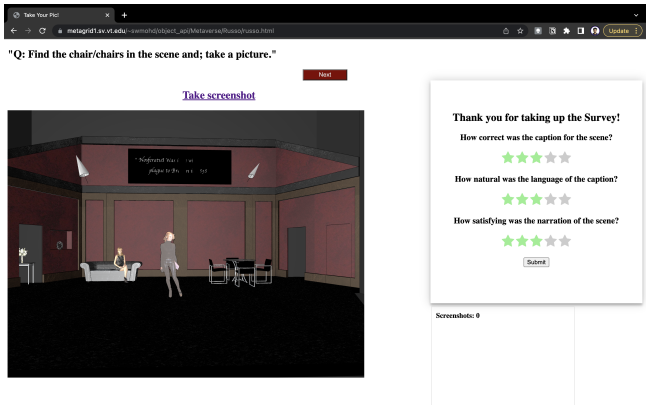

Figure 3: Stimuli example: Task Type Search; scene courtesy VT Theater Department (Chris Russo)

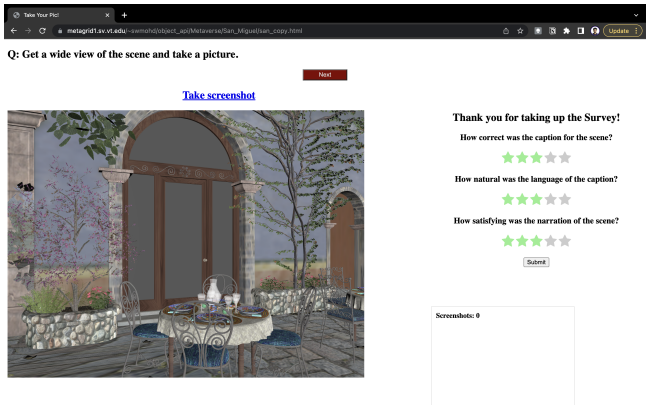

Figure 4: Stimuli example: Task Type Summarize

The second independent variable in this experiment was Algorithm. We designed four narration algorithms in two categories: Object-centric (Count, Prominence) and Environment-centric (LTR, BTT). The algorithmic narrations for the above screenshots are as follows:

- COUNT: 'There are: THREE CHAIRS.' (Fig 4)
- PROMINENCE: 'The most prominent objects from biggest to smallest are: DINING TABLE, CHAIR, PLATE, WINE GLASS.' (Fig 5)
- LEFT-TO-RIGHT (LTR): 'The objects in the image from left to right are: SOFA, PERSON, PERSON, CHAIR, TABLE.' (Fig 4)
- BOTTOM-TO-TOP (BTT): 'The objects in the image from bottom to top are: CHAIR, DINING TABLE, PLATE, AND WINE GLASS.' (Fig 5)

Table 1: Questions asked for each trial

Survey Questions
How correct was the caption for the scene?
How natural was the language of the caption?
How satisfying was the narration of the scene?

Table 2: Language Distribution

Languages	Data
English	58.97%
Hindi	12.82%
Spanish	5.13%
French	2.56%
Arabic	1.28%
Other	19.23%

5.2 Evaluation Procedure

After listening to narrations generated by four different algorithms, the users were asked three questions (Table 1). Before starting the survey, the users were asked to fill out a pre-study questionnaire which consisted of questions related to demographics, gender, age etc. As the user study was conducted online, participants were required to navigate the scene in order to capture the appropriate screenshot as per the questions asked, using the 'take screenshot' button. To facilitate this process, a pre-survey training session was conducted to familiarize the users with the scene navigation.

During the training session, users were instructed on the proper use of the controls and the steps required for the survey. The controls were explained to the users both verbally and in written format. Key instructions were provided to ensure that all users were adequately prepared for 3D navigation and scene controls. When the user affirmed they understood the controls, the study proceeded. Users were shown 24 random 3D scene and algorithm combinations, each with a rating form.

