



# Understanding How Blind Users Handle Object Recognition Errors: Strategies and Challenges

Jonggi Hong

Department of Computer Science  
Stevens Institute of Technology  
Hoboken, NJ, USA  
jhong8@stevens.edu

Hernisa Kacorri

College of Information, UMIACS  
University of Maryland, College Park  
College Park, MD, USA  
hernisa@umd.edu

## Abstract

Object recognition technologies hold the potential to support blind and low-vision people in navigating the world around them. However, the gap between benchmark performances and practical usability remains a significant challenge. This paper presents a study aimed at understanding blind users' interaction with object recognition systems for identifying and avoiding errors. Leveraging a pre-existing object recognition system, URCam, fine-tuned for our experiment, we conducted a user study involving 12 blind and low-vision participants. Through in-depth interviews and hands-on error identification tasks, we gained insights into users' experiences, challenges, and strategies for identifying errors in camera-based assistive technologies and object recognition systems. During interviews, many participants preferred independent error review, while expressing apprehension toward misrecognitions. In the error identification task, participants varied viewpoints, backgrounds, and object sizes in their images to avoid and overcome errors. Even after repeating the task, participants identified only half of the errors, and the proportion of errors identified did not significantly differ from their first attempts. Based on these insights, we offer implications for designing accessible interfaces tailored to the needs of blind and low-vision users in identifying object recognition errors.

## CCS Concepts

• **Human-centered computing** → **Human computer interaction (HCI)**; **Empirical studies in interaction design**; **Ubiquitous and mobile devices**.

## Keywords

object recognition errors, camera-based assistive technology, blind, visual impairment

### ACM Reference Format:

Jonggi Hong and Hernisa Kacorri. 2024. Understanding How Blind Users Handle Object Recognition Errors: Strategies and Challenges. In *The 26th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '24)*, October 27–30, 2024, St. John's, NL, Canada. ACM, New York, NY, USA, 15 pages. <https://doi.org/10.1145/3663548.3675635>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).  
*ASSETS '24*, October 27–30, 2024, St. John's, NL, Canada

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 979-8-4007-0677-6/24/10  
<https://doi.org/10.1145/3663548.3675635>

## 1 Introduction

The field of computer vision has made significant strides, achieving considerable benchmarking rates in object recognition tasks. Yet, despite these advancements, real-world applications often encounter substantial discrepancies between expected and observed performance [21, 62]. Factors such as complex tasks, resource limitations (e.g., mobile device processing), and inputs that deviate from the training data (e.g., classifying images with personal items or cluttered backgrounds collected by a user) pose persistent challenges, leading to higher-than-anticipated error rates in practical scenarios [9]. Moreover, the vulnerability of object recognition systems to adversarial attacks further compounds these challenges [31, 47]. While image classifiers hold potential for supporting the blind community in day to day tasks, they are hindered by their inability to effectively convey recognition errors, especially when tactile or olfactory verification is impractical (e.g., distant objects or scenes). Thus, despite advancements, the gap between benchmark performance and real-world usability remains a critical concern for assistive object recognition systems.

In this work, we explore the challenges that blind users face when handling object recognition errors and the strategies they use to overcome them. Specifically, we conduct a user study with 12 blind and low-vision participants, using a two-pronged approach: a semi-structured remote interview and a hands-on error identification task in participants' homes. In the interview, we aim to answer the following research question: "What are the experiences of blind and low-vision users with error handling in camera-based assistive technologies?" Participants describe how often they verify recognition results, the frequency with which they encounter errors, the importance they place on these errors, and the challenges they face in identifying them. To better contextualize their responses, we discuss their confidence in photo composition, the frequency of use, and the purposes for each of their camera-based assistive technologies. The interview is then followed by the experiment with an error identification task, where we aim to answer the following research questions: "How do blind and low-vision users identify and respond to object recognition errors, and what are the relationships between recognition error types, decision-making time, confidence levels, and task repetition?" Participants interact twice with URCam, an object recognition iOS app that we developed for this experiment. We fine-tuned the underlying model to recognize 15 object stimuli relevant to our study. The app provides object labels or a 'Don't know' response when the recognition confidence is low. To better contextualize the results, we report the accuracy of URCam during the task and manually code the strategies participants use for capturing photos when URCam responds with 'Don't know.'

Findings from our study provide insights on blind users' interaction with error-prone object recognition technologies. Interviews indicate that many participants preferred to independently review photo quality and identify errors in camera-based assistive technologies. They often triangulated information using contextual cues, their remaining vision, multiple trials, or other AI apps, rather than seeking sighted assistance. Although the frequency of encountered errors varied among participants, most expressed concern about misrecognitions. However, some did not report difficulty in identifying these errors. During the error identification task, we observed that participants could identify, on average, only half of the errors, with most of these being false positives. Notably, participants strategically adjusted viewpoints, backgrounds, and object size to avoid the "Don't know" predictions, often rotating objects or the camera to reveal different angles. We found that participants tended to make decisions more quickly when they felt more confident about the accuracy of the predictions. Comparing participants' first and second attempts at the same task, we did not observe a significant difference in the proportion of errors identified. However, there was a notable decrease in time spent to make a decision during the second attempt. Additionally, participants' certainty regarding recognition correctness decreased in subsequent attempts, attributed mainly to inconsistent recognition outcomes among similar objects.

The contributions of this work are the following: (1) Providing insights into blind users' experiences in assessing the quality of photos and handling errors in camera-based assistive technology. (2) Characterizing the challenges encountered by blind people in using object recognition technologies, particularly in error identification and user's confidence. (3) Suggesting practical implications for the design of object recognition systems, with a focus on error-handling mechanisms, based on empirical findings.

## 2 Related Work

Object recognition, encompassing both object detection and classification [12, 73], has been the subject of active research for decades, representing fundamental and inherently challenging problems within computer vision. Object detection specifically seeks to ascertain the precise location and dimensions of objects within an image, often represented through bounding boxes [91, 92]. On the other hand, image classification aims to determine whether certain objects, belonging to predefined classes, are present within an image or not [48, 57]. Both object detection and image classification find application in a diverse array of fields, including accessibility. Just within the context of technologies for blind and low vision people, the focus of this paper, there are a myriad of publications. In a recent review by Gamage *et al.*, the breakdown highlights the various assistive tasks where this technology is being utilized, covering a wide range of contexts from *handling object and devices, orientation and mobility, communication and information, personal care and protection, cultural and sports activities*, to *personal medical treatment* [26]. Given the inherently error-prone nature of this technology, understanding and designing for user interactions with prediction errors is critical. Below we synthesize prior literature that discuss this in the context of assistive technologies for the blind and more broadly.

**Table 1: Characteristics of related studies on errors in AI-infused assistive technology juxtaposed with ours.**

		[59]	[75]	[18]	[74]	[37]	[10]	Ours
<b>Input/People</b>	Blind & low vision	6, 100	7	15	22, 13	12	20	12
	Sighted		235			12		
<b>Input/Methods</b>	Photo	•	•				•	•
	Speech					•		
	Other			•	•	•		
<b>Methods</b>	Interview	•			•	•	•	•
	Survey		•		•			
	Focus group			•				
	Crowdsourcing		•					
	Lab study	•				•		•
<b>Task</b>	Image captioning	•	•					
	Speech recognition					•		
	Object recognition							•
	Navigation				•			
	Obfuscation						•	
	Controlling a car			•				

### 2.1 Interactions with Errors in AI Technology in the Context of the Blind Community

Previous research has consistently demonstrated the significant impact of errors on the experiences of blind users. Table 1 illustrates previous research examples concerning the ramifications of errors in AI-infused assistive technology. For instance, safety concerns regarding malfunctions in autopilot systems of self-driving vehicles pose a primary apprehension for blind individuals who are encouraged to use such vehicles autonomously [18, 19]. Similar concerns arise in systems where error risks are less critical than those in self-driving vehicles but still consequential. For instance, studies have shown that minor errors in navigation systems can lead to frustration and disorientation, even when the destination is just a few meters away (*e.g.*, [74]). Prior studies also highlighted the need for blind users to distinguish and handle errors when understanding images with AI-based image descriptions [30, 51]. These findings underscore the importance of user-error interaction interfaces that provide contextual information and predictions from machine learning models to help blind users accurately assess error causes and severity.

Similarly, errors significantly impact blind users' experiences with object recognition systems, as blind individuals often rely solely on system outputs due to the challenge of verifying them [64, 75]. Consequently, understanding the implications of errors in AI-infused assistive technology is critical. Research has highlighted instances where such errors have led to adverse outcomes. For example, blind users tend to overtrust automatically generated captions on social media images, even when the captions are incorrect and nonsensical [59]. While some errors in blind navigation systems are manageable in familiar environments, they become problematic when they can lead to embarrassing situations with bystanders [1, 55]. Moreover, errors in image recognition systems used for controlling household objects can pose safety threats. Consequently, robust safety mechanisms are essential for such

tools. Given the significance of error handling in object recognition systems for blind users, this work delves into and delineates the challenges they face in identifying and recovering from object recognition errors with the number of participants, methods, and task contextualized within this literature.

## 2.2 Interactions with Errors in AI-infused Technology in a Broader Context

While errors are easy to tell in some applications where users can understand the outcome from the system and ground truth easily (e.g., navigating familiar routes with a way-finding system), the outcome from the system may not be clearly perceived due to the characteristics of the task, a poorly designed interface, the complexity of the information, or poor concentration caused by a high workload [45, 46]. For example, the ground truth may not be available immediately when the outcome is provided by the system (e.g., medical diagnosis, weather prediction). The ground truth may not be straightforward to the user if the system handles data in an unfamiliar work domain [76]. Therefore, many researchers have worked on developing user interfaces for AI-infused systems aimed at effectively managing errors and aiding users in navigating discrepancies between system outputs and desired outcomes. Noteworthy efforts include strategies to temper user expectations regarding AI system performance [44, 58], alongside the presentation of user-friendly interfaces tailored to address errors arising from diverse AI-infused applications.

