PSY 6990

WOZ Study for Visual Search Application for Visually Impaired

By

Nisarg Dave, Erin Richie, Sivaramakrishnan Sriram

PSY 6990: Explanations in AI Systems

Project 2

Overview

Abstract

Visual search applications have the potential to be used for more than just web searches. After understanding how visual search works and the current research, we proposed that it could be used to help the visually impaired navigate. However, we were unsure whether users would trust the system or not. We simulated a visual search system in a Wizard of Oz style study using Google Glass, a set of headphones, and a text to speech application. We then asked participants to close their eyes and walk through an obstacle course while wearing our apparatus. After reviewing the results, we found that most of our 12 participants trusted the system.

Introduction

Visual Search Systems are currently used to sort and search images on the web within applications such as Google. However, they have to potential to identify items in real time and explain what is in a user's path. In the world of AI & machine intelligence, explainability is can become an issue when designing such an apparatus. Our proposed visual search system is combination of several domains. Computer vision technology is growing and now computers can even identify specific humans [1]. Integrating computer vision and AI isn't an easy task. The integrated system becomes challenging when one factors in the explanations of the apparatus to the user. These explanations are important, for if the user doesn't understand the system- they may not trust it.

Hardware:

Ideally, the apparatus would consist of one headset which has a camera, microphone and speaker within it. The headset can see things during the night using infrared dot projectors. The user would wear the headset, turn it on and a camera would start capture the environment. The software will analyze the results and will produce the description or messages to the user. Then the user would be able to respond to the instruction.

Software:

Our primary system is the AI model based on deep convolutional neural network coupled with recurrent neural network for natural language generation. The software takes input from the camera and analyzes it using CNN. CNN finds, what objects are in the image. CNN finds description of the image and then RNN translates those description features in the natural language. Hence, user will come to know about What's going on around there.

System Architecture:

The architecture of system is much complex but yet we can describe it concisely in a very neat and clear manner.

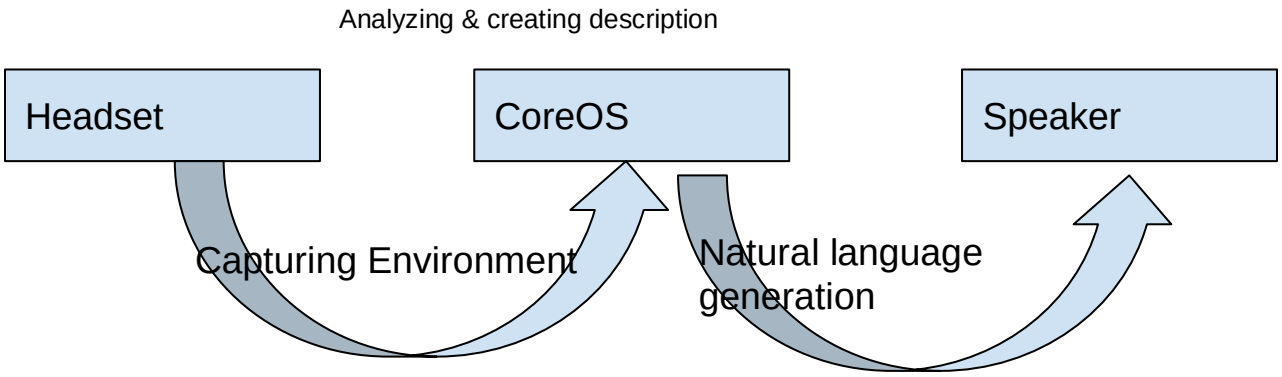

Figure 1: Abstract System Architecture

Inferential Mechanism:

The major hypothesis here is, "Finding not too technical explanation and getting system more understable for others"

The system is much more complex, because of architectural structures of neural network. The system is made using two different neural networks, CNN & RNN. The CNN is here for analyzing the video or image input. RNN is coupled with CNN to get the input from CNN and generating natural language equivalent to those learned knowledge.