6 RESULTS AND ANALYSIS

6.1 Study Demographics

In this study, data was collected from 24 trials that were presented to each of the 44 volunteer participants recruited from Virginia Tech; all were age 18 years or older. Recruitment was carried out via email, with the consent form and associated terms and conditions shared with potential participants. Prior to commencing the survey, all participants were asked to complete a demographic questionnaire (Tables 2 and 3). In this study, data regarding the vision status of participants was collected. The recruitment process did not exclude individuals with normal vision, and efforts were made to include participants with and without vision problems, as well as those who use corrective lenses. This approach was taken to evaluate the robustness of the system in terms of its ability to accommodate the needs of all users, regardless of their vision status. The inclusion of participants with normal vision also allowed for a more comprehensive evaluation of the system's efficacy.

Table 3: Vision Distribution

Vision	Data
Myopia	54.35%
Hyperopia	2.17%
Normal	43.48%

6.2 Two-Way ANOVA with Repeated Measures

This is a 2x4 within-subjects experimental design with two independent variables: task and algorithm. The study was repeated measures and randomized conditions where users experienced 3 trials under each condition. The dependent variables were users' subjective ratings of correctness, naturalness, and satisfaction. We used repeated measures ANOVA to determine the effects of both task types and algorithms on user ratings. From two-way ANOVA, we can test three hypotheses. The Null Hypothesis being:

(1) There is no significant effect of task type on user ratings of the narration.
(2) There is no significant effect of algorithm type on user ratings of the narration.
(3) There is no significant interaction of tasks and algorithms both on user ratings of the narration.

Figures 6 and 7 show how the user ratings of Correctness were found to be significantly influenced by the algorithms *($p = .003$)* and their interaction with the task *($p < .001$)*. However, the task type alone did not have a significant effect on the ratings. The analysis revealed that the Count, Prominence, and Bottom-To- Top algorithms consistently performed well in all tasks. Among them, the Prominence and Bottom-To-Top algorithms received marginally higher ratings in the Search tasks. On the other hand, the Left-to-Right algorithm received poor ratings in the Search tasks but significantly improved to achieve competitive ratings in the Summarize tasks (Figure 7).

Similar rating patterns were observed for the user ratings of Naturalness, with algorithms playing a significant role *($p < .001$)* and dominating the interaction effect with the task *($p = .036$)*. However, the task type alone did not have a significant effect on the ratings. The analysis revealed that the Count, Prominence, and Bottom-To-Top algorithms consistently performed well in all tasks. The Left-to-Right algorithm received poor ratings in the Search tasks, but it improved significantly to achieve moderate ratings in the Summarize tasks (Figures 8 and 9). The algorithm 1 (Count) performed best overall.

The user ratings of Satisfaction were found to be significantly influenced by the algorithms *($p < .001$)* and their interaction with the task *($p < .001$)*. Additionally, the task type had a weakly significant effect *($p = .046$)*, with users being less satisfied with the Summarize tasks. Based on the analysis, it was observed that the Count, Prominence, and Bottom-Up algorithms consistently performed well in all tasks. Among them, the Prominence and Bottom-Up algorithms received slightly higher ratings in the Search tasks. On the other hand, the Left-to-Right algorithm received poor ratings in the Search tasks, but it significantly improved to achieve competitive ratings in the Summarize tasks (Figures 10 and 11).