Gesture recognition technology has found utility in controlling an array of devices, from visual displays [50] and robots [63] to wearable devices [66, 83], and even in virtual reality interactions [22, 33, 38]. Despite considerable advancements in gesture recognition accuracy and usability, input recognition errors persist, significantly detracting from user experiences [49]. Research endeavors have thus delved into comprehending the ramifications of these errors, uncovering that user tolerance is frequently shaped more by the context of interaction than solely by system performance. Remarkably, users may tolerate recognition error rates of up to 40% before opting for alternative interaction modes over gestures [43]. Moreover, endeavors to alleviate the detrimental impacts of gesture recognition errors have encompassed various strategies, such as real-time error detection and adaptive model adjustments based on discerning whether the erroneous inputs stem from user mistakes or recognition errors [77].

Similarly, in the context of speech recognition systems, while significant progress has been made in minimizing errors under controlled environments, practical challenges such as speaker variability and ambient noise persist [29, 40, 65]. These errors manifest in various forms, including failure to detect speech, misrecognition, or incorrect handling of recognized speech [37, 68]. Studies have revealed that users overlook more than half of speech recognition errors in the absence of visual cues [37]. To address this, researchers have explored techniques for automated error detection in speech recognition outputs, ranging from visually highlighting potentially erroneous words to employing neural network-based predictive models [16, 25, 27, 28, 82]. However, despite advancements, predicting speech recognition errors remains an ongoing research area, with current methods achieving moderate precision and recall rates.

Beyond gesture and speech, researchers are also endeavoring to mitigate errors in other AI-infused applications, including robotics [56] and autonomous vehicles [85]. In these domains, where safety and reliability are paramount, error-handling mechanisms play a critical role in ensuring smooth operation and user trust [7, 71]. Strategies such as fault tolerance, redundancy, and fail-safe mechanisms are being explored to minimize the impact of errors and safeguard against catastrophic failures [13, 60, 86]. Moreover, advancements in simulation and testing methodologies enable researchers to systematically evaluate robustness and user experience with errors in real-world deployment scenarios [2, 3, 69]. Overall, the quest for error-resilient AI-infused systems represents a multifaceted and interdisciplinary endeavor, requiring collaboration across domains to achieve the vision of intelligent, trustworthy technology.

## 3 Methods

To gain insight into blind people's challenges and strategies in handling errors in AI-infused applications for object recognition, we carry out a comprehensive two-phase user study. The study first encompasses a semi-structured interview that captures participants' experience with camera-based assistive tools. The interview is then followed by an object recognition task, where participants are asked to identify errors when interacting with a mobile application in their homes. We adopt this two-pronged approach from a prior study by Hong *et al.* [37] looking at challenges and strategies adopted by blind people when reviewing automatic speech recognition errors. Our study was approved by the Institutional Review Board at our *anonymized institution* (IRB number *anonymized*). Participants were compensated at a 15\$/hour rate for a total of \$26.21 on average (\$23 – 29,  $SD = 1.73$ ).

### 3.1 Participants

We recruited 12 blind participants (6 women, 6 men, 0 nonbinary) from campus email lists and local organizations. As shown in Table 2, their age ranged from 32 to 70 ( $M = 54.3$ ,  $SD = 15.2$ ). Three participants reported being totally blind, five having some light perception, and four being legally blind. P1 and P2 reported “*an auditory processing disorder*” and difficulty hearing “*very high sounds*”, respectively. Yet, all participants indicated that they faced no problems in using a screen reader. All mentioned using smartphones several times a day. All participants were right-handed except for one, who was left-handed (P4). When asked to report their levels of familiarity with machine learning, two participants reported being somewhat familiar, eight being slightly familiar, and two being not familiar at all. We used a 4-point scale for this question, where *not familiar at all* indicated that participants have never heard of machine learning, *slightly familiar* that they have heard of it but don't know what it does, *somewhat familiar* that they have a broad understanding of what it is and what it does, and *extremely familiar* that they have extensive knowledge on machine learning. All questions are available in Appendix A.

### 3.2 Procedure

The study is conducted over two days that may be up to 7 days apart. On the first day, participants engage in a semi-structured interview

ID	Age	Gender	Level of vision	Onset	Familiarity with ML*
P1	39	Female	Light perception	Birth	Not familiar at all
P2	67	Male	Legally blind	55	Slightly familiar
P3	62	Female	Totally blind	Birth	Somewhat familiar
P4	32	Male	Legally blind	20	Slightly familiar
P5	66	Male	Light perception	46	Slightly familiar
P6	61	Male	Light perception	41	Somewhat familiar
P7	70	Male	Legally blind	Birth	Slightly familiar
P8	50	Female	Legally blind	45	Slightly familiar
P9	69	Female	Totally blind	55	Not familiar at all
P10	66	Female	Light perception	Birth	Slightly familiar
P11	33	Female	Light perception	Birth	Slightly familiar
P12	36	Male	Totally blind	Birth	Slightly familiar

\*ML: Machine learning

**Table 2: Participants' demographics and background.**

and answer questions related to demographics, and technology experience. On the second day, they complete a recognition task with an object recognition application engineered by our team that aims to serve as a testbed. The app is called URCam. During this session, participants interact with URCam and a set of given object stimuli. They attempt to identify any recognition errors that the app might have made and express their confidence.

**3.2.1 Semi-structured interview.** The interview lasted 51 minutes on average (18 – 90m,  $SD = 21.37$ ). It was completed remotely over Zoom and recorded for later analysis. Beyond demographics, participants responded to questions about:

- frequency of using a mobile device, taking photos, reviewing photos, and changing settings of the camera;
- purpose of taking photos, subjects included, applications and devices used, and confidence on photo composition;
- frequency of use of a camera-based assistive application, its usefulness, and device;
- frequency of verifying the recognition results of a camera-based assistive application, encountering errors, importance of errors, and difficulty of identifying the errors;
- strategy of taking photos with an assistive application, degree of understanding how that application works.

As shown in Appendix A, questions assessing frequency are categorized into two groups. The first includes those answerable with an absolute 7-point scale, adopted from Rosen *et al.* [72] (ranging from 'never' to 'several times a day'). For example, 'How often do you take photos or record a video?' The second group includes those suited to a relative 6-point scale (from 'never' to 'always') [20] *e.g.*, 'How often do you encounter misrecognitions when you use Seeing AI?'

**3.2.2 Error identification task.** Given a set of object stimuli and an iPhone 8 device with an object recognition app, participants are asked to try to identify the objects using the application. When deployed in real-world environments, object recognition errors are typically confounded by blurred images, viewpoints with low discriminative characteristics, cluttered backgrounds, low saliency, and more importantly partially included or out-of-frame objects of interest [14, 23, 54]. Thus, we do not conduct this session in our lab,



**Figure 1: Object stimuli in our study from Kacorri *et al.* [42]: baking soda, caramel coffee, Cheetos, chewy bars, chicken broth, coca-cola, diced tomatoes, diet coke, dill, Fritos, Lacroix apricot, Lacroix mango, Lays, oregano, roast coffee.**

but move the study to the homes of blind participants. As in Lee *et al.* [52], all study materials are delivered at home, and instructions are conveyed via Zoom. Each participant received a laptop where the Zoom call is set up for remote communication. Furthermore, participants are provided with Vuzix Blade smart glasses, featuring an integrated camera and initiated with the Zoom call. The smart-glasses can both enable real-time access to participants' first-person perspectives and allow for recordings of observations for subsequent data analysis. At the beginning of the task, the experimenter presents a list of 15 objects for reference (Figure 1). During each trial, participants randomly select an object, capture its image, and obtain a label from the object recognition app, which is communicated via synthesized speech. Upon hearing "Don't know" from the app, indicating that it failed to recognize any object in the photo, participants proceed to capture additional photos until the app provides a label for an object. Subsequently, participants indicate whether the recognition was accurate and express their confidence level in their judgment of correctness for the recognition. After completing the initial 15 trials with all objects (*Attempt 1*), participants repeat the process with the objects in a randomized order (*Attempt 2*), totaling 30 trials. Participants are encouraged to think aloud throughout the task. Upon task completion, participants provide feedback on the difficulty level and the strategies they employ for identifying errors.


### 3.3 Object Stimuli

For the error identification task, we utilize a fixed set of 15 objects across all participants (Figure 1). We adopt similar stimuli to those previously employed in a study examining the interaction of blind users with a teachable object recognizer by Kacorri *et al.* [42]. We adopt their methodology, which involved the selection of objects to encompass a variety of shapes, sizes, materials, and visual similarities. While some products, such as baking soda, chicken broth, diced tomatoes, and diet coke, featured logos or images on their containers that differed slightly from those used in the prior study due to design updates, the fundamental aspects affecting participants' tactile perception, such as shape, material, and weight, remained consistent across all objects.

### 3.4 URCam: An Object Recognition App

For the error identification task, we build an object recognition app, called URCam, that serves as a testbed; the software used as a



basis for experimentation. URCam is fine-tuned using the images of objects in Figure 1. The base model of the object recognizer is InceptionV3 [81], originally trained on the ImageNet dataset [24]. The dataset for fine-tuning comprises photos captured by nine blind participants in a previous study by Lee *et al.* [53], where they trained a teachable object recognizer. Their dataset includes 225 images for each object, totaling 3375 images. Although other existing datasets (e.g., [14]) provide images collected by blind and low-vision people, they did not include fine-grained labels for the specific objects in our study. Therefore, we opted for the dataset collected with blind participants that included those objects. Fine-tuning involves 500 iterations of gradient descent with a learning rate of 0.01. During the identification task, our study participants interact with URCam on an Apple iPhone 8. As shown in Figure 2, upon pressing the  *Scan* item button, the app transmits the image to a server via HTTP, where the fine-tuned object recognition model generates predictions regarding the image's label, subsequently relaying it back to the participant's device via voice and visual display.