As CNN is very complex and it's coupling with RNN makes it even more complex. The lots of nodes in each and every hidden layer. We can't even get insights of what's going on in each and every level. So We can follow other solution for this, we can create the layer on top of neural net to describe what it's going to do?

Making sensible inferences and making it explainable using another layer is feasible idea.

Adding Explainable Interface:

CNN uses many convolution and pooling layers. Each layer nodes are responsible for different tasks. We can create one layer to achieve desired explainability on top of current system.

Mapping the current feature vectors and identifying the major components using the max pooling. The system uses feature mapping criteria for finding possible explanation for working of neural network.

The questions that we can answer using this mechanism are,

a. How system is making prediction?
b. Why system made this prediction? And why not other?
c. Does the user trust each prediction?

The foundation of these explicit explanation lies in causation. Causal reasoning can lead the knowledge mining further and can introduce the new linkages. The causation can control the system's core explanations so explanations should be defined as per causation[2].

Perception and retrospection is the key terms for our system. We need to take care of possible explanations by forward and backward reasoning[3].

Explanation type needed for each and every possible domain based explanations[4]

1. Formal explanation in working of the system
2. Material type explanations for internal component of software.
3. Final type explanations for explaining the goal and objective.

(irrespective of presence & absence of cause)

Learning while explaining in case based reasoning. It adapts to the problem solving in each step[5][6].

Using narration for each rules and methods. For each & every stakeholder it will be easier to perform task. The generation of narrative rules will be done while learning the explanations so it's totally auto encoding the system using relevant inputs.

We are focusing on the human centric justification and explanation so we are considering various hierarchical orders for providing the explanations[9].

Methods

Apparatus:

Users provided with the Google glass and a headset. The users need to wear it and walk with closed eyes in the environment. The environment set-up contains several objects such as chair, table, board & window. The users need to walk, one researcher walks with the user and will play the instruction. We created five sample sound files using "Speak4Me"- a text to speech application. For example, this allowed the researcher to play the "Chair" command when user approached the chair. In this way, the participant thought the apparatus was instructing them, when it was actually the researcher.

Google Glass:

The Google Glass is an Android Wearable device that was developed by Google in 2013. This is an smart glass which helps you to do basic things without taking your phone out of your pocket. The glass consist of three things. The spectrum, the touchpad and the wifi sensor system.

Figure 2: Google Glass

The spectrum serves as the display of the glass. So, this is the one and only output device of the glass which can display all the information that the glass receives. The touchpad allows the user to navigate through all the apps that the glass uses. The wifi sensor helps the glass to pair and sync with your smartphone.

The glass is paired with the phone using the “My Glass” app . All the notifications and the content that get displayed on the spectrum is controlled by this app. The app acts as the Operating system of the glass. The touchpad have functions for navigating through the apps. Just like when you swipe up it shows notifications . When you swipe down it shows the closes any app that is open on the glass. Double tap will wake the glass. Single tap is equal to single touch. And many more options depending on the app that is in.

Figure 3: Example of using the Google Glass touchpad.

Headset:(SkullCandy Method Wireless)

The main disadvantage of the Google Glass is that it doesn't have an audio output. The main purpose of this project is to help visually impaired people. We wanted to develop an AI system that can describe the things that surround it at any point of time. So, we gave that audio output option through the headset. The headset is paired with same smartphone in which the google glass is paired. So, whatever the audio output that comes from the system is heard through the Headset.

Figure 4: Skull Candy Headset

Location:

Library Instructor room, J. Robert Van Pelt and John and Ruanne Opie Library, Michigan Technological University, Houghton MI-49931.

We choose the Library Instruction room as it had variety of objects that can be used for the experiment. The space has many chairs and tables, a board and three sides were covered by wall. One side had a long window . So, it was a perfect location that can help our system to describe many different things as possible.

Figure 5: Location of Testing, Opie Van Pelt Library

The participant was given the Google glass and the headset . They were instructed that "Our project was about developing an AI system that could describe the things that surround us for blind people." "Please participate in our experiment and tell us what do you feel and how much you trust this system? "The participants were asked to close their eyes , wear the glass and the headset. They were instructed that "Whenever an obstacle comes in your way, the system will tell you what the obstacle is and you should stop immediately and take an another path." Just to ensure that they don't get hurt by hitting on these object we cautioned them to stop immediately once they hear about any obstacle through the headset.