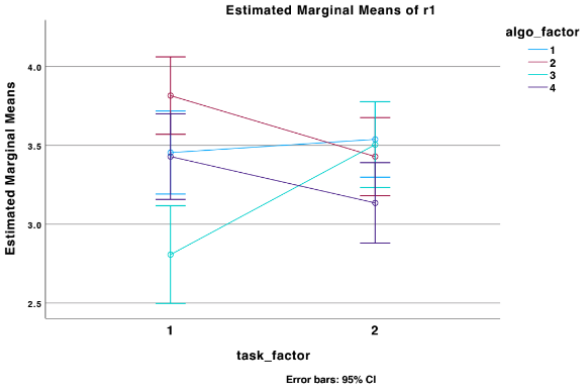

Figure 5: Task vs Algorithm Performance for Correctness Rating

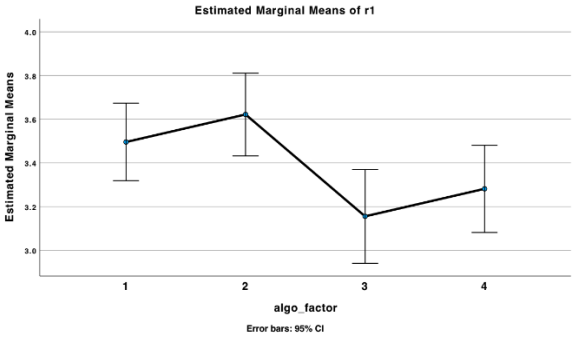

Figure 6: Algorithm Performance for Correctness Rating

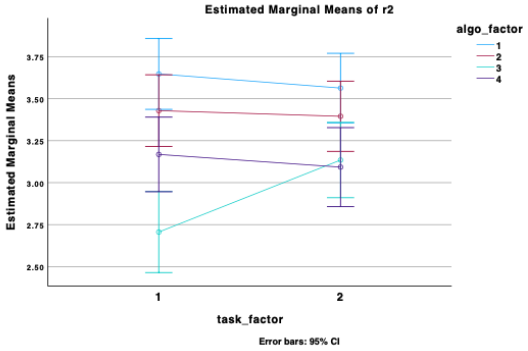

Figure 7: Task vs Algorithm Performance for Naturalness Rating

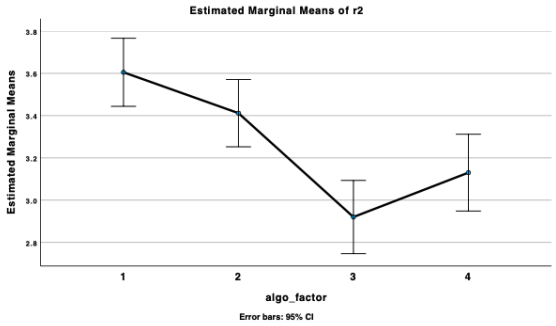

Figure 8: Algorithm Performance for Naturalness Rating

Based on the analysis of all tasks, trials, and subjective ratings, it was observed that the Object-centric algorithms outperformed the Environment-centric algorithms. Furthermore, among the Object-centric algorithms, both the Count and Prominence algorithms performed equally well. Conversely, the Left-Right algorithm was rated the lowest among all the algorithms. These results were consistent across all participants and were particularly evident among the myopic participants.

7 POST-HOC TESTS

Following the ANOVA analysis, we see that different algorithms and tasks have a statistically significant effect on user ratings. In order to parse out the individual effects of each group, Tukey's HSD and the General Linear Model were employed to identify significant pairs formed between the task and algorithm categories. This involved conducting multiple pairwise comparisons, commonly referred to as Post-Hoc comparisons.

It is worth noting that for the purpose of this study, algorithm 1 and algorithm 2 (Count and Prominence) were grouped under a single category termed 'Object centric', while algorithm 3 and algorithm 4 (Left to Right and Bottom to Top) were grouped under 'Environment centric' or Spatial Algorithms. Furthermore, both algorithm categories were tested in conjunction with the search tasks to evaluate the impact of each category on the ratings. The graphs and table presented in the following section indicate the results of these analyses, where OC represents Object-centric and EC represents Environment-centric.

7.1 Captioning Grammar by Task Type

Post-hoc tests are conducted to determine whether there are significant differences between groups when the ANOVA indicates a significant result. These tests control the error rate, either between groups or family-wise. Post-hoc tests adjust the p-values (using the Bonferroni adjustment) or critical values (using Tukey's HSD) to ensure appropriate levels of statistical significance. The purpose of post-hoc tests is to identify which specific groups exhibit significant differences, as the ANOVA only indicates the presence of differences between groups. By controlling the error rate, post-hoc tests help to reduce the likelihood of false positives or type I errors, and thus provide more reliable and accurate results.