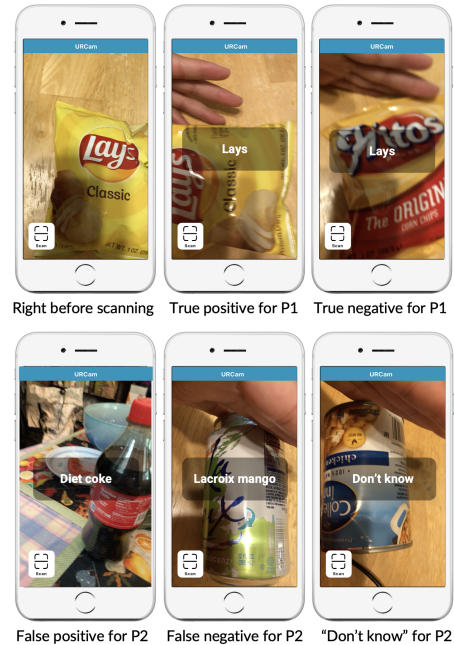
To differentiate between objects within our training set and those the app hasn't encountered previously, we employ a technique that assesses the model's discriminative capacity by measuring the entropy of its confidence scores [88]. Specifically, we establish a threshold for both the entropy value and the confidence score to determine instances where the model should refrain from providing a predicted label and instead output "Don't know". If the entropy value exceeds 2.0 or the confidence score falls below 0.4, the application synthesizes the phrase "Don't know" rather than presenting a predicted label. With this precautionary measure, the model strives to abstain from delivering potentially misleading or inaccurate predictions when it lacks sufficient confidence in its discriminatory abilities.

### 3.5 Data Analysis

The responses from the semi-structured interview and tasks are captured via Zoom. We transcribe these responses to enable a comprehensive analysis of the participants' experience and feedback. We also explore how participants handle application uncertainty (i.e., "Don't know") and deal with potential misrecognitions during the error identification task.

**3.5.1 Semi-Structured Interview.** We use a thematic coding approach to find the major themes in the participants' responses [17]. To reduce the subjectivity, two researchers cooperate to code the responses. One of the researchers transcribes the responses. With the transcribed data, the two researchers code the responses independently and create initial codebooks. They compare the two codebooks and code data to resolve the disagreements through consensus. After resolving the disagreements (a total of 35 out of 373 answers), they establish a shared codebook and code the data. In the final codebook, the responses of 17 open questions in the semi-structured interview include a total of 153 codes.

**3.5.2 Error Identification Task.** We manually annotate the images captured by participants and compare these annotations with the object recognition results recorded by the app to assess the accuracy of object recognition during the task. We categorize the trials based on how well participants identify any object recognition errors by



**Figure 2:** A series of screenshots from URCam that was deployed in the study, where participants P1 and P2 experienced correct, incorrect, and uncertain predictions communicated via a "Don't know" message.

analyzing their responses captured in the video recordings. During this analysis, if participants cannot tell whether the recognition was correct or incorrect, which happened for a total of 7 trials, we interpret this as them perceiving that the recognition can be incorrect but being very uncertain about it. Specifically, we group the trials into:

**True positive:** The object recognition is *correct* and the participant perceive it as *correct*.

**False positive:** The object recognition is *incorrect*, but the participant perceive it as *correct*.

**True negative:** The object recognition is *incorrect* and the participant perceive it as *incorrect*.

**False negative:** The object recognition is *correct*, but the participant perceive it as *incorrect*.

We examine the correlation between participants' confidence levels, and trial completion time, along these 4 groups. Trial completion time is manually measured through video analysis, which involves recording the elapsed time from when the app provided the recognition result to when the participant reports its correctness to the experimenter. Additionally, we investigate any adjustments in participants' strategies for capturing photos when they receive a "Don't know" response from the URCam. We categorize the adjustments by looking at variation in *background*, *viewpoint*, *illumination*, and *object size*; a coding scheme adopted by Hong *et al.* [36].

## 4 Insights from the Interview

The central themes explored during the interview encompass blind people's experiences with photography or video recording and their interaction with camera-based assistive applications. Our discussion delve into various aspects, such as how blind people assess the quality of their photographs, the motivations behind their photography, and the methods they employ to discern inaccuracies within camera-based assistive apps.

### 4.1 Capturing and Reviewing Photos

By delving into participants' experiences with capturing photos or videos, our goal is to uncover the degree of integration of these technologies into the daily routines of blind people. Furthermore, through an exploration of the techniques participants employed to manipulate camera settings, we aim to uncover insights into their approach for capturing photos that would allow them to achieve their goal be it sharing them with others or completing visual tasks. We find that all participants consistently engage in photography activities, each capturing photos at least once a month, as depicted in Figure 3. This aligns with findings from a previous study indicating that BLV people actively use cameras for daily tasks [39]. One of the primary reasons for using a camera was to share images or videos via social media or video calls as shown in prior studies [39, 78]. The majority (8 out of 12) report taking photos or videos more frequently than several times a week. One reason for using a camera was to share photos or engage in video calls. For instance, P4 explained, "Video calls, share photos, I take videos of bands as I play songs. I've got a YouTube channel with several hundred videos of shows I've gone to." Additionally, using assistive technology was cited as another reason, as described by P9: "Sometimes I'll check to see what SeeingAI will say. Just curious to know, what the app will say about. I've used glasses with an app Aira. [...] if you're in like Walgreens, you can connect with Aira and they will tell you what's on the shelf." During their photographic endeavors, participants tend to maintain consistent camera settings and environmental conditions. The majority (8 out of 12) of participants reported never altering their camera settings. Among the participants who did make adjustments (4 out of 12), modifications primarily aimed to optimize lighting conditions. Specifically, three participants sought out locations with ample natural light, while one experimented with flash settings. For instance, P7 sought to evade shadows, stating, "I'll strategically reposition them to ensure optimal lighting without an excess of shadows or other visual distractions." Additionally, one participant (P8) explored varying camera angles to enhance their photographic outcomes. P8 expressed a preference for home photography due to the favorable lighting conditions and exploring camera angles, explaining, "when I'm home, I feel it gives me the maximum amount of light and I get the best pictures. [...] I might move it around a couple of times so that it'll describe it in the most detailed way."

We also posed questions regarding how often they review their photos, as this practice may influence the quality of their images. In general, participants did not frequently review their photos. The majority (8 out of 12) reported checking their photos several times a month or less. **Most participants reviewed their photos independently without sighted help.** Participants who identified

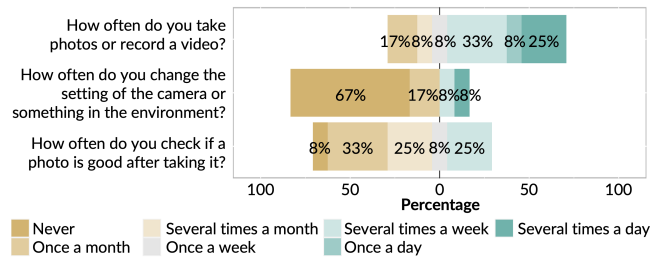


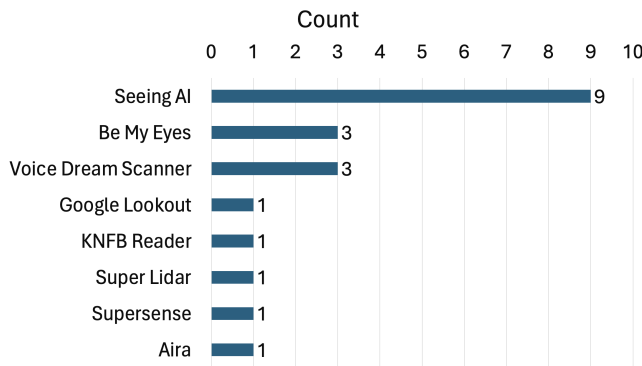
Figure 3: Participants' experience in taking photos.

as legally blind ( $N = 4$ ) predominantly relied on their own visual assessment. They utilized automatically generated image descriptions from assistive tools like Seeing AI and iOS's built-in image captioning function ( $N = 5$ ). For instance, P12 would judge the quality of their photo based on text recognition results, stating, "What's relevant are the OCR results I get from it. Especially if there is a garbled section that doesn't fall into a normal OCR error pattern, then I know the photo's not good." A possible reason for independent reviewing behavior could be concerns about privacy issues when sharing their photos with sighted people [8, 80, 87, 90]. Few (3 out of 12) participants sought assistance from sighted individuals in their vicinity and only one (P1) utilized remote assistance through apps like Aira [6] and BeMyEyes [15].

To provide context for understanding the motivations behind participants' photography, we asked questions about the subjects they captured in their photos. Participants cited documents for text recognition ( $N = 10$ ), people ( $N = 9$ ), objects ( $N = 8$ ), food ( $N = 6$ ), landscapes ( $N = 5$ ), and miscellaneous items such as a scene and a bill ( $N = 4$ ). Similarly, the most prevalent purposes for taking photos or recording videos were for text recognition ( $N = 10$ ), video calls ( $N = 8$ ), and object recognition ( $N = 5$ ). These responses diverge somewhat from the findings of a previous study conducted by Jayant *et al.* [39] in 2011, which suggested that blind individuals primarily took photos to capture friends or family for leisure, while their most sought-after camera function was text recognition. **This result indicates the increasing prevalence of computer vision-based assistive applications among blind and low-vision people.** However, many participants still found image framing challenging ( $N = 9$ ), a difficulty highlighted in prior studies [4, 39, 53]. For instance, P1 and P5 expressed concerns such as, "Making sure the information I'm trying to capture is in the frame of the camera," and "I don't know how far away from the object to hold the phone", respectively. Participants also identified other difficulties such as maintaining focus on the object ( $N = 2$ ), stabilizing the camera ( $N = 2$ ), adjusting lighting conditions ( $N = 2$ ), and orienting objects correctly ( $N = 2$ ).