Procedure:

The WOZ study experiment was conducted to ensure what people think about our AI system. So, we decided to conduct the experiment with 10 of our friends. We choose the people with a criteria that they will not be having any idea about Intelligence systems, So that they can experience the product as a common user. We will see how an experiment was conducted for a single participant.

The five sound files made matched the objects that we found in that room. These are audio outputs of .wav formats which will describe the objects in a single word like “Chair” ,”Board”, etc. All these outputs were stored in the same smartphone which is paired with the google glass and the headset. The starting point for all the participants were same.

As the participant starts walking around the room, one of us will be following the participant. As soon as the participant approaches any object , we played that object’s audio output from the smartphone. This audio output will be heard by the participant through the headset. So, the participant will come to know that there is an object in front of them was able to adjust their path. Again when they approached some other object, we played that object’s audio output and the same procedure follows on. In this way the participants were able to navigate through the room without hitting any object, simulating the real world scenario of helping a blind person to navigate around obstacles.
The participants were asked to give the feedback on what they felt about the system. A google form which contained specific set of questions and rating criteria was given to them.

Results:

As explained in our methods section, our participants were asked 11 questions upon completing the course and then gave them the option to add additional comments and/or provide their first name for data sorting purposes. Results are listed below in the order in which they were asked on the survey. Please note that for rating purposes we used a semantic scale with the option to choose a rating of 1-5. We typically made a 5 rating the relatively “positive” end of the scale and 1 the relatively “negative” end of the scale. We received the following responses:

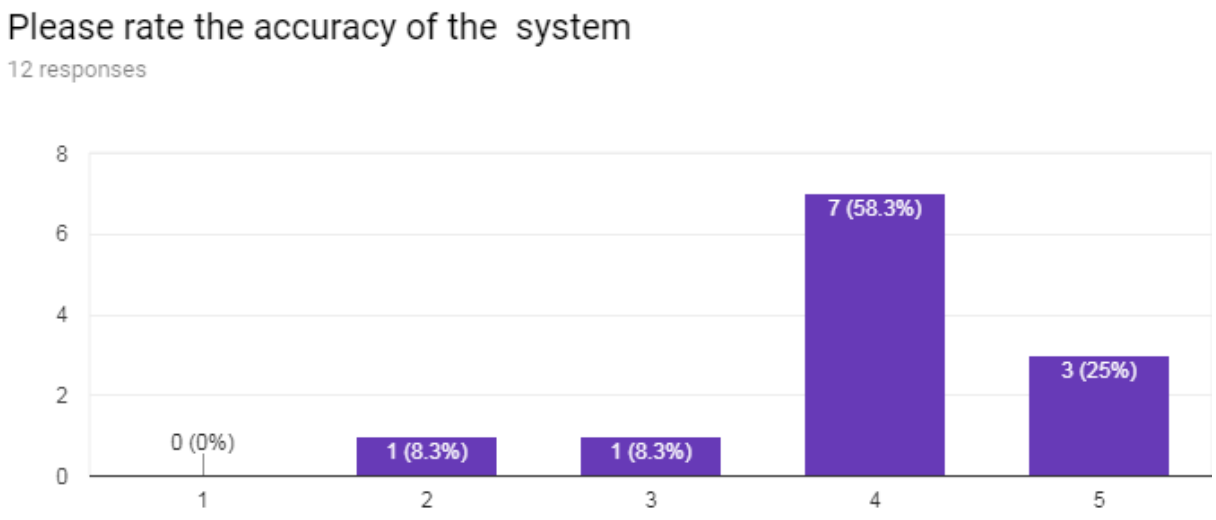

Figure 6: Accuracy Rating Response
Where 1 is not at all accurate and 5 is very accurate

