Object Detection Featuring 3D Audio Localization for Microsoft HoloLens

A Deep Learning based Sensor Substitution Approach for the Blind

Martin Eckert, Matthias Blex and Christoph M. Friedrich

University of Applied Sciences and Arts Dortmund, Department of Computer Science, 44227 Dortmund, Germany

Keywords: Sensor Substitution, Spatial Audio, Object Detection, Convolutional Neural Networks, Mixed Reality.

Abstract: Finding basic objects on a daily basis is a difficult but common task for blind people. This paper demonstrates the implementation of a wearable, deep learning backed, object detection approach in the context of visual impairment or blindness. The prototype aims to substitute the impaired eye of the user and replace it with technical sensors. By scanning its surroundings, the prototype provides a situational overview of objects around the device. Object detection has been implemented using a near real-time, deep learning model named YOLOv2. The model supports the detection of 9000 objects. The prototype can display and read out the name of augmented objects which can be selected by voice commands and used as directional guides for the user, using 3D audio feedback. A distance announcement of a selected object is derived from the HoloLens's spatial model. The wearable solution offers the opportunity to efficiently locate objects to support orientation without extensive training of the user. Preliminary evaluation covered the detection rate of speech recognition and the response times of the server.

1 INTRODUCTION

According to World Health Organization (WHO, 2014), 285 million people worldwide are estimated to be visually impaired; a subset of this estimated population which amounts to 39 million people are diagnosed blind.

This paper aims to use modern technology to help people with visual impairment and blindness. This can improve everyday life quality and help with environmental orientation and interaction. Blind people face many challenges when navigating and interacting with objects and other beings. Particularly in unfamiliar environments, this may prove difficult with unpredictable and uncertain events such as the misplacement of an object. What simply seems like forgetting the location of your coffee cup, is suddenly an ordeal of sorted interactions to find the missing object.

Within this work, a prototype to detect and identify objects in digital images, using deep learning technologies was developed. The prototype, features a HoloLens setup for the user and can be used to identify objects in the users environment and describe them via voice output.

To improve usability of the framework, Microsoft HoloLens provides not only a mobile mixed reality solution that covers image capturing, microphone and 3D audio playback functionalities. Furthermore, it provides a spatial model of the user surroundings which can be used for distance calculation between an object in the spatial model and the HoloLens. This hardware was used within the scope of rapid prototyping, the use of less expensive consumer oriented devices is possible for the final implementation.

The user can issue a voice command to initiate an environment scan that is analyzed by the server running a deep learning network. Following this analysis, the identified objects are articulated in the form of a voice output or “spatial hearing” for the user. Single objects are selectable by voice commands and augmented with additional information on direction and distance.

The 3D audio playback functionality of the HoloLens enables the user to locate a selected object by audio signals, virtually emitted from the object. The possibility of spatial hearing that the HoloLens offers, enables the user to navigate in relation to the object of interest.

The paper is structured as follows, section 2 covers significant research in the field of sensor substitu-

Eckert, M., Blex, M. and Friedrich, C.
Object Detection Featuring 3D Audio Localization for Microsoft HoloLens - A Deep Learning based Sensor Substitution Approach for the Blind.
DOI: 10.5220/0006655605550561
In *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2018) - Volume 5: HEALTHINF*, pages 555-561
ISBN: 978-989-758-281-3

tion with a focus on blind and visually impaired people. Section 3 describes the approach and programmatic design. In section 4 a preliminary evaluation of the device is given. The last section 5 lists pros and cons of the developed prototype and provides ideas for future possibilities.

2 RELATED WORK

Bach-y-Rita conducted a study that shows how quickly blind people can train and develop a visual understanding of the environment around them through the substitution with other senses, e.g. audio or haptics (Bach-y-Rita, 2004). In this way, his statement justifies that the eyes are not used to see, but serve only to absorb information. The brain then assembles the information into an internal image. In light of this research, it is arguable that the source of data is irrelevant and can be replaced.

2.1 Orientation Aids for People with Impaired Vision

In (Meijer, 1992), an experimental system named “The vOICe” translated image representations into time-multiplexed sound patterns. The system worked with a standard video camera and the image was translated into a gray-scale image.