### 4.2 Handling Image Recognition Errors

To gain insights into the experiences and preferences regarding camera-based assistive applications, we conducted a comprehensive inquiry into the apps they regularly utilize. Participants reported using a total of eight camera-based assistive apps, with inquiries aimed at elucidating their experiences with each. Across 20 participant-app pairs, the predominant choice was Seeing AI,

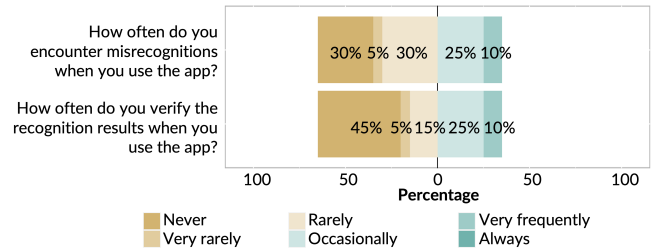


**Figure 4: Camera-based assistive apps the participants have used regularly.**

as depicted in Figure 4. Additionally, participants employed other apps offering text and object recognition functionalities, including Google Lookout, KNFB Reader, Super Lidar, Supersense, and Voice Dream Scanner. Aira and Be My Eyes were also utilized for obtaining remote sighted assistance. Participants varied in the frequency of app usage, with some employing them several times a day ( $N = 5$ ), several times a week ( $N = 7$ ), several times a month ( $N = 5$ ), or once a month ( $N = 3$ ). When asked about the frequency of encountering misrecognitions, responses varied, ranging from very frequently ( $N = 2$ ) and occasionally ( $N = 5$ ) to rarely ( $N = 6$ ), very rarely ( $N = 1$ ), and never ( $N = 6$ ), as shown in Figure 5. However, it's noteworthy that participants might not have perceived all errors. Thus, the reported frequency of errors could be lower than the actual frequency.

We also inquired about participants' strategies for capturing "good" photos when using each camera-based assistive app. To capture quality photos, participants employed strategies such as adjusting the distance and orientation of the camera ( $N = 9$  and  $N = 7$ , respectively) and centering objects in the camera frame ( $N = 7$ ). This reflects the perceived challenge of image framing mentioned earlier. Additionally, participants utilized computer-generated feedback for blind photography ( $N = 8$ ), such as the audio tone system described by P12, a user of Voice Dream Scanner, who stated, "It has this system where the louder and steadier the audio tone is, the better you are. There's a certain tone. You've got the perfect picture and you snap it." P1 highlighted comparable feedback from Seeing AI for taking a photo of a person, stating, "I listen to the prompts. It'll tell me if the face is at the bottom left or top right. Or face is at center. When I hear that. That's when I push the button."

We delved deeper into how participants addressed potential recognition errors they may have experienced with these apps. We queried participants about the frequency with which they validated predictions from the apps (Figure 5). In the majority of cases, participants reported never verifying outputs while using the apps ( $N = 9$ ). Many of them expressed trust in the app's outputs without validation ( $N = 7$ ), exemplified by statements such as "if it says it's a \$5 bill, I believe it" (P2, Seeing AI), "I assume it's correct when it reads it to me" (P6, Seeing AI), and "(I rarely verify the recognition results) because it's pretty accurate." (P12, Google Lookout). This



**Figure 5: Participants' responses about frequency of encountered errors and verification of the output from the apps.**

response aligns with findings from prior studies indicating that blind users tend to trust computer-vision systems [59]. Some participants refrained from validating outputs because they found errors easy to detect ( $N = 6$ ). Particularly with text recognition apps, they could identify errors if the outputs did not make sense. For instance, P11, who never verified outputs from Seeing AI and Voice Dream Scanner, stated "If it tells me a certain thing, I'll know that it actually meant certain numbers. The errors that are sometimes made, they kind of have patterns if you know what it is." This response aligns with the findings of a study by Guerreiro *et al.* [32], which suggests that errors are often acceptable when users understand the imperfections of the technology. When recognizing objects, participants compared app outputs with their expectations based on object textures, shapes, and weights. For instance, P6, who never validated outputs from Seeing AI, mentioned "[...] I could say sometimes it does get the canned soup name wrong, but I guess I don't consider it wrong enough to call it wrong." Some participants verified outputs occasionally ( $N = 5$ ), rarely ( $N = 3$ ), or very rarely ( $N = 1$ ). The most common reason for verifying results was uncertainty with a single output, prompting the need for multiple trials to make a decision ( $N = 8$ ). For example, P3 explained, "if I'm consistently not getting a result with Seeing AI, then I'll see if KNFB Reader will give me results."

**While the frequency of encountering errors varied among the participants, the majority expressed concern about the misrecognitions.** We delved into the impact of errors in camera-based assistive technology on users' experiences with it. In most cases, participants either agreed ( $N = 13$ ) or strongly agreed ( $N = 3$ ) that they cared about the misrecognitions from the apps, as depicted in Figure 6. Sometimes, however, they did not prioritize error correction because they could understand the outputs even with some errors. For instance, errors in text recognition did not significantly alter the meaning of the texts, or the apps were not utilized for sensitive or critical tasks. P8 (Seeing AI) expressed this sentiment, stating, "It's not the most important thing, because I'm not using it for something critical." When asked if there were situations where they cared more about errors, participants often cited text recognition scenarios involving important content such as bills, currency, expiration dates, or other crucial numbers ( $N = 11$ ). For instance, P1, a Be My Eyes user, explained, "if they don't see the expiration date properly on something and it's expired, you know, I could get sick." Other critical situations included reading directions for tasks ( $N = 5$ ) and reviewing important documents ( $N = 5$ ). P9 (Voice Dream Scanner) provided examples of such documents, stating,



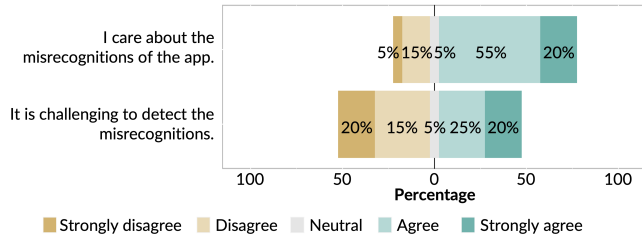


Figure 6: Participants' responses about handling errors.

“probably when it’s something that is connected to legal documents, financial statements, legal financial statements.” Responses regarding the difficulty of identifying misrecognitions varied. **In half of all cases ( $N = 10$ ), participants disagreed or strongly disagreed that identifying errors was challenging when they could easily detect them using contextual cues such as surrounding text or object textures.** For example, P1 (Be My Eyes) remarked, “if they’re wrong, I know they’re wrong. So it’s not really a challenge to identify that it’s a misrecognition for me.” P12 (Seeing AI) similarly commented, “I can catch the errors as they come up because often, it’s not wrong enough for me to not be able to figure out what it says.” Conversely, in other cases, participants ( $N = 9$ ) found errors less distinguishable and challenging to identify. P8 (Seeing AI), acknowledging the possibility of missing errors, expressed, “If it’s wrong, I wouldn’t know. [...] I don’t even know whether it’s wrong or true.” Additionally, P9 recounted instances where sighted individuals detected errors from Voice Dream Scanner that she had missed, stating, “There have been occasions when I didn’t detect anything and a sighted person may have indicated there was something that I just did not get.” When asked how they identified errors, the majority ( $N = 10$ ) of participants relied on contextual cues. For example, P1, using Seeing AI for text recognition, mentioned, “If the information reading isn’t very clear, if I can tell that it’s only reading a part of something then I have to readjust it.” Similarly, P6, identifying objects with Seeing AI, explained, “if I get a soup, and it’s not pronouncing the type of soup, that type of thing.” This behavior contrasts with the strategies of blind users in handling errors in navigation systems, where the majority sought sighted assistance when they encountered errors [32]. In other cases, participants sought clarification from sighted individuals ( $N = 5$ ) or verified app outputs through multiple trials ( $N = 5$ ).

## 5 Error Identification Results

We conducted a comprehensive evaluation of participants’ experience with identifying errors within the context of object recognition. Our analysis centered on discerning patterns in participants’ error-handling behavior throughout the task. Additionally, we examined the influence of repeated object recognition efforts on error handling by comparing the two attempts. Furthermore, participants’ feedback provided valuable insights into their attitudes toward errors encountered in object recognition.

### 5.1 Identifying Object Recognition Errors

Across the 30 trials in the first and 15 in the second attempt, the average accuracy of object recognition stood at 0.76 ( $SD = 0.10$ ).

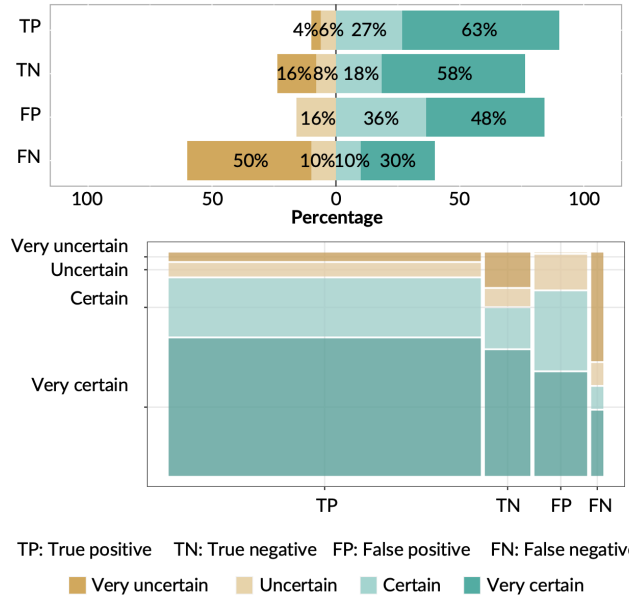


Figure 7: Likert chart (top) and mosaic plot (bottom) of certainty levels and trial categories. The size of the rectangles in the mosaic plot corresponds to the number of trials.

Table 3: Participants’ strategies to overcome “Don’t know”.