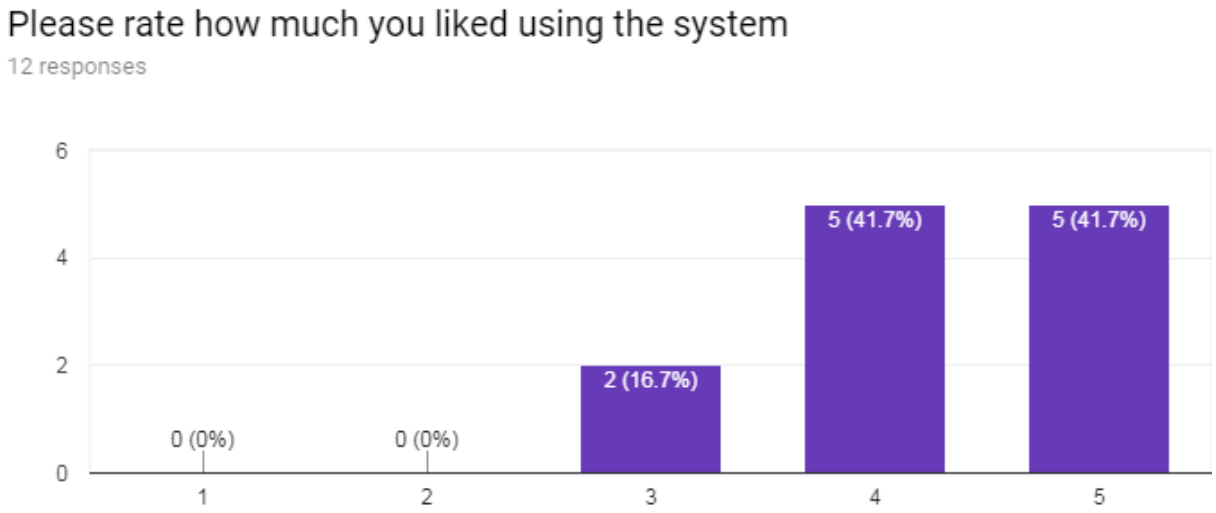

Figure 7: Likability Rating Response:
Where 1 is "I did not like using the system" and 5 is "I really enjoyed using the system"

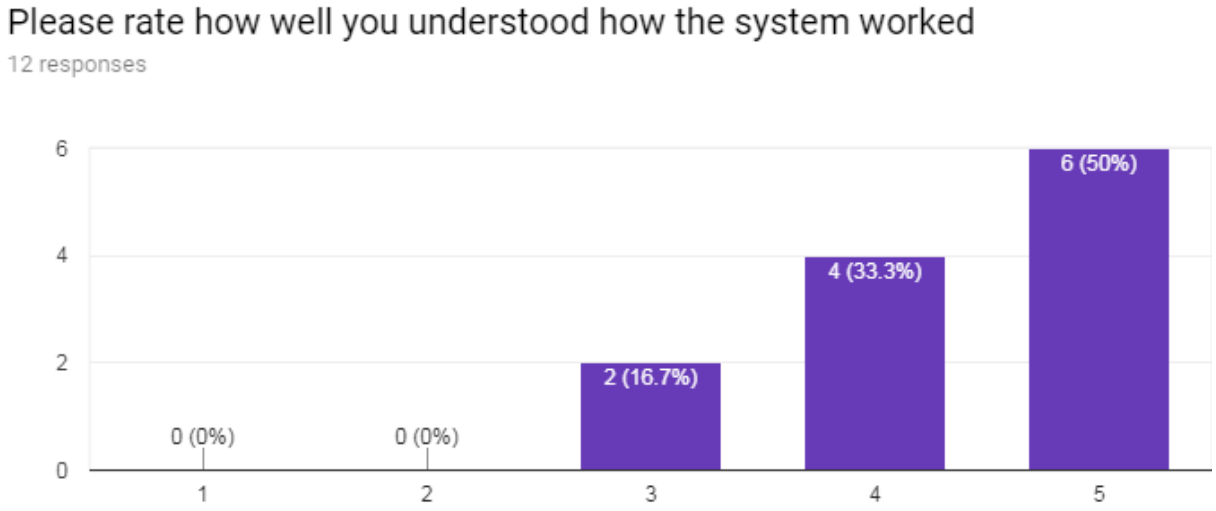

Figure 8: Understanding of System Rating Response:
Where 1 is "I did not understand the system at all" and 5 is "I completely understood the system"