The acquired image is processed column by column from left to right and an audio output is generated for each column. Each pixel is assigned a tone and the pitch is determined by the y-coordinate, the volume by the intensity. When the first column is played, an acoustic signal sounds. This beep tells the user that scanning of the image starts from the left edge. Thus, the user can estimate which horizontal position the individual elements have in the picture and at which point in time the picture is played back. The solution is solid but requires the user to be trained on the system for a longer time. It is offered free of charge as Android, Windows and Web application (Meijer, 2017).

Another system is “SoundView”, it detects objects by means of attached bar-codes and returns the information (Nie et al., 2009). Found objects are announced by speech output. The implementation is facilitated through a camera, an element that provides digital signals and a pair of headphones. The barcodes on the objects must have a fixed size. This allows the calculation of the distance to found objects.

With the development of smaller communication devices, (Sudol et al., 2010) presented a mobile computer-aided visual assistance device. Based on a regular cell phone, the image is sent via network to a computer, the image recognition takes place and is fed back to the user by using a text-to-speech engine. With this technique the user can identify numbers on a dollar bill or CD covers.

”EyeCane”, first introduced by (Maidenbaum et al., 2012), measures the real-world distance between the device and the object it is pointed at. The development was influenced by the previous works of (Meijer, 1992). To measure the distance, the device uses an Infra-red (IR) beam and creates a corresponding audio signal. Maidenbaum goes a step further and improves the idea that with point-distance information in a virtual 3D setting. This development means the cane can also be used to make virtual environments accessible for the blind.

”EyeMusic” on the other hand uses tones from musical instruments to represent colors of objects. It extends the work of (Meijer, 1992) by the use of a musical representation of the image. Five musical instruments are used, each of which is represented by one color. Red for Reggae Organ, Green for Rapmans Reed, Blue for Brass Instruments, White for Choir and Yellow for String Instruments (Abboud et al., 2014). The project heavily collaborated with psychologists to fit the needs of blind and visually impaired people.

“Sonic Eye” is a portable device that sends ultrasound waves through a speaker attached to the head (Sohl-Dickstein et al., 2015). These are reflected from objects in the direction of sound and recorded by a stereo microphone. The normally inaudible ultrasound signal is slowed down and played as a hearable sound to the user.

The App “Be-my-Eyes” helps users send pictures into a distributed network of people who volunteer to classify images and report the information back to the original users (Wiberg, 2015). Instead of Artifical Intelligence (AI), the system uses human intelligence to detect objects. This, however, is not without delay as estimated waiting times for object detection is generally two minutes.

2.2 Convolutional Neural Networks for Object Detection

Computational object detection is used to identify regions of an image that belong to a real-world object. Classic methods for object detection make use of algorithms like Histogram of oriented gradients (HOG) (Dalal and Triggs, 2005) and Scale-invariant feature transform (SIFT) (Lowe, 1999) features. The recent development in deep learning had also influenced the field of computational object detection. Nowadays, Convolutional Neural Networks (CNNs) are used to detect and classify objects within an image or video stream in nearly real-time (Girshick, 2015).

The You Only Look Once (YOLO) approach to object detection was developed by (Redmon et al., 2016) and offers real-time object detection.

Opposed to the traditional procedure where regions are identified and classified, YOLO recognizes the regions and calculates class probabilities in a single step. In this work an improved version of You Only Look Once v2 (YOLOv2) is used (Redmon and Farhadi, 2017).

3 METHOD

The software design is based on a client-server architecture and depicted in figure 1. To calculate predictions for the classification of objects, a consumer grade Graphic Processing Units (GPUs) (e.g. NVIDIA Titan X) is required for this prototype. Thus a client-server architecture is necessary.

Trough it specification the HoloLens proved to be a suitable platform for the proof of concept. The hard and software specification allowed quick implementation and offered necessary debugging options.

On the client side, the HoloToolkit (Microsoft, 2017a) integrates with Unity3D (Unity Technologies, 2017) and Visual Studio Integrated Development Environment (IDE) (Microsoft, 2017b). The HoloToolkit controls cameras, audio input, 3D audio output, as well as server-communication.

Additional to the audio and gesture control the HoloLens provides a ”Clicker”, a small gray device with a button, connected via Bluetooth.