Code	Strategy	Cases
<b>Object size</b>	Adjust the camera distance for better framing	11
	Rotate the object to show different sides	119
<b>Background</b>	Move the object to another place	23
	Hide the background objects with a paper	1
<b>Viewpoint</b>	Move the camera to display other sides of the object	29
	Change the way of holding the object	17
	Rotate the object to change perspective	5
	Rotate the camera (portrait and landscape)	1
	No change	10

Participants encountered 7.33 incorrect recognitions on average ( $SD = 2.99$ ), experiencing more false positives ( $M = 3.67, SD = 2.46$ ) than false negatives ( $M = 0.83, SD = 1.03$ ). When looking at whether participants could distinguish between correct and incorrect recognitions, we find that on average, participants successfully identified 21.83 ( $SD = 2.82$ ) correct (true positives) and 3.17 ( $SD = 2.44$ ) incorrect (true negatives) recognition results. However, participants identified errors at a proportion of 0.49 on average ( $SD = 0.32$ ), indicating **that they could detect less than half of the errors**. This outcome is consistent with the over-reliance on image recognition results observed in a previous study by MacLeod *et al.* [59]. The low rate of error identification can be attributed to the challenge of distinguishing objects within the same category that share similar shapes, textures, and weights (e.g., coca-cola and diet coke) when limited visual information is available.

When looking at participants’ strategies for recovery from “Don’t know” predictions, we find that they cluster around varying object



size, background, and viewpoint (shown in Table 3). This is exciting as none of the participants reported having machine learning expertise. Yet, these patterns underscore **participants' awareness of the potential impact of object's size, viewpoint and background on the performance of the object recognition model**, drawing from parallels to how humans recognize objects independent of size, viewpoint, location, and illumination [67]. On average, the object recognition app provided a "Don't know" response in 8.2 trials ( $SD = 4.17$ ) out of 30 trials, totaling 216 cases; a "Don't know" response would often be followed by subsequent a "Don't know" responses with an average of 2.01 ( $SD = 0.89$ ) occurrences. As detailed in Table 3, when participants encountered "Don't know," the most prevalent (116 cases) approach to circumvent it was rotating the object to display its other side, thereby varying the viewpoint in the image, a strategy also prevalent among sighted non-experts in prior work [36]. The second most common approach (29 cases) also involved adjusting the viewpoint, with participants moving the camera instead of the object. Additionally, participants occasionally (23 cases) altered the background of the image by relocating the object to different positions.

We examine participants' certainty around the error identification task by looking at their responses for each trial where they indicate their confidence in their judgment of the model prediction. Overall, participants expressed varying levels of certainty, reporting being "very certain," "certain," "uncertain," and "very uncertain" across 17.67 ( $SD = 5.71$ ), 7.83 ( $SD = 5.77$ ), 2.25 ( $SD = 2.83$ ), and 1.75 ( $SD = 2.01$ ) trials, respectively. As shown in Figure 7, we find that participants reported being either certain or very certain in 90% of true positive trials; trials where the object recognition is correct and the participant perceive it as such. This seems promising. Yet, they also reported being either certain or very certain in 84% of false positive trials; trials where the object recognition is incorrect but the participant perceive it as such. In contrast, participants reported being either certain or very certain in 76% of true negative trials, and 40% of false negative trials. This trend underscores a **tendency for heightened certainty when participants perceived recognition outcomes as correct**. Overall, these findings indicate a prevalent inclination among participants to place trust in the predictions from the object recognizer.

Through analysis of trial completion time, we observe that **participants tend to make quicker decisions regarding the correctness of predictions when they were very certain** ( $M = 3.91s$ ,  $SD = 2.70$ ) compared to when they were just certain ( $M = 8.48s$ ,  $SD = 5.28$ ), uncertain ( $M = 7.87s$ ,  $SD = 2.71$ ), or very uncertain ( $M = 7.69s$ ,  $SD = 8.30$ ). A small correlation was observed between the level of certainty and the trial completion time, as indicated by the Pearson Correlation Coefficient ( $r = 0.27$ ).

## 5.2 Identifying Errors a Second Time

In a real-world scenario, participants tend to interact with a recognition application and similar objects over a long period and often learn to anticipate failures. In Section 5.1, we present aggregated observations from both attempts. To understand even at a small scale the effect of repeated use of the object recognition application on handling incorrect recognitions, in this section we compared the two attempts in the error identification task. Overall, we find

that **the proportion of errors identified by the participants was not significantly<sup>1</sup> different across the two** with it being at 0.51 on average ( $SD = 0.40$ ) for the first and 0.46 ( $SD = 0.36$ ) for the second attempt. Regarding the level of certainty in the correctness of the recognitions, in the second attempt, **participants were certain or very certain for a smaller proportion of trials** across all four categories compared to the first attempt, as shown in Figure 8. One of the reasons for this difference was inconsistent recognition results with the same object across the first and second attempts, supported by P9's response: "the second time around, they gave me different information. So then I became uncertain about trusting what it was telling me."

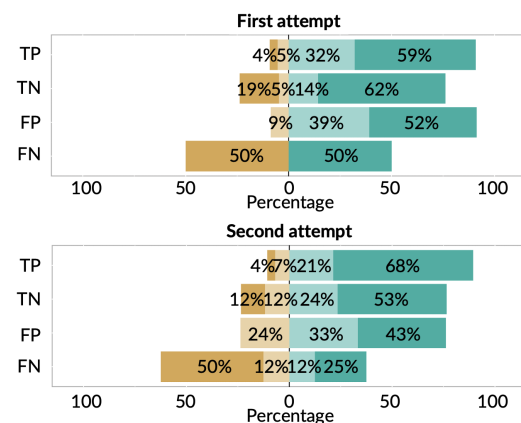
Furthermore, in the second attempt, **the trial completion time significantly<sup>2</sup> decreased** to 4.22 seconds ( $SD = 2.33$ ), compared to the first, where we recorded a longer duration of 6.75 seconds ( $SD = 3.15$ ). This discrepancy suggests a meaningful variation between the attempts. Possible explanations for this observed difference could be attributed to participants' increased familiarity with the task procedure in the second attempt, as well as quicker decision-making based on prior experience with the task in the first attempt.

## 5.3 Subjective Feedback

While participants missed around half of the errors, they generally perceived identifying errors as not challenging, confirming the finding from a prior study that BLV users have mixed feelings with both confidence and concerns regarding identifying errors in private object detection [89]. When asked about the difficulty, the majority disagreed ( $N = 5$ ), with some strongly disagreeing ( $N = 3$ ). For instance, P8, who has low vision, was able to discern correct and incorrect predictions based on their vision and the textures of the object. Other participants identified errors by comparing

<sup>1</sup>We did not observe a statistically significant difference in the results of repeated measures Analysis of Variance (ANOVA) with Aligned Rank Transform (ART) regarding the number of errors and the proportion of errors ( $p > .05$ ).

<sup>2</sup>The results of the repeated measures ANOVA with ART exhibited a statistically significant difference ( $F_{1,9} = 9.67$ ,  $p = .013$ ,  $\eta^2 = 0.52$ ).



TP: True positive TN: True negative FP: False positive FN: False negative

Very uncertain Uncertain Certain Very certain

**Figure 8: Percentages of the certainty levels across the categories of the trials in the first and second attempts.**

predictions across multiple trials. For example, P10 explained, “*I didn’t recognize a mistake until the second similar object appeared. So like the two cans of the Lacroix apricot and Lacroix mango, one of them was incorrect because it was telling me apricot both times.*” Errors were sometimes evident to participants because predicted and true objects had distinct textures, shapes, or weights, as noted by P12: “*[...] for example, the diced tomatoes versus the chicken broth, chicken broth is more liquid. It was easy to identify that it was wrong.*” On the other hand, three participants strongly agreed that identifying errors was challenging. Among them, two mentioned that the recognition results were inconsistent with an object, making it difficult to determine their correctness. P9 said “*Two things that seem similar, but the first time they said they were the same, and then the next time putting them back, they said something different on one of them. So now I’m not sure. So I strongly agree, it was difficult for me to tell us it was in error.*” Another participant mentioned that it was challenging to remember all objects explained at the beginning of the study, which complicated the decision-making regarding the correctness of recognition results.

## 6 Discussion

Our user study, exploratory in nature, shows both promising results and future research directions for supporting blind users’ interactions with error-prone AI-infused technologies. In this section we discuss lessons learned and limitations that may affect the generalizability of our findings.

### 6.1 Implications

**Enable users to leverage their expertise in reviewing errors independently.** The findings from the interview have shed light on an interesting trend: most participants expressed a preference for evaluating the quality of their photographs without the assistance of sighted individuals or remote sighted aid services, such as Be My Eyes or Aira, when using camera-based assistive technologies. This preference seems to stem from a fundamental aspect of the utilization of AI-based systems – namely, the desire to carry out visual tasks independently when sighted assistance is unavailable. It further shows the preference of blind and low-vision users to utilize their expertise in assistive technology, such as integrating recognition results from multiple AI apps to identify errors. This personalized approach to using assistive technology was highlighted in a previous study [34]. This observation underscores a crucial need within the blind community: the ability for individuals to autonomously assess the quality of their photos, taking into account factors such as framing, background clutter, and blurriness.

Addressing this challenge will likely require innovative approaches, particularly in the realm of computer vision. Developing techniques that can accurately quantify the quality factors of photographs without relying on visual cues accessible only to sighted individuals holds great promise in this regard. Such techniques could potentially leverage advanced algorithms and machine learning models to analyze various aspects of a photograph, from composition to sharpness, and provide meaningful feedback to blind and low-vision users. While some initial strides have been made in this area, such as image descriptors for blind users to assess photos for training

personalized object recognition systems [35] and real-time feedback for blind users to capture high-quality photos [4, 61], there remains a need for further investigation. Specifically, it is essential to evaluate the effectiveness of these descriptors in identifying errors and providing actionable insights to users.

**Incorporate the context and recognition system type in designing intuitive user interfaces.** In our interviews, participants delineated diverse approaches to pinpointing errors in both object and text recognition. When utilizing text recognition, they relied heavily on contextual cues, as errors often manifested as deviations from the surrounding text’s logical flow. In contrast, with object recognition, participants leveraged intrinsic object properties such as weight and texture to gauge recognition accuracy. Additionally, certain applications with vision language models like Be My AI [5] furnish detailed image descriptions, enriching user experience. However, this potentially introduces more complex challenges in error detection due to the longer and more descriptive texts [11], compared to the simple object labels in URCam. While our study primarily delved into object recognition, participant feedback underscores the pivotal role of recognition system type and contextual understanding in crafting user interfaces for error detection in camera-based assistive technologies.