We then asked participants to explain how they thought the system worked and received the following quotes:

- *When we are walking around by closing our eyes, the system tells us whether there is a chair, table, board, door are there in front of you or not.*
- *I hear voice of Object near me.*
- *Was given a warning when there was an obstacle in my path.*
- *very accurate*
- *When I walk it will guide you whether the obstacle is there or not*
- *Yes, it did work well in during most of the situations, however; there were a couple of situations where it showed some time lag.*
- *Sometimes it lags.*
- *It worked when I was near to an obstacle in my path.*
- *I guess it converts the data captured from the Google glass, reads the text and converts it to speech using some neural net.*
- *I think the system worked well. However, if the feedback was a few seconds faster I feel like I would have avoided the obstacles more.*
- *The system was really good at locating the objects and known what the object was, but once I was told about the object I did not know where the best place was to go (if I did not already know the room)*
- *The map of the room was recorded into the glass which used text-to-speech conversion to guide me through the room using the earphones and using the glass as sensor.*

We then asked participants why they thought they received the instructions they did from the apparatus and received the following quotes:

- *it sounds as "Chair" if chair is in front of you. it sounds as table if it is in front of you.*
- *System recognized the object and said its name*
- *To avoid walking into the objects.*
- *for chair and wall in front on me*
- *Yes it gave the instructions but there's a time lag*
- *Because, it wanted to prevent me from the objects that were obstructing my path.*
- *Yes*
- *I guess it gave the instructions as it was programmed for particular object recognition.*
- *Should give navigation directions*
- *To help prevent me from walking in to the obstacles (e.g. walking in to chairs or the wall)*
- *I did not get instructions just a name of an object in front of me*
- *It gave me instructions whenever it felt there was an obstacle ahead of me and it advised me to stop and turn to any other direction.*

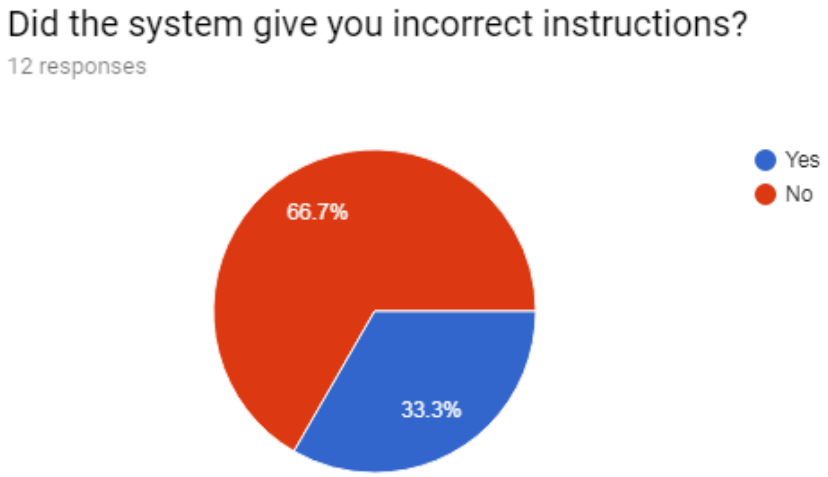

Figure 9: Response to Whether or Not the System Gave Incorrect Instruction

We then asked for explanation as to why those who received incorrect instruction thought they received incorrect instruction and received the following quotes:

- *NA (3)*
- *N/A (3)*
- *If chair and table both are in front of us, then it just sound either 'chair' or 'table'.*
- *Time lag while saying the information*
- *May be due to the distance between the object and the system.*
- *May be due to limited complexity of program or limited database available to software.*
- *I did not notice any incorrect instructions.*
- *How the system knew what object was there, I don't know.*