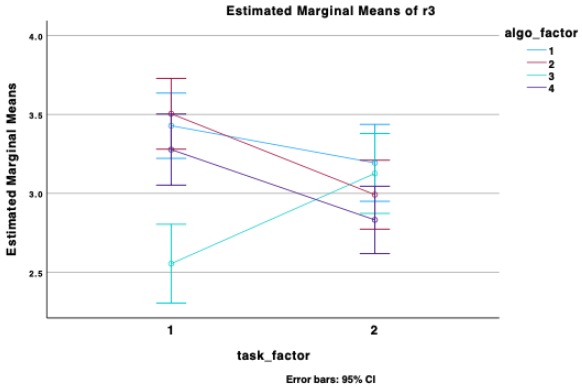

Figure 9: Task vs Algorithm Performance for Satisfaction Rating

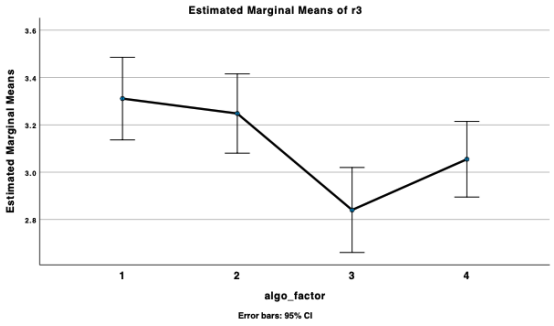

Figure 10: Algorithm Performance for Satisfaction Rating

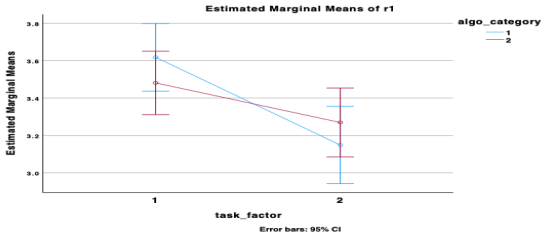

Figure 11: Group Impact on Rating One - Correctness

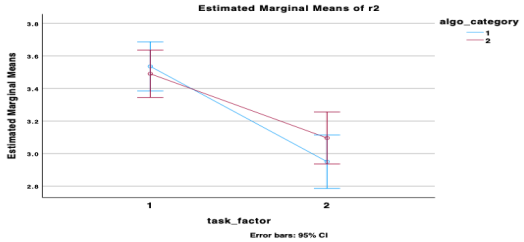

Figure 12: Group Impact on Rating Two - Naturalness

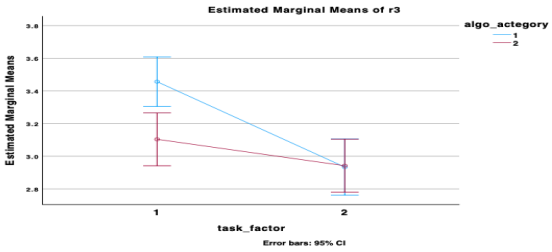

Figure 13: Group Impact on Rating Three - Satisfaction

In the post-hoc analysis, we aggregated the algorithm types into two categories: Object-centric (Count and Prominence) and Environment-centric (Left-to-Right and Bottom-to-Top) algorithms. For the Correctness ratings, the task type had a significant effect *($p < .001$)*, with users rating the Summarize tasks lower than the Search tasks (Figure 12). Similarly, for the ratings of Naturalness, users rated all interfaces significantly lower in the Summarize tasks *($p < .001$)* (Figure 13). In the user ratings of Satisfaction, the task had a significant effect, with the Summarize tasks being rated lower overall *($p < .001$;)*. Additionally, there was a significant effect of the algorithm, where Object-centric algorithms were rated significantly better (more satisfying) than Environment-centric algorithms for Search tasks *($p = .012$)*.

7.2 Descriptive Statistics

In order to examine the distribution of responses, we calculated descriptive statistics. These showed that algorithm 2 (Prominence) received consistently higher ratings for the search task in terms of correctness. On the other hand, algorithm 4 (BTT) was perceived as performing relatively better for the summarize task. The results suggest that algorithm 1 (Count) is the highest-performing algorithm for both task types, followed by algorithm 2. In contrast, algorithm 3 (LTR) consistently received lower ratings for both the search and summarize tasks.

The results indicate that algorithm 2 (Prominence) received consistently higher ratings for the search task in terms of satisfaction, followed by algorithm 1 (Count). For the summarize task, all four algorithms received consistently equal ratings. However, algorithm 3 (LTR) was perceived as performing worse for the summarize task, followed by algorithm 4 (BTT).

Based on the correlation coefficients and significant values, it can be concluded that there is a strong correlation between vision status and ratings two and three (Naturalness and Satisfaction). These findings suggest that the ratings received are significantly

effected by user vision status, indicating an increased subjective value for narrations in that population.