Consequently, our findings offer valuable insights for designing intuitive interfaces tailored to object recognition error identification with images. For instance, our findings suggest that providing descriptive information about different facets of an object in an image could mitigate error occurrences, as evidenced by participants frequently resorting to rotating objects to avoid “*Don’t know*” response from the URCam app during the error identification tasks. On the other hand, real-time camera-based assistive technologies may present unique challenges. We expect that blind users would employ different strategies for avoiding and validating errors since they can observe the effects of their camera framing immediately. This immediate feedback loop could encourage adaptive behaviors, such as repositioning the camera or altering the angle of capture to ensure better recognition accuracy. Future research should explore these adaptive strategies in depth, examining how real-time feedback influences user interaction patterns and error mitigation techniques.

**Enable users to understand the performance of the object recognizer.** While participants missed approximately half of the errors, their collective perception of error identification as non-challenging was notable. When probed about the difficulty level, the majority of participants disagreed, with some expressing strong disagreement. This observation underscores the nuanced difficulty inherent in comprehending both the overall performance metrics of the object recognizer (*i.e.*, error rate) and pinpointing individual errors, a challenge compounded for blind and low-vision users. This corroborates findings from prior studies that showed the blind and low-vision users’ tendency to exhibit an overtrust on the output AI-based assistive technologies such as image recognition systems [59] and automatic speech recognition [37].

While studies within the domain of Explainable AI have demonstrated the potential efficacy of elucidating the certainty and rationale behind machine learning model outputs in enhancing performance understanding and usability [79, 84], many of these studies rely on visual information such as heatmap [41] and plots [70]

inaccessible to blind users or have not been assessed with blind individuals. Consequently, to facilitate error identification effectively, forthcoming research endeavors must prioritize the development of methodologies that enable blind and low-vision users to assess the performance of object recognition systems.

## 6.2 Limitations

**Variability between confined study conditions and real-world experience.** A notable limitation of our study lies in the potential disparity between error identification under confined conditions and real-world usage scenarios. In the user study setting, participants were confined to a specific study setup in their home environment with limited variables such as lighting, background, and framing. Typically, they would place the study materials on a table and sit nearby. In a way, they were restricted in their ability to move around freely to find optimal positions for capturing photos, which could influence the quality of the image and subsequently impact error identification. Furthermore, a somewhat ‘staged’ indoor setting may not fully replicate the diverse conditions encountered in real-world scenarios, such as varying lighting conditions, backgrounds, and the presence of outdoor elements. Participants’ level of familiarity and experience with the application may also differ between a one-off and real-world usage contexts. While participants received guidance and instructions during the user study, their experience in using the application in real-world settings may vary, potentially affecting their proficiency in error identification. Last, the number and types of objects encountered in real-world scenarios may differ from the stimuli in the study. Real-world scenarios often involve a wider variety of objects and contexts, presenting unique challenges for error identification.

**Single-session limitation and potential longitudinal variability.** An inherent limitation of our study is that the error identification task was conducted within a single session at participants’ homes. While this approach allowed us to gather valuable data in a naturalistic setting, it may not fully capture the evolution of participants’ error identification abilities over time. Indeed, our observations revealed differences between participants’ performance in the first and second attempts of the error identification task. This discrepancy suggests that participants’ understanding of the object recognizer’s performance, the characteristics of objects, and optimal photo-taking techniques may have improved with repeated exposure and experience. Consequently, a longitudinal study spanning multiple sessions could provide deeper insights into how participants’ error identification abilities evolve over time.

Therefore, while our study provides valuable initial insights into error identification in a single-session context, future research employing longitudinal methodologies could offer a more comprehensive understanding of the development and refinement of error identification experience and expertise that blind users build while interacting with their camera-based assistive technologies.

## 7 Conclusion

We explored the experiences of blind and low-vision people regarding photo-taking, usage of camera-based assistive systems, and error identification within these systems. Through semi-structured interviews, we uncovered that participants predominantly utilize

photo-taking for the purpose of utilizing camera-based assistive systems, rather than solely for capturing memories or sharing with others. Additionally, participants revealed their inclination towards independently reviewing photo quality and identifying errors, despite acknowledging the challenging nature of these tasks for approximately half of the participants. Furthermore, our empirical investigation through error identification tasks provided valuable insights into the challenges associated with identifying object recognition errors. The results indicated that participants successfully identified only around 50% of the errors, predominantly employing viewpoint, background, and object size alterations within images to mitigate errors. Additionally, we observed that the certainty regarding recognition correctness could be adversely affected by inconsistent recognition outcomes in subsequent interactions. These findings significantly contribute to our understanding and quantification of the challenges in identifying object recognition errors within assistive technologies.

## Acknowledgments

We thank Kyungjun Lee, Ebrima Jarjue, and Ernest Essuah Mensah, who were students at the University of Maryland at the time of the data collection and contributed to the remote study protocol. Jonggi Hong initiated this work at the University of Maryland, College Park. This material is based upon work supported by the National Science Foundation under Grant No. 1816380. Hernisa Kacorri was additionally supported by the National Institute on Disability, Independent Living, and Rehabilitation Research (NIDILRR), ACL, HHS under Grant No. 90REGE0008 and 90REGE0024.

## References

- [1] Ali Abdolrahmani, William Easley, Michele Williams, Stacy Branham, and Amy Hurst. 2017. Embracing errors: Examining how context of use impacts blind individuals’ acceptance of navigation aid errors. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 4158–4169.
- [2] Afsoon Afzal, Deborah S Katz, Claire Le Goues, and Christopher S Timmerley. 2020. A study on the challenges of using robotics simulators for testing. *arXiv preprint arXiv:2004.07368* (2020).
- [3] Afsoon Afzal, Deborah S Katz, Claire Le Goues, and Christopher S Timmerley. 2021. Simulation for robotics test automation: Developer perspectives. In *2021 14th IEEE conference on software testing, verification and validation (ICST)*. IEEE, 263–274.
- [4] Dragan Ahmetovic, Daisuke Sato, Uran Oh, Tatsuya Ishihara, Kris Kitani, and Chieko Asakawa. 2020. Recog: Supporting blind people in recognizing personal objects. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.
- [5] Be My AI. 2024. *Introducing: Be My AI*. <https://www.bemyeyes.com/blog/introducing-be-my-ai>
- [6] Aira. 2024. *Your Life, Your Schedule, Right Now*. <https://aira.io>
- [7] Ighoyota Ben Ajenaghughure, Sonia Claudia da Costa Sousa, and David Lamas. 2020. Risk and Trust in artificial intelligence technologies: A case study of Autonomous Vehicles. In *2020 13th International Conference on Human System Interaction (HSI)*. IEEE, 118–123.
- [8] Taslima Akter, Bryan Dosono, Tousif Ahmed, Apu Kapadia, and Bryan Semaan. 2020. "I am uncomfortable sharing what I can't see": Privacy Concerns of the Visually Impaired with Camera Based Assistive Applications. In *29th USENIX Security Symposium (USENIX Security 20)*. 1929–1948.
- [9] Michael A Alcorn, Qi Li, Zhitao Gong, Chengfei Wang, Long Mai, Wei-Shinn Ku, and Anh Nguyen. 2019. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4845–4854.
- [10] Rahaf Alharbi, Robin N Brewer, and Sarita Schoenebeck. 2022. Understanding emerging obfuscation technologies in visual description services for blind and low vision people. *Proceedings of the ACM on Human-Computer Interaction*, 6, CSCW2 (2022), 1–33.