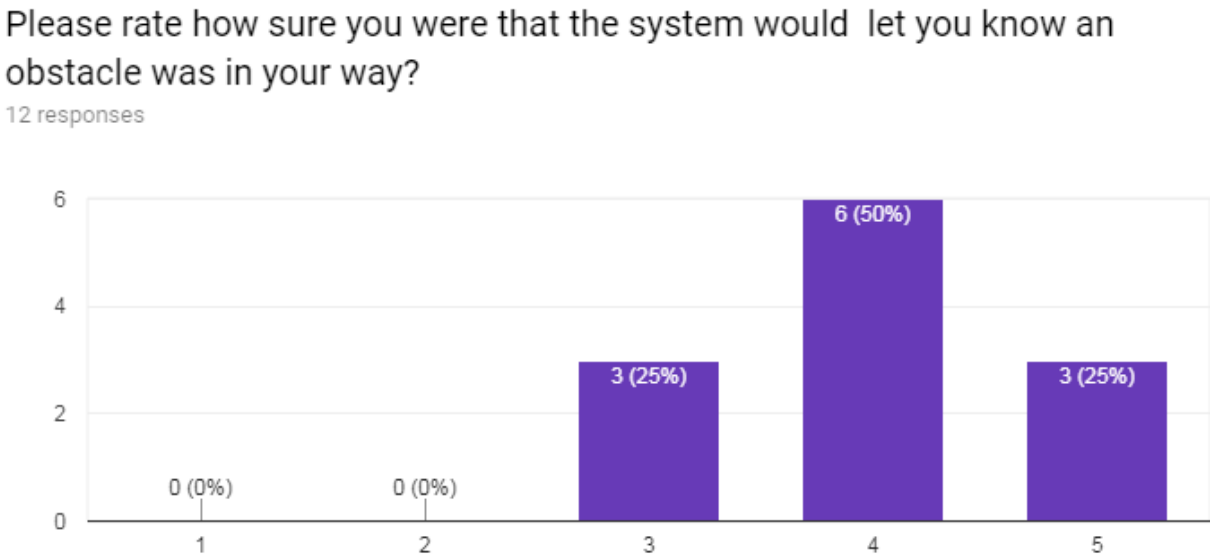

Figure 10: Understanding of System Trust Rating Response:
Where 1 is “Very Unsure” and 5 is “Absolutely Sure”

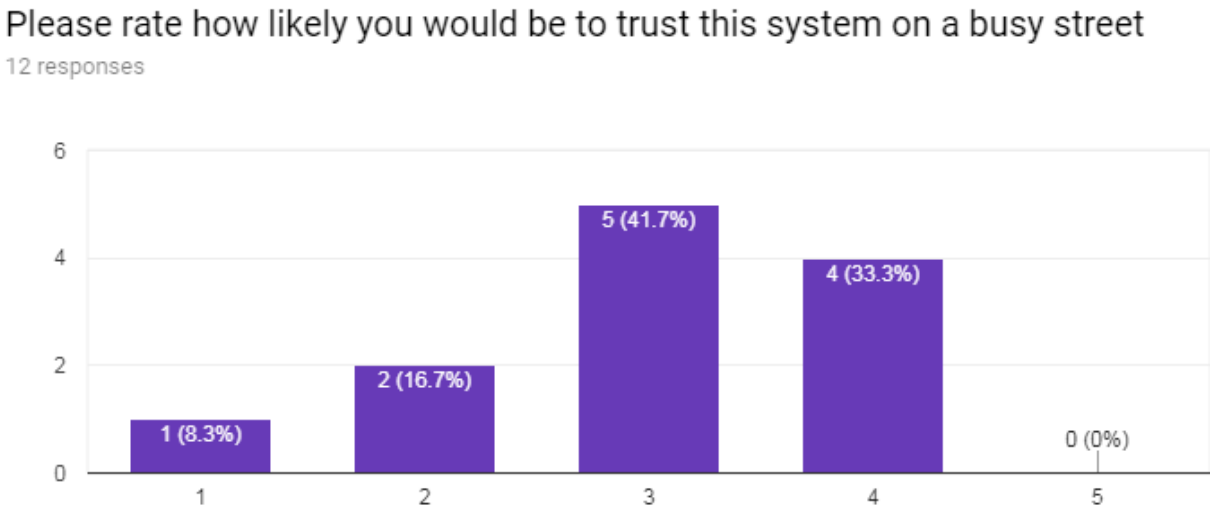

Figure 11: Trust on Busy Street Rating Response:
Where 1 is “Not at all likely” and 5 is “Extremely Likely”