Vision Status and Rating Two (Naturalness). The results indicate that vision status group 4 provided the highest ratings for the naturalness of the narration, with over 50% of the group rating the narration between 3 to 5. The majority of users within this group rated the narration a 3. Approximately 30% of users with normal vision rated the naturalness of the narration as 3. Over 50% of the group rated the narration 3 or higher. These findings suggest that individuals with normal vision generally found the naturalness of the description to be good.

Vision Status and Rating Three (Satisfaction). We explored the various vision categories and their corresponding ratings for satisfaction, which pertains to how satisfied users were after hearing the narration from the system. The results suggest that users within vision status group 4 (Myopia) reported the highest satisfaction ratings for the narration, with over 60% of the group providing a rating of 3 or higher. The most common rating provided by users in this group was a 3; however, a notable number of users within this group expressed satisfaction with the caption they heard. Given the results indicating that the Myopic user group reported higher ratings for naturalness and satisfaction, we conducted a deeper analysis to determine which algorithms performed best for this group and received the highest ratings for different rating questions overall. Algorithm 1 (Count) was best in the Naturalness rating, as rated by Myopic (vision 4) group participants. Algorithm 1 (Count) was best in Satisfaction rating, as rated by Myopic (vision 4) group participants.

8 CONCLUSION AND FUTURE WORK

8.1 Contributions

We have developed an audio captioning service that incorporates YOLO for object detection and X3DOM's capabilities for 3D rendering in a Web page. The navigation feature of the solution enables users to view the scene from their own perspective, take screenshots of the scene, and receive captions based on one of four captioning algorithms: count, prominence, left-to-right, and bottom-to-top. To evaluate the narration performance of the system with this diverse set of algorithms and narration techniques, a user study was conducted with 44 participants recruited from Virginia Tech. The online study was designed to assess the system's performance with different combinations of algorithms and scenes.

Our study aimed to investigate which type of grammar is better for captioning a scene for Summarize tasks and Search tasks. Upon analyzing the ratings and conducting statistical analysis, we discovered that participants on average rated the Summarize task type higher than the Search tasks. We also observed that the ratings for spatial algorithms, namely algorithm 3 and algorithm 4, were higher for Summarize tasks than for Search tasks. Therefore, we can conclude that spatial or Environment-centric algorithms are more effective for comparison or Summarization tasks. Additionally, we found that the Environment-centric Left-To-Right algorithm received the lowest ratings overall.

We also observed that the Object-centric Count and Prominence captions (algorithm 1 and algorithm 2) were rated better for Search tasks. On the other hand, the Environment-centric captions (algorithm 3 and algorithm 4) received higher ratings in the Summarize task type. Specifically, Environment-centric captions performed better than Object-centric captions in the Summarization tasks. We also found that the Environment-centric Left-To-Right captions received the lowest ratings. These findings support our hypothesis that Object-based caption grammars are better for Search tasks, while Environment-based caption grammars are better for Summarize tasks. These results are promising, and may provide an advantageous paradigm to further develop deterministic audio caption grammars for 3D scene content.

8.2 Limitations

The online user study was conducted on desktop computers using a Web browser. In order to optimize scene rendering on the browser, certain strategies were enforced, such as reducing texture quality and optimizing geometry to achieve better performance in less time. However, due to the reduced texture quality in some scenes, certain objects could not be accurately identified by the algorithms, leading to a decrease in system accuracy. Additionally, navigating the scene on the browser with a trackpad or mouse proved to be challenging for some users; in addition, some scenes took several seconds to load, impacting the overall experience for remote users.

This work demonstrates an end-to-end service-based method for captioning Web3D scenes; similarly 'the Metaverse'. We have simply shown a method - auditory postcards, which can be improved in many ways; for example, our tasks and task types are still coarse, we don't use client-side depth buffer information, or cloud-based large language models; these are topics of future work. We collected information about our participants' vision status and performed an analysis to investigate the correlation between the vision group and the algorithm of choice. However, as we did not gather specific data about the severity or degree of visual impairment, we cannot draw any definitive conclusions at this time. Therefore, future research could explore the user's captioning preferences based on their task context and the type and magnitude of their visual impairment.