The server runs YOLOv2 (Redmon, 2017b), on top of darknet (Redmon, 2017a) implemented in C and Compute Unified Device Architectur (CUDA) (NVIDIA, 2017).

For the best results in the indoor-setting, a pretrained YOLOv2 model is used. The details of the software stack and further resources are indicated in the table 1.

3.1 Server-side Object Detection

After the integrated Camera takes a regular Red Green Blue (RGB) image, in JPG format with the resolution of 896 × 504, the image is sent to the server for analysis.

Table 1: Software stack showing the software used to create the prototype. Development IDEs and server-side CNNs are included.

Development stack	
Visual Studio	(Microsoft, 2017b)
Unity3D	(Unity Technologies, 2017)
HoloToolkit	(Microsoft, 2017a)
Server stack	
CentOS	(Red Hat, Inc., 2017)
Flask	(Ronacher, 2017)
darknet	(Redmon, 2017a)
Pexpect	(Quast and Thomas, 2017)
YOLOv2	(Redmon, 2017b)
Client	
Windows Holographic Platform	(Microsoft, 2017c)

When the server receives the image, it performs object detection. A pre-trained model of YOLOv2 is used and the results are converted to JavaScript Object Notation (JSON) format (El-Aziz and Kannan, 2014) and returned to the client.

YOLOv2 follows the approach of “only look once” for object recognition applied to each image. As a result, bounding boxes are created and the probability that an object belongs to a certain class inside the bounding box is calculated.

YOLOv2, trained on the 2007 and 2012 PASCAL Visual Object Classes (VOC) datasets, has proven to be the fastest detector based on the VOC dataset (Redmon et al., 2016), demonstrating a solid detection result that is twice as precise as real-time detectors such as 100Hz Deformable Parts Model (DPM) (Girshick et al., 2015) and 30Hz DPM.

The server is returning 2D coordinates to the HoloLens, then 3D real-world coordinates of the object are calculated on the device based on the received 2D data and the virtual 3D model, generated by the HoloLens. The virtual 3D model is mapped by a real-time process, supported by the HoloLens Time-of-Flight (TOF) system and visible light cameras. This is necessary to provide distance announcements for the individual recognized objects.

The estimation of real world coordinates is achieved by using the spatial model of the HoloToolkit. While moving, the TOF and RGB cameras are mapping the surroundings and create a spatial model of the room. Based on this model the distance to the selected object can be calculated. However through the low resolution of the spatial model the distance can only be estimated.

If the user chooses an individual object by voice command, a pink virtual cube is placed in the virtual scene of the HoloLens display to mark the selected object. The virtual scene is generated by scanning and mapping the environment while HoloLens is active.