We then asked participants what scenarios they would feel comfortable using this apparatus in and received the following quotes:

- *chair table door board are the common things that the system can easily and accurately predict*

- *Object Recognition*
- *A slow environment, not an environment that is fast.*
- *walking and daily activities*
- *home*
- *It can be trusted fully in the scenarios having comparatively lesser obstructions.*
- *Every*
- *Right now at very less crowded environment*
- *Indoor obstacles like Wall, Chair, Table, etc.*
- *In lower traffic areas with fewer people around.*
- *I would need to be in an environment that I knew well*
- *I would use this in an area where there is not much of a danger to move as I feel it still requires more work.*

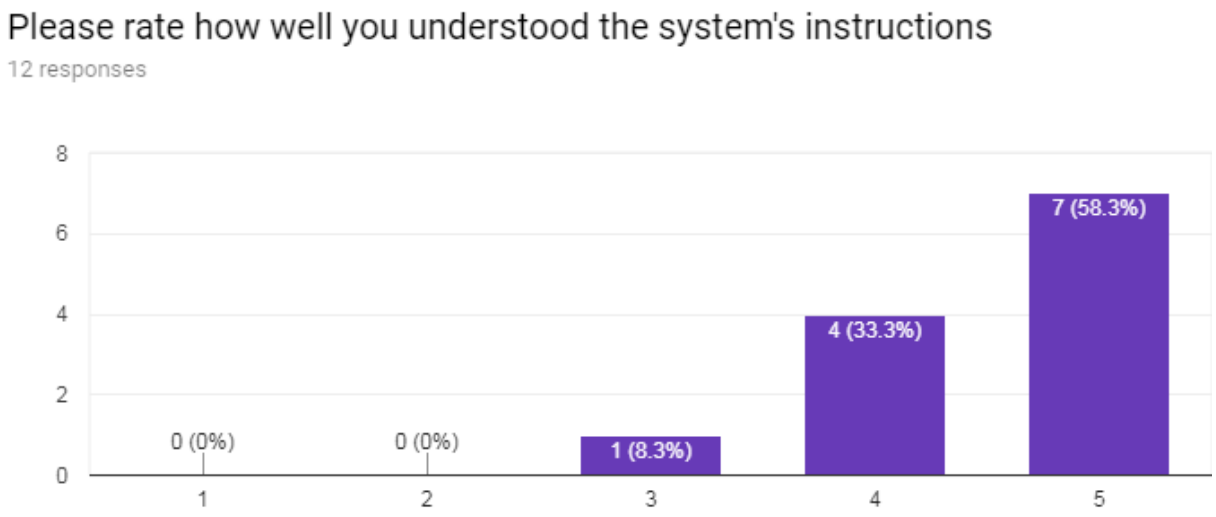

Figure 12: Understanding of System Instructions Rating Response:
Where 1 is “I did not understand the instructions at all” and 5 is “I completely understood the instructions”

Finally, here are the open comment quotes we received:

- *Try to add more features in next version of the system.*
- *System works fine, but needs a lot of training*
- *There was a small lag in the instructions. If I was walking faster it would not have worked.*
- *good project,,just additional comment for new directions will be helpful*
- *There's a time lag while saying the information*