- [11] Akhter Al Amin, Saad Hassan, Matt Huenerfauth, and Cecilia Ovesdotter Alm. 2023. Modeling Word Importance in Conversational Transcripts: Toward improved live captioning for Deaf and hard of hearing viewers. In *Proceedings of the 20th International Web for All Conference*. 79–83.
- [12] Alexander Andreopoulos and John K Tsotsos. 2013. 50 years of object recognition: Directions forward. *Computer vision and image understanding* 117, 8 (2013), 827–891.
- [13] Jyotika Athavale, Andrea Baldovin, Ralf Graefe, Michael Paulitsch, and Rafael Rosales. 2020. AI and reliability trends in safety-critical autonomous systems on ground and air. In *2020 50th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W)*. IEEE, 74–77.
- [14] Reza Akbarian Bafghi and Danna Gurari. 2023. A new dataset based on images taken by blind people for testing the robustness of image classification models trained for imagenet categories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16261–16270.
- [15] BeMyEyes. 2024. *Lend you eyes to the blind*. <http://www.bemyeyes.org/>
- [16] Larwan Berke, Christopher Caulfield, and Matt Huenerfauth. 2017. Deaf and hard-of-hearing perspectives on imperfect automatic speech recognition for captioning one-on-one meetings. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*. 155–164.
- [17] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.
- [18] Robin N Brewer and Vaishnav Kameswaran. 2018. Understanding the power of control in autonomous vehicles for people with vision impairment. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*. 185–197.
- [19] Julian Brinkley, Brianna Posadas, Julia Woodward, and Juan E Gilbert. 2017. Opinions and preferences of blind and low vision consumers regarding self-driving vehicles: Results of focus group discussions. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*. 290–299.
- [20] Sorrel Brown. 2010. Likert scale examples for surveys. *ANR Program evaluation, Iowa State University, USA* (2010).
- [21] Yang Trista Cao, Kyle Seelman, Kyungjun Lee, and Hal Daumé III. 2022. What's Different between Visual Question Answering for Machine? Understanding? Versus for Accessibility? *arXiv preprint arXiv:2210.14966* (2022).
- [22] Taizhou Chen, Lantian Xu, Xianshan Xu, and Kening Zhu. 2021. Gestonhmd: Enabling gesture-based interaction on low-cost vr head-mounted display. *IEEE Transactions on Visualization and Computer Graphics* 27, 5 (2021), 2597–2607.
- [23] Tai-Yin Chiu, Yinan Zhao, and Danna Gurari. 2020. Assessing image quality issues for real-world problems. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3646–3656.
- [24] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255.
- [25] Rahhal Errattahi, Asmaa El Hannani, and Hassan Ouahmane. 2018. Automatic speech recognition errors detection and correction: A review. *Procedia Computer Science* 128 (2018), 32–37.
- [26] Bhanuka Gamage, Thanh-Toan Do, Nicholas Seow Chiang Price, Arthur Lowery, and Kim Marriott. 2023. What do Blind and Low-Vision People Really Want from Assistive Smart Devices? Comparison of the Literature with a Focus Study. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility* (<conf-loc>, <city>New York</city>, <state>NY</state>, <country>USA</country>, <conf-loc>) (ASSETS '23). Association for Computing Machinery, New York, NY, USA, Article 30, 21 pages. <https://doi.org/10.1145/3597638.3608955>
- [27] Sahar Ghannay, Nathalie Camelin, and Yannick Esteve. 2015. Which ASR errors are hard to detect. In *Errors by Humans and Machines in Multimedia, Multimodal and Multilingual Data Processing (ERRARE 2015) Workshop, Sinaia, Romania*. 11–13.
- [28] Sahar Ghannay, Yannick Esteve, and Nathalie Camelin. 2015. Word embeddings combination and neural networks for robustness in asr error detection. In *2015 23rd European Signal Processing Conference (EUSIPCO)*. IEEE, 1671–1675.
- [29] Sharon Goldwater, Dan Jurafsky, and Christopher D Manning. 2010. Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication* 52, 3 (2010), 181–200.
- [30] Ricardo E Gonzalez Penuela, Jazmin Collins, Cynthia Bennett, and Shiri Azenkot. 2024. Investigating Use Cases of AI-Powered Scene Description Applications for Blind and Low Vision People. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–21.
- [31] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).
- [32] João Guerreiro, Eshed Ohn-Bar, Dragan Ahmetovic, Kris Kitani, and Chieko Asakawa. 2018. How context and user behavior affect indoor navigation assistance for blind people. In *Proceedings of the 15th International Web for All Conference*. 1–4.
- [33] Sarthak Gupta, Siddhant Bagga, and Deepak Kumar Sharma. 2020. Hand gesture recognition for human computer interaction and its applications in virtual reality. *Advanced Computational Intelligence Techniques for Virtual Reality in Healthcare* (2020), 85–105.
- [34] Jaylin Herskovitz, Andi Xu, Rahaf Alharbi, and Anhong Guo. 2023. Hacking, switching, combining: understanding and supporting DIY assistive technology design by blind people. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.
- [35] Jonggi Hong, Jaina Gandhi, Ernest Essuah Mensah, Farnaz Zamiri Zeraati, Ebrima Jarjue, Kyungjun Lee, and Hernisa Kacorri. 2022. Blind Users Accessing Their Training Images in Teachable Object Recognizers. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility* (Athens, Greece) (ASSETS '22). Association for Computing Machinery, New York, NY, USA, Article 14, 18 pages. <https://doi.org/10.1145/3517428.3544824>
- [36] Jonggi Hong, Kyungjun Lee, June Xu, and Hernisa Kacorri. 2020. Crowdsourcing the Perception of Machine Teaching. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.
- [37] Jonggi Hong, Christine Vaing, Hernisa Kacorri, and Leah Findlater. 2020. Re-viewing Speech Input with Audio: Differences between Blind and Sighted Users. *ACM Trans. Access. Comput.* 13, 1, Article 2 (April 2020), 28 pages. <https://doi.org/10.1145/3382039>
- [38] Yi-Jheng Huang, Kang-Yi Liu, Suiang-Shyan Lee, and I-Cheng Yeh. 2021. Evaluation of a hybrid of hand gesture and controller inputs in virtual reality. *International Journal of Human-Computer Interaction* 37, 2 (2021), 169–180.
- [39] Chandrika Jayant, Hanjie Ji, Samuel White, and Jeffrey P Bigham. 2011. Supporting blind photography. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*. 203–210.
- [40] Hui Jiang. 2005. Confidence measures for speech recognition: A survey. *Speech communication* 45, 4 (2005), 455–470.
- [41] Weina Jin, Xiaoxiao Li, Mostafa Fatehi, and Ghassan Hamarneh. 2023. Guidelines and evaluation of clinical explainable AI in medical image analysis. *Medical Image Analysis* 84 (2023), 102684.
- [42] Hernisa Kacorri, Kris M. Kitani, Jeffrey P. Bigham, and Chieko Asakawa. 2017. People with Visual Impairment Training Personal Object Recognizers: Feasibility and Challenges. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) (CHI '17). Association for Computing Machinery, New York, NY, USA, 5839–5849. <https://doi.org/10.1145/3025453.3025899>
- [43] Maria Karam and MC Schraefel. 2006. Investigating user tolerance for errors in vision-enabled gesture-based interactions. In *Proceedings of the working conference on Advanced visual interfaces*. 225–232.
- [44] Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 2019. Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-User Expectations of AI Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) (CHI '19). Association for Computing Machinery, New York, NY, USA, 1–14. <https://doi.org/10.1145/3290605.3300641>
- [45] Tom Kontogiannis. 1999. User strategies in recovering from errors in man-machine systems. *Safety Science* 32, 1 (1999), 49–68.
- [46] Tom Kontogiannis and Stathis Malakis. 2009. A proactive approach to human error detection and identification in aviation and air traffic control. *Safety Science* 47, 5 (2009), 693 – 706. <https://doi.org/10.1016/j.ssci.2008.09.007>
- [47] Alexey Kurakin, Ian Goodfellow, Samy Bengio, et al. 2016. Adversarial examples in the physical world.
- [48] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malcolli, Alexander Kolesnikov, et al. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision* 128, 7 (2020), 1956–1981.
- [49] Ben Lafreniere, Tanya R. Jonker, Stephanie Santosa, Mark Parent, Michael Glueck, Tovi Grossman, Hrvoje Benko, and Daniel Wigdor. 2021. False positives vs. false negatives: The effects of recovery time and cognitive costs on input error preference. In *The 34th Annual ACM Symposium on User Interface Software and Technology*. 54–68.
- [50] Chanhwi Lee, Jaehan Kim, Seoungbae Cho, Jinwoong Kim, Jisang Yoo, and Soonchul Kwon. 2020. Development of real-time hand gesture recognition for tabletop holographic display interaction using azure kinect. *Sensors* 20, 16 (2020), 4566.
- [51] Jaewook Lee, Jaylin Herskovitz, Yi-Hao Peng, and Anhong Guo. 2022. ImageExplorer: Multi-layered touch exploration to encourage skepticism towards imperfect AI-generated image captions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–15.
- [52] Kyungjun Lee, Jonggi Hong, Ebrima Jarjue, Ernest Essuah Mensah, and Hernisa Kacorri. 2022. From the lab to people's home: lessons from accessing blind participants' interactions via smart glasses in remote studies. In *Proceedings of the 19th international web for all conference*. 1–11.
- [53] Kyungjun Lee, Jonggi Hong, Simone Pimento, Ebrima Jarjue, and Hernisa Kacorri. 2019. Revisiting blind photography in the context of teachable object recognizers. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*. 83–95.