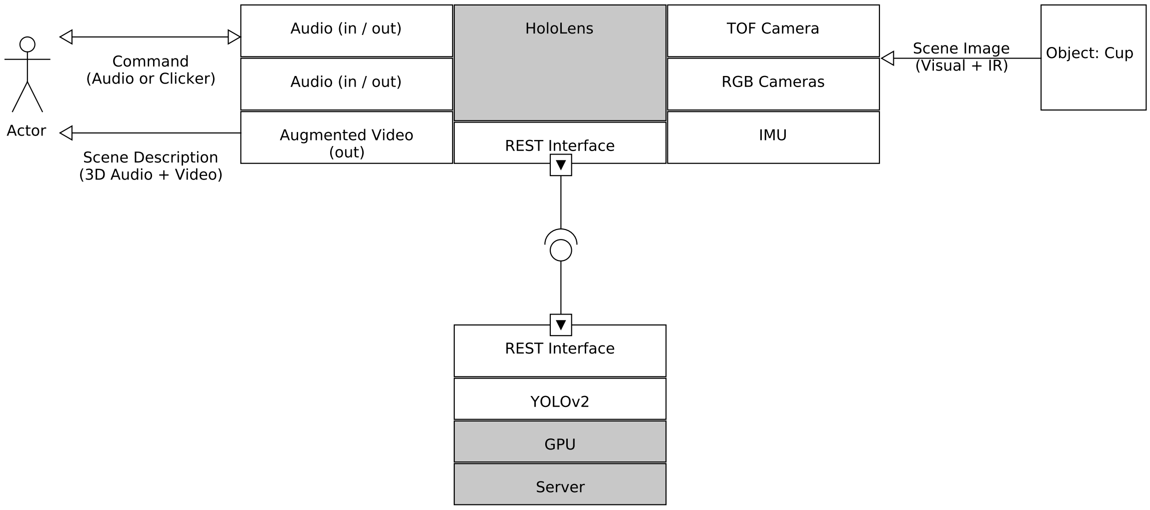

Figure 1: Diagram of the technical architecture. HoloLens gathers Images and Information through sensors. Information is forwarded to the server running object detection with YOLOv2. Information is transferred back to HoloLens. User receives 3D audiovisual feedback on scanned scene and can continue to interact with the system. The Inertial measurement unit (IMU) is used for head tracking.

This scene is then accessed to place objects within the virtual environments, respecting real-live boundaries and create a higher immersion factor while using the device. This enables the software to fix an object in the virtual world even if the user moves the device. Within this work this function is used to calculate distances to objects and highlight selected objects to support people with visual impairments.

Client communication is implemented with a server via Internet connection. This communication has been implemented via Representational State Transfer (REST) interface (Costa et al., 2014) which also ensures that multiple clients can communicate with the server.

3.2 User Interaction

In order to make the operation of the program suitable for everyday use, the possibility of control via voice commands has been integrated. All functions of the program can be executed by voice command.

Firstly, the user must be given an opportunity to take a photo so that it can be processed. This action is triggered by activating the HoloLens via voice command “Scan” or pressing of the “Clicker”, a bluetooth-connected button device, as an alternative robust hardware trigger.

If objects are found in the recorded image, their names are played back as voice output. Table 2 shows the voice commands that can be issued by the user and following actions taken on client- and server-side.

Understanding of audio commands is implemented by using the Speech Recognition Platform, which is part of Microsoft’s Universal Windows Platform (UWP). The audio speech output is realized on the UWP *SpeechSynthesizer* which is bridged to Unity3D. Spatial 3D audio playback is also implemented by using the UWP *SpatialSoundManager*. All three functionalities can be used in Unity3D by using the Microsoft HoloToolkit (Microsoft, 2017a).

At the moment only English is supported, which limits recognition performance for non-native speakers.

The voice output informs the user between individual steps and guide him through the program. For example, if a photo has been taken, the user is informed that there is a short waiting time for object detection. As soon as results are returned, the list of found objects with their names is announced automatically. Figure 2 represents a situation where the user issues a voice command to initiate the object detection.

The locate function, triggered by naming the corresponding object, positions a virtual sound source in the position of the found object. Using the placed sound source, the loudspeakers of HoloLens are used to determine the direction and distance of the selected object.

4 EVALUATION

The following section covers the preliminary evaluation of the prototype. For this position paper, the

Table 2: User, Client and Server interaction. First column contains the voice commands issued by the user. Second and third column describe the actions executed by client and server. *Note that in the fourth line the identifier “Cup” can be replaced by any other Object recognized by YOLOv2 in the scene.

User	Client	Server
“Scan”	• records image • sends it to server • reads list of objects aloud	• receives picture • start object detection • return list of recognized objects
“Objects”	• rereads list of objects aloud	
“Cup”*	• highlights cup with pink cube	
“Distance”	• announces distance to selected object • start 3D audio ping	
“All”	• selects or deselects all found objects	

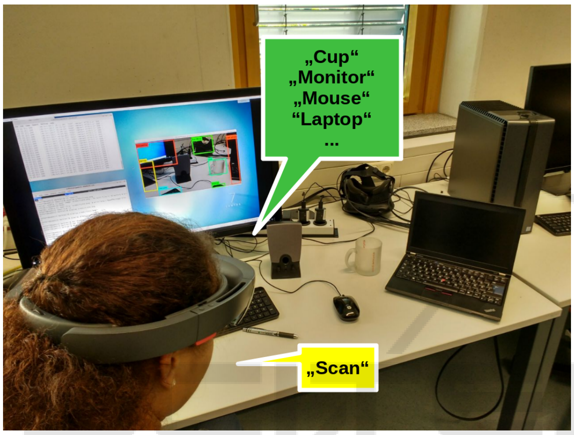

Figure 2: User issuing the “Scan” command using voice or the HoloLens Clicker to detect objects in the field of view. The detected objects are announced by the HoloLens voice output. The Debug interface can be seen in the background.

quality of the speech recognition, the efficiency of the object detection and the response times of the system have been evaluated.