- *I would say that is should assist a person with directions too, because the direction of the object is also important.*
- *It was helpful.*
- *You can program this smart glass more complex so that it will take less computational delay.*
- *It's a great project and the team have made a good initial model.*
- *I think the concept is extremely useful and with clearer instructions and or a couple of milliseconds faster on the feedback I would have felt more comfortable.*
- *The only comment that I would give is that to say what to do when I came across the object. Otherwise it was cool and good a notifying my about an obstacle*
- *I guess it needs to specify more accurate location of the obstacle like if its straight ahead or left or right.*

Discussion:

Overall, most participants liked the system and felt it was accurate. As shown in Figures 6 and 7. It is interesting that most participants rated that they understood how the system worked (see Figure 8), but the explanation they provided was not accurate.This information agrees with some of the discussions we have had in class debating if it matters whether users have an accurate explanation or not as long as they think they have an accurate explanation. Based on the ratings that participants liked the system and thought they understood how it worked even though they did not, an argument can be made that the explanation of a system does not necessarily need to be accurate to be effective.

It is also interesting that there was a difference in the confidence our participants had that the system would let them know an object was in their way and whether or not they would use the system on a busy street. Logically, it seems that if you were certain the system was telling you what was in your path, you would have no problem trusting it would do that in a busier area- but this is not what the data shows. Most comments on where participants would trust using the system included an indoor or slower paced area. Some participants mentioned that the programming would need to be updated before they would trust the system in a higher risk situation. This indicates that while users may have trust for the system after using it just once, they have already formed limitations and understandings as to when and where that level of trust is acceptable.

We included the question on whether or not participants understood the instructions the system gave them as a check for outliers that completely misunderstood the system- since all participants rated a 3 or higher as seen in Figure 12, we can have an understanding that all participants understood what the system was instructing them to do and were able to follow those instructions.

From the overall comments we learned that participants felt there was a lag in instruction. This could be operator error since our researcher was literally pushing a button as they watched the participant draw closer to an obstacle. This could also be a lag due to bluetooth signal. It is possible that there was a significant time difference from the moment we pushed the button to the moment the sound hit the participant's ear. It's possible that the lag accounted for why some would trust the system in lower risk situations rather than higher risk.

Conclusions:

Even though participants liked and felt they understood the system presented, there were still limitations as to when and where they would trust it. Most participants rated that they were confident the system would tell them something was in the way, but would not use the system in a busy street. This lead us to conclude that even after one use of a system, users are already deciding when to trust a system and the limitations they have within that trust. Even though most participants did not have an accurate understanding of how the system worked, they felt they had an accurate understanding- leading us to believe that it doesn't necessarily matter if a user's explanation of a system is correct as long as they think they have an accurate understanding.

References:

1. Rutkin, Aviva Hope. "Retina-Inspired Camera Goes on Sale." *MIT Technology Review*, MIT TechnologyReview,23Aug.2013 ,www.technologyreview.com/s/518586/a-camera-that-sees-like-the-human-eye/.
2. Hoffman, R. R., & Klein, G. (2017). Explaining explanation, part 1: theoretical foundations. *IEEE Intelligent Systems*, *32*(3), 68-73
3. Hoffman, R. R., Mueller, S. T., & Klein, G. (2017). Explaining Explanation, Part 2: Empirical Foundations. *IEEE Intelligent Systems*, *32*(4), 78-86.
4. Lombrozo, T. (2006). The structure and function of explanations. *Trends in cognitive sciences*, *10*(10), 464-470.
5. Doyle, D., Tsymbal, A., & Cunningham, P. (2003). *A review of explanation and explanation in case-based reasoning*. Trinity College Dublin, Department of Computer Science.
6. Sørmo, F., Cassens, J., & Aamodt, A. (2005). Explanation in case-based reasoning–perspectives and goals. *Artificial Intelligence Review*, *24*(2), 109-143.
7. marques brownlee. "Google Glass Explorer Edition: Explained!" *YouTube*, YouTube, 30 Aug. 2013, www.youtube.com/watch?v=elXk87IKgCo&t=188s.
8. "Glass – Glass." *Glass*, www.x.company/glass/.
9. Or Biran, Kathleen Mckeown (2017). Human centric justification for machine learning predictions. International joint conference on Artificial Intelligence.