8.3 Future Work

In the future, we will utilize Head-Mounted Displays (HMDs) for both virtual and augmented environments where users can immerse or augment the scene and interact with the objects while hearing the narrations. Most object-detection systems can only process a single image plane capture (not stereo). However, with access to the depth buffer, and/or a depth camera in an AR headset, we believe accuracy and satisfaction could be improved. The captions could also be improved by utilizing more sophisticated natural language processing models on top of the pre-trained network for object detection used in this study. However, great care must be taken in the grammatical construction, since captions must be reliable, accurate, and reproducible. The ultimate goal would be to generate more accurate and natural descriptions for users to access and understand their 3D environment and to complete their tasks.

REFERENCES

Dragan Ahmetovic, Federico Avanzini, Adriano Baratè, Cristian Bernareggi, Gabriele Galimberti, Luca A Ludovico, Sergio Mascetti, and Giorgio Presti. 2019. Sonification of rotation instructions to support navigation of people with visual impairment. In

2019 IEEE International Conference on Pervasive Computing and Communications (PerCom. IEEE, 1–10.

Jesse Anderson. 2022. *How Can a Blind Person Use Virtual Reality.* Retrieved February 21, 2023 from https://equalentry.com/how-can-a-blind-person-use-virtual-reality/

Nida Aziz, Tony Stockman, and Rebecca Stewart. 2019. An investigation into customisable automatically generated auditory route overviews for pre-navigation. Georgia Institute of Technology.

Johannes Behr, Peter Eschler, Yvonne Jung, and Michael Zöllner. 2009. X3DOM: A DOM-Based HTML5/X3D Integration Model. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/1559764.1559784

Johannes Behr, Yvonne Jung, Timm Drevensek, and Andreas Aderhold. 2011. Dynamic and Interactive Aspects of X3DOM. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/2010425.2010440

J. Behr, Y. Jung, J. Keil, T. Drevensek, M. Zoellner, P. Eschler, and D. Fellner. 2010. A Scalable Architecture for the HTML5/X3D Integration Model X3DOM. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/1836049.1836077

Pramod Chundury, Biswaksen Patnaik, Yasmin Reyazuddin, Christine Tang, Jonathan Lazar, and Niklas Elmqvist. 2021. Towards understanding sensory substitution for accessible visualization: An interview study. *IEEE transactions on visualization and computer graphics* 28, 1 (2021), 1084–1094.

Angela Constantinescu, Karin Müller, Monica Haurilet, Vanessa Petrausch, and Rainer Stiefelhagen. 2020. Bring the environment to life: A sonification module for people with visual impairments to improve situation awareness. In *Proceedings of the 2020 International Conference on Multimodal Interaction*. 50–59.

Tilman Dingler, Jeffrey Lindsay, and Bruce N Walker. 2008. Learnabiltiy of sound cues for environmental features: Auditory icons, earcons, spearcons, and speech. International Community for Auditory Display.

Niklas Elmqvist. 2023. *Visualization for the Blind.* Retrieved February 18, 2023 from https://interactions.acm.org/archive/view/january-february-2023/visualization-for-the-blind

Mark Everingham, S. M. Eslami, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. 2015. The Pascal Visual Object Classes Challenge: A Retrospective. 111, 1 (2015). https://doi.org/10.1007/s11263-014-0733-5

Mexhid Ferati, Steve Mannheimer, and Davide Bolchini. 2011. Usability evaluation of acoustic interfaces for the blind. In *Proceedings of the 29th ACM international conference on Design of communication*. 9–16.

Michele Geronazzo, Alberto Bedin, Luca Brayda, Claudio Campus, and Federico Avanzini. 2016. Interactive spatial sonification for non-visual exploration of virtual maps. *International Journal of Human-Computer Studies* 85 (2016), 4–15.

Sneha Gupta, Suchismita Chakraborti, R Yogitha, and G Mathivanan. 2022. Object Detection with Audio Comments using YOLO v3. In *2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*. IEEE, 903–909.

Andreas Hub, Joachim Diepstraten, and Thomas Ertl. 2003. Design and development of an indoor navigation and object identification system for the blind. *ACM Sigaccess Accessibility and Computing* 77-78 (2003), 147–152.

Brian FG Katz, Slim Kammoun, Gaëtan Parseihian, Olivier Gutierrez, Adrien Brilhault, Malika Auvray, Philippe Truillet, Michel Denis, Simon Thorpe, and Christophe Jouffrais. 2012. NAVIG: Augmented reality guidance system for the visually impaired: Combining object localization, GNSS, and spatial audio. *Virtual Reality* 16 (2012), 253–269.