4.1 Speech Recognition

In order to test how well voice commands are recognized and processed, the debug output for each recognized voice command has been used. The voice commands that should be recognized were limited to “Scan”, “Start”, “Stop” and “Chair”. Each word was pronounced 25 times by a non-native speaker. The word “Chair” is only a placeholder and could be replaced by any other object known in the set of object classes. In this test, 83 % of the voice commands were detected correctly and processed immediately. There is a problem with multi-word terms such as “potted plant” or “edible fruit” being pronounced. The multi-word term problem has not been part of this evaluation.

4.2 Object Detection

The object detection of YOLOv2 as tested in the publication (Redmon, 2017b) is 65.5 % correct on the VOC 2007 dataset.

The remaining 34.5 % of errors are split into 19.0 % Localization errors, 6.75 % similar object class detection errors, 4.75 % background detection errors and lastly, 4 % of errors are classified as other.

The YOLO9000 model features 9000 object classes (Redmon and Farhadi, 2017) and enables the object detection to find a larger variety of everyday objects. The detection data for objects has been derived from Microsoft’s Common Object in Context (COCO) and the detection task is based on the ImageNet dataset (Russakovsky et al., 2015).

YOLO9000 gets 19.7 mAP on the ImageNet detection validation set (Redmon and Farhadi, 2017).

4.3 Response Times

In order to measure the processing time of the results on the client, the time between the reception of results and processing was run 100 times on a NVIDIA Titan X GPU. The image is 896 × 504 pixels in size and contains three objects that are found. During the test runs, the position of the HoloLens was not changed. The results are summarized as the median of 100 runs. The accumulated response time was approximately one second. This result can be divided into an average of 627 ms for network transport, 312 ms for object detection and client processing took 101 ms.

5 CONCLUSION

This paper shows, that recent technology advances in deep learning and mixed reality hardware allow faster development of assistive technologies. The proposed system offers many possibilities to simplify the everyday life of those who are visually impaired or blind and can be used without any previous training on the device.

The following section discusses problems and opportunities on how a continuation of this work could evolve. The augmented scene shows inaccuracies, especially on small objects. This leads to inaccurate highlighting of some selected objects due to missing calibration.

Another problem is that the compact hardware of the HoloLens is rather unpleasant and uncomfortable after long periods of wear. Limited battery capacity and a permanently required network connection also limit the mobility of the prototype.

The next version of the HoloLens is supposedly deep learning capable (Microsoft Research Blog, 2017). Considering that the computing power of the new version is strong enough, object detection could be performed directly on the HoloLens.

For this reason the HoloLens was used during the prototyping process, the use of less costly and more durable hardware with well attuned specifications needs to be considered.

In order to provide a quicker overview for people with less severe visual impairment, texts with the names of object classes could be displayed in the field of view.

In addition to object recognition, there is extra information that can be obtained from captured images. Furthermore, the spatial awareness of the HoloLens would make it possible to warn the user if he/she is standing in front of a wall or obstacle at a certain distance. As proven by (Garon et al., 2016), the resolution of depth information can be increased using an external depth sensor.

Other applications already recognize signs or bank notes (Sudol et al., 2010), using various Optical Character Recognition (OCR) frameworks. OCR software could be combined into the prototype to extend these functionalities. The program could also be extended to include recognition of humans, signs or texts in real-time.

AUTHORS CONTRIBUTION

MB implemented the software and co-authored the paper, ME authored the manuscript and tested the implementation. CMF coined the idea, supervised the work and co-authored the manuscript. All authors approved the final version.

REFERENCES

Abboud, S., Hanassy, S., Levy-Tzedek, S., Maidenbaum, S., and Amedi, A. (2014). EyeMusic: Introducing a "visual" colorful experience for the blind using auditory sensory substitution. *Restorative Neurology and Neuroscience*, 32(2):247–257.

Bach-y-Rita, P. (2004). Tactile sensory substitution studies. *Annals-New York Academy Of Sciences*, 1013:83–91.

Costa, B., Pires, P. F., Delicato, F. C., and Merson, P. (2014). Evaluating a representational state transfer (REST) architecture: What is the impact of rest in my architecture? In *Proceedings of the IEEE/IFIP Conference on Software Architecture, (ICSA)*, pages 105–114.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (CVPR)*, pages 886–893.