- [54] Kyungjun Lee and Hernisa Kacorri. 2019. Hands holding clues for object recognition in teachable machines. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.
- [55] Kyungjun Lee, Daisuke Sato, Saki Asakawa, Hernisa Kacorri, and Chieko Asakawa. 2020. Pedestrian detection with wearable cameras for the blind: A two-way perspective. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.
- [56] Dewen Liu, Changfei Li, Jieqiong Zhang, and Weidong Huang. 2023. Robot service failure and recovery: Literature review and future directions. *International Journal of Advanced Robotic Systems* 20, 4 (2023), 17298806231191606.
- [57] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. 2020. Deep learning for generic object detection: A survey. *International journal of computer vision* 128, 2 (2020), 261–318.
- [58] Olga Lukashova-Sanz, Martin Dechant, and Siegfried Wahl. 2023. The Influence of Disclosing the AI Potential Error to the User on the Efficiency of User-AI Collaboration. *Applied Sciences* 13, 6 (2023), 3572.
- [59] Haley MacLeod, Cynthia L Bennett, Meredith Ringel Morris, and Edward Cutrell. 2017. Understanding blind people's experiences with computer-generated captions of social media images. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 5988–5999.
- [60] Carl Macrae. 2022. Learning from the failure of autonomous and intelligent systems: Accidents, safety, and sociotechnical sources of risk. *Risk analysis* 42, 9 (2022), 1999–2025.
- [61] Maniratnam Mandal, Deepti Ghadiyaram, Danna Gurari, and Alan C Bovik. 2023. Helping Visually Impaired People Take Better Quality Pictures. *IEEE Transactions on Image Processing* (2023).
- [62] Daniela Massiceti, Camilla Longden, Agnieszka Slowik, Samuel Wills, Martin Grayson, and Cecily Morrison. 2023. Explaining CLIP's performance disparities on data from blind/low vision users. *arXiv preprint arXiv:2311.17315* (2023).
- [63] M Meghana, Ch Usha Kumari, J Sthuthi Priya, P Mrinal, K Abhinav Venkat Sai, S Prashanth Reddy, K Vikranth, T Santosh Kumar, and Asisa Kumar Panigrahy. 2020. Hand gesture recognition and voice controlled robot. *Materials Today: Proceedings* 33 (2020), 4121–4123.
- [64] Meredith Ringel Morris. 2020. AI and Accessibility. *Commun. ACM* 63, 6 (2020), 35–37.
- [65] Chelsea Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, and Jichen Zhu. 2018. Patterns for how users overcome obstacles in voice user interfaces. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–7.
- [66] Nooruddin Nooruddin, Rahool Demhani, and Nizamuddin Maitlo. 2020. HGR: Hand-gesture-recognition based text input method for AR/VR wearable devices. In *2020 IEEE international conference on systems, man, and cybernetics (SMC)*. IEEE, 744–751.
- [67] Thomas J Palmeri and Isabel Gauthier. 2004. Visual object understanding. *Nature Reviews Neuroscience* 5, 4 (2004), 291. <https://doi.org/10.1038/nrn1364>
- [68] Cathy Pearl. 2016. *Designing voice user interfaces: Principles of conversational experiences*. O'Reilly Media, Inc.
- [69] Jaume R Perello-March, Christopher G Burns, Roger Woodman, Mark T Elliott, and Stewart A Birrell. 2021. Driver state monitoring: Manipulating reliability expectations in simulated automated driving scenarios. *IEEE transactions on intelligent transportation systems* 23, 6 (2021), 5187–5197.
- [70] Biswajeet Pradhan, Abhirup Dikshit, Saro Lee, and Hyesu Kim. 2023. An explainable AI (XAI) model for landslide susceptibility modeling. *Applied Soft Computing* 142 (2023), 110324.
- [71] Kaspar Raats, Vaike Fors, and Sarah Pink. 2020. Trusting autonomous vehicles: An interdisciplinary approach. *Transportation Research Interdisciplinary Perspectives* 7 (2020), 100201.
- [72] Larry D Rosen, Kelly Whaling, L Mark Carrier, Nancy A Cheever, and Jeffrey Rokkum. 2013. The media and technology usage and attitudes scale: An empirical investigation. *Computers in human behavior* 29, 6 (2013), 2501–2511.
- [73] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115, 3 (2015), 211–252.
- [74] 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 *The 21st international acm sigaccess conference on computers and accessibility*. 222–235.
- [75] Elliot Salisbury, Ece Kamar, and Meredith Morris. 2017. Toward scalable social alt text: Conversational crowdsourcing as a tool for refining vision-to-language technology for the blind. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 5.
- [76] Abigail J Sellen. 1994. Detection of everyday errors. *Applied Psychology* 43, 4 (1994), 475–498.
- [77] Naveen Senthilnathan, Ting Zhang, Ben Lafreniere, Tovi Grossman, and Tanya R Jonker. 2022. Detecting input recognition errors and user errors using gaze dynamics in virtual reality. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–19.
- [78] Woosuk Seo and Hyunggu Jung. 2021. Understanding the community of blind or visually impaired vloggers on YouTube. *Universal Access in the Information Society* 20 (2021), 31–44.
- [79] Donghee Shin. 2021. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies* 146 (2021), 102551.
- [80] Abigale Stangl, Emma Sadjo, Pardis Emami-Naeini, Yang Wang, Danna Gurari, and Leah Findlater. 2023. "Dump it, Destroy it, Send it to Data Heaven": Blind People's Expectations for Visual Privacy in Visual Assistance Technologies. In *Proceedings of the 20th International Web for All Conference*. 134–147.
- [81] C. Szegegy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2818–2826. <https://doi.org/10.1109/CVPR.2016.308>
- [82] Yik-Cheung Tam, Yun Lei, Jing Zheng, and Wen Wang. 2014. ASR error detection using recurrent neural network language model and complementary ASR. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2312–2316.
- [83] Puchuan Tan, Xi Han, Yang Zou, Xuecheng Qu, Jiangtao Xue, Tong Li, Yiqian Wang, Ruizeng Luo, Xi Cui, Yuan Xi, et al. 2022. Self-powered gesture recognition wristband enabled by machine learning for full keyboard and multicommand input. *Advanced Materials* 34, 21 (2022), 2200793.
- [84] Eric S Vorm. 2018. Assessing demand for transparency in intelligent systems using machine learning. In *2018 Innovations in Intelligent Systems and Applications (INISTA)*. IEEE, 1–7.
- [85] Junhong Wang, Yun Li, Zhaoyu Zhou, Chengshun Wang, Yijie Hou, Li Zhang, Xiangyang Xue, Michael Kamp, Xiaolong Luke Zhang, and Siming Chen. 2022. When, where and how does it fail? a spatial-temporal visual analytics approach for interpretable object detection in autonomous driving. *IEEE Transactions on Visualization and Computer Graphics* 29, 12 (2022), 5033–5049.
- [86] Chenyun Wu, Rabia Sehab, Ahmad Akrad, and Cristina Morel. 2022. Fault diagnosis methods and Fault tolerant control strategies for the electric vehicle powertrains. , 4840 pages.
- [87] Jingyi Xie, Rui Yu, He Zhang, Sooyeon Lee, Syed Masum Billah, and John M Carroll. 2024. BubbleCam: Engaging Privacy in Remote Sighted Assistance. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–16.
- [88] Guangxiao Zhang, Zhuolin Jiang, and Larry S Davis. 2012. Online semi-supervised discriminative dictionary learning for sparse representation. In *Asian conference on computer vision*. Springer, 259–273.
- [89] Lotus Zhang, Abigale Stangl, Tanusree Sharma, Yu-Yun Tseng, Inan Xu, Danna Gurari, Yang Wang, and Leah Findlater. 2024. Designing Accessible Obfuscation Support for Blind Individuals' Visual Privacy Management. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–19.
- [90] Zhuohao Jerry Zhang, Smirity Kaushik, JooYoung Seo, Haolin Yuan, Sauvik Das, Leah Findlater, Danna Gurari, Abigale Stangl, and Yang Wang. 2023. {ImageAlly}: A {Human-AI} Hybrid Approach to Support Blind People in Detecting and Redacting Private Image Content. In *Nineteenth Symposium on Usable Privacy and Security (SOUPS 2023)*. 417–436.
- [91] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. 2019. Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems* 30, 11 (2019), 3212–3232.
- [92] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. 2023. Object detection in 20 years: A survey. *Proc. IEEE* 111, 3 (2023), 257–276.

## A Interview Questions

In this section, you'll find the questions posed to the participants during the user study. If a question has multiple-choice options, they're listed in square brackets after the question.

### A.1 Demographic Information

- What is your age?
- What is your gender or gender identity? [woman, man, non-binary]
- What is your occupation?
- What is your dominant hand? [left, right]
- What phone do you use? Do you use the screen reader (e.g., VoiceOver)?

#### A.1.1 Visual Impairments.

- Do you have visual impairments? [yes, no]
- Describe your current level of vision.

- For how many years have you had this level of vision ability?

#### A.1.2 Hearing Impairments.

- Do you have hearing impairments? [yes, no]
- Describe your current level of hearing ability.
- For how many years have you had this level of hearing ability?

#### A.1.3 Motor Impairments.

- Do you have motor impairments? [yes, no]
- Describe your current level of motor ability.
- For how many years have you had this level of motor ability?

### A.2 Technology Experience

- How often do you use a mobile device? [never, once a month, several times a month, once a week, several times a week, once a day, several times a day]
- How would you classify your level of familiarity with machine learning? [
  - not familiar at all (have never heard of machine learning)
  - slightly familiar (have heard of it but don't know what it does)
  - somewhat familiar (I have a broad understanding of what it is and what it does)
  - extremely familiar (I have extensive knowledge of machine learning)
]

### A.3 Photo-taking Experience

- How often do you take photos or record a video? [never, once a month, several times a month, once a week, several times a week, once a day, several times a day]
- How often do you change the setting of the camera or something in the environment? For example, sitting at the same table, light condition, or using flash. [never, once a month, several times a month, once a week, several times a week, once a day, several times a day]
  - (if not "never") Please describe for what tasks and why.
- How often do you check if a photo is good after taking it? [never, once a month, several times a month, once a week, several times a week, once a day, several times a day]
  - Do you have a strategy for checking a photo?
- Which of the following do you capture with a camera? (select all that apply) [document, people, landscapes, food, objects, others]
  - (for each, ) How often do you capture it with a camera?
- On what devices do you interact with a camera like a smartphone, computer, smart glasses, or other devices?
- With what applications or tasks do you use a camera? For example, posting on social media, video calls, assistive technologies, etc.
  - Why do (or don't) you use a camera with these applications or tasks?
- When you take a photo, how often do you feel confident that it was good? [never, very rarely, rarely, occasionally, very frequently, always]

- What challenges do you face when taking photos or broadly manipulating a camera?

### A.4 Experience with Image-Based Assistive Tools

(The following questions are asked for each application from the question above "With what applications or tasks do you use a camera?")

- How often do you use the app/tool? [never, once a month, several times a month, once a week, several times a week, once a day, several times a day]
- How often do you use the app/tool when you don't have access to sighted help? [never, once a month, several times a month, once a week, several times a week, once a day, several times a day]
  - Can you provide some examples of when this occurs?
- How often would you notice that the app/tool was wrong after the fact? [never, once a month, several times a month, once a week, several times a week, once a day, several times a day]
- How often do you encounter misrecognitions when you use the app/tool? [never, very rarely, rarely, occasionally, very frequently, always]
- How often do you verify the recognition results when you use the app/tool? [never, very rarely, rarely, occasionally, very frequently, always]
  - Why?
- I find the app/tool to be useful. [strongly disagree, disagree, neither agree nor disagree, agree, strongly agree]
- I care about the misrecognitions of the app/tool?
  - Why?
- Are there some situations in which you care about the misrecognitions more than others?
- It is challenging to detect the misrecognitions. [strongly disagree, disagree, neither agree nor disagree, agree, strongly agree]
- On what devices do you use the app/tool?
- For what tasks do you use the app/tool?
- What mechanisms do you use to detect the misrecognitions if any?
- What kinds of objects do you typically try to recognize with the app/tool?
- What is your strategy for taking good photos when using the app/tool?
- Do you have a sense of how the app/tool works and how it is able to recognize the object?
- How did you learn to use the app/tool when you first installed it?
- Do you like any functions or specific interactions with the app/tool?
- Do you dislike any functions or specific interactions with the app/tool?
- Are you aware of any mobile applications that allow you to personalize them by giving photos of objects or people that you care about?

- (If yes) List them. Which of them have you used before?  
Can you tell me a bit more about your experience?

### **A.5 Post-Task Questions**

We asked the following questions to the participants after the error identification task.

- It was difficult to identify errors made by the object recognizer. [strongly disagree, disagree, neither agree nor disagree, agree, strongly agree]
  - Why?
- How did you know when an object was incorrectly recognized?