Nitin Kumar and Anuj Jain. 2022. A Deep Learning Based Model to Assist Blind People in Their Navigation. *Journal of Information Technology Education: Innovations in Practice* 21 (2022), 095–114.

Min Lin, Qiang Chen, and Shuicheng Yan. 2013. Network In Network. https://doi.org/10.48550/ARXIV.1312.4400

Mansi Mahendru and Sanjay Kumar Dubey. 2021. Real time object detection with audio feedback using Yolo vs. Yolo_v3. In *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. IEEE, 734–740.

Sergio Mascetti, Lorenzo Picinali, Andrea Gerino, Dragan Ahmetovic, and Cristian Bernareggi. 2016. Sonification of guidance data during road crossing for people with visual impairments or blindness. *International Journal of Human-Computer Studies* 85 (2016), 16–26.

Nicholas Polys, Jessica Hotter, Madison Lanier, Laura Purcell, Jordan Wolf, W. Cully Hession, Peter Sforza, and James D. Ivory. 2017. Finding Frogs: Using Game-Based Learning to Increase Environmental Awareness. In *Proceedings of the 22nd International Conference on 3D Web Technology* (Brisbane, Queensland, Australia) *(Web3D '17)*. Association for Computing Machinery, New York, NY, USA, Article 10, 8 pages. https://doi.org/10.1145/3055624.3075955

Nicholas Polys, Cecile Newcomb, Todd Schenk, Thomas Skuzinski, and Donna Dunay. 2018. The Value of 3D Models and Immersive Technology in Planning Urban Density. In *Proceedings of the 23rd International ACM Conference on 3D Web Technology* (Poznań, Poland) *(Web3D '18)*. Association for Computing Machinery, New York, NY, USA, Article 13, 4 pages. https://doi.org/10.1145/3208806.3208824

Nicholas F Polys, Kathleen Meaney, John Munsell, and Benjamin J Addlestone. 2021. X3D Field Trips for Remote Learning. In *The 26th International Conference on 3D Web Technology*. 1–7.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2015. You Only Look Once: Unified, Real-Time Object Detection. https://doi.org/10.48550/ARXIV.1506.02640

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2014. ImageNet Large Scale Visual Recognition Challenge. https://doi.org/10.48550/ARXIV.1409.0575

Manaswi Saha, Alexander J Fiannaca, Melanie Kneisel, Edward Cutrell, and Meredith Ringel Morris. 2019. Closing the gap: Designing for the last-few-meters wayfinding problem for people with visual impairments. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*. 222–235.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2014. Going Deeper with Convolutions. https://doi.org/10.48550/ARXIV.1409.4842

Christoph Urbanietz, Gerald Enzner, Alexander Orth, Patrick Kwiatkowski, and Nils Pohl. 2019. A radar-based navigation assistance device with binaural sound interface for vision-impaired people. Georgia Institute of Technology.

Karen Vines, Chris Hughes, Laura Alexander, Carol Calvert, Chetz Colwell, Hilary Holmes, Claire Kotecki, Kaela Parks, and Victoria Pearson. 2019. Sonification of numerical data for education. *Open Learning: The Journal of Open, Distance and e-Learning* 34, 1 (2019), 19–39.

Bruce N Walker and Lisa M Mauney. 2010. Universal design of auditory graphs: A comparison of sonification mappings for visually impaired and sighted listeners. *ACM Transactions on Accessible Computing (TACCESS)* 2, 3 (2010), 1–16.

WHO. 2022. *Vision Impairment and blindness.* Retrieved February 18, 2023 from https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment

Limin Zeng, Markus Simros, and Gerhard Weber. 2017. Camera-based mobile electronic travel aids support for cognitive mapping of unknown spaces. In *Proceedings of the 19th international conference on human-computer interaction with mobile devices and services*. 1–10.

Minye Zhan, Rainer Goebel, and Beatrice de Gelder. 2018. Ventral and Dorsal Pathways Relate Differently to Visual Awareness of Body Postures under Continuous Flash Suppression. *eNeuro* 5, 1 (2018). https://doi.org/10.1523/ENEURO.0285-17.2017 arXiv:https://www.eneuro.org/content/5/1/ENEURO.0285-17.2017.full.pdf