El-Aziz, A. A. and Kannan, A. (2014). JSON encryption. In *Proceedings of the International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–6.

Garon, M., Boulet, P.-O., Doironz, J.-P., Beaulieu, L., and Lalonde, J.-F. (2016). Real-Time High Resolution 3D Data on the HoloLens. In *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct), 2016*, pages 189–191.

Girshick, R. (2015). Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision,(ICCV)*, pages 1440–1448.

Girshick, R., Iandola, F., Darrell, T., and Malik, J. (2015). Deformable part models are convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 437–446.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Proceedings of the 7th IEEE International Conference on Computer Vision, (ICCV)*, volume 2, pages 1150–1157.

Maidenbaum, S., Arbel, R., Abboud, S., Chebat, D., Levy-Tzedek, S., and Amedi, A. (2012). Virtual 3D shape and orientation discrimination using point distance information. In *Proceedings of the 9th International Conference on Disability, Virtual Reality & Associated Technologies, (ICDVRAT)*, pages 471–474.

Meijer, P. B. (1992). An experimental system for auditory image representations. *IEEE Transactions on Biomedical Engineering*, 39(2):112–121.

Meijer, P. B. (2017). Augmented reality and soundscape-based synthetic vision for the blind. https://www.seeingwithsound.com [Accessed: October 6, 2017].

Microsoft (2017a). HoloToolkit. http://www.webcitation.org/6u0tUB8dz [Accessed: October 5, 2017].

Microsoft (2017b). Microsoft Visual Studio. https://www.visualstudio.com/ [Accessed: October 5, 2017].

Microsoft (2017c). Windows Holographic Platform. https://www.microsoft.com/de-de/hololens [Accessed: October 5, 2017].

Microsoft Research Blog (2017). Second version of HoloLens HPU will incorporate AI coprocessor for implementing DNNs. http://www.webcitation.org/6u0iWSp42 [Accessed: October 5, 2017].

Nie, M., Ren, J., Li, Z., Niu, J., Qiu, Y., Zhu, Y., and Tong, S. (2009). SoundView: an auditory guidance system based on environment understanding for the visually impaired people. In *Proceedings of the IEEE Conference on Engineering in Medicine and Biology Society, (EMBC)*, pages 7240–7243.

NVIDIA (2017). CUDA Zone — NVIDIA Developer. https://developer.nvidia.com/cuda-zone [Accessed: October 5, 2017]".

Quast, J. and Thomas, K. (2017). pexpect. https://pexpect.readthedocs.io/en/stable/# [Accessed: October 5, 2017].

Red Hat, Inc. (2017). CentOS 7. https://www.centos.org/ [Accessed: October 5, 2017].

Redmon, J. (2017a). darknet. https://pjreddie.com/darknet/ [Accessed: October 5, 2017].

Redmon, J. (2017b). YOLOv2. https://pjreddie.com/darknet/yolo/ [Accessed: October 5, 2017].

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (CVPR)*, pages 779–788.

Redmon, J. and Farhadi, A. (2017). YOLO9000: Better, Faster, Stronger. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, USA*.

Ronacher, A. (2017). flask - A microframework based on Werkzeug, Jinja2 and good intentions. https://github.com/pallets/flask [Accessed: October 6, 2017].

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.

Sohl-Dickstein, J., Teng, S., Gaub, B. M., Rodgers, C. C., Li, C., DeWeese, M. R., and Harper, N. S. (2015). A device for human ultrasonic echolocation. *IEEE Transactions on Biomedical Engineering*, 62(6):1526–1534.

Sudol, J., Dialameh, O., Blanchard, C., and Dorcey, T. (2010). Looktel — A comprehensive platform for computer-aided visual assistance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 73–80.

Unity Technologies (2017). Unity3D. https://unity3d.com/ [Accessed: October 5, 2017].

WHO (2014). WHO factsheets: Visual impairment and blindness. http://www.who.int/mediacentre/factsheets/fs282/en/ [Accessed: October 5, 2017].

Wiberg, H. J. (2015). Be My Eyes. http://bemyeyes.com [Accessed: October 5, 2017].