Virtual World Navigation Hat – Development of Sensory Substitution Device through Point Cloud Projection with IoT Sensors for the Visually Impared

Jason D'Souza*
University of San Agustin
jdsouza@usa.edu.ph

Ethel Herna Pabito
University of San Agustin

ChenLin Wang
University of San Agustin

Vince Ginno Daywan
University of San Agustine

Abstract

[TO BE DONE AFTER RESEARCH PAPER IS NEARING COMPLETION]

Table of Contents

1. Introduction

[To be paragraphed or in bulletin points]

Main problem: The visually impaired or the blind navigating the world without sight

Main solution: See through Sound, construct a Sensory Substitution Device to translate visual and environmental information into audio information or any additional senses

Main uniqueness: Virtual world though Point Cloud Projection through a modular operating or interface system, this is to solve previous problems of existing Sensory Substitution Devices

2. Review of Related Literature

[To be paragraphed or in bulletin points]

Closest to the main processes of this project

- **AuralVision** *(https://auralvision.org)* – convert visual depth (though Time-of-flight(ToF) laser sensors) into stereo sound (though bone-conduction)
- **Eyesynth's NIIRA** *(https://realsenseai.com/case-studies/powering-sonic-vision-for-the-blind/ and https://www.eejournal.com/article/21st-century-real-time-navigation-for-the-visually-impaired/)* – translates spatial data (through RealSense D415 giving dept and semantic scene info) into audio (through bone-conduction)
- **Unfolding Space Glove** *(https://pubmed.ncbi.nlm.nih.gov/35271009/)* – converts relative distance (from hand to objects giving depth and position) to vibratory feedback

- **Virtual Whiskers** *(https://arxiv.org/abs/2408.14550)* – finds safe paths or obstacle density (through a camera and AI) into multiple vibration units in the vibration belt
- **Depth Sonification Research** *(https://arxiv.org/abs/2304.05462)* – Collects depth information (through LiDAR or stereo-depth camera) into deep repetition rate (not device-specific but specifies best to encode depth sensor data)
- **SoundSight App** *(https://link.springer.com/article/10.1007/s12193-021-00376-w)* – Mobile app that translates both color and depth (through LiDAR or thermal cameras) into audio has volume for depth and timbre for color or other features

Closest to the design of the project

- **Object Detection and Voice Assisted Navigation: Smart Hat** (https://journal.iba-suk.edu.pk:8089/SIBAJournals/index.php/sjcms/article/view/1535/469) – uses object detection on the camera located on the smart hat which is then provided to the voice assistance to the user
- **Clearway Companion – an AI powered AI for Visually Impaired** *(https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4809356)* – also uses object detection but also takes into account distance estimation, and motion detection (via PIR motion sensors) to provide assistance through haptic notifications and user-friendly audio feedback.

Problems of existing solutions

- **Limited generalizability beyond labs** *(https://cognitiveresearchjournal.springeropen.com/articles/10.1186/s41235-020-00240-7)* – Sensory Substitution Devices (SSDs) often perform well in controlled experiments but fail to translate into practical, real-world use.
- **User comfort, ergonomics, and aesthetics** *(https://journals.sagepub.com/doi/full/10.3233/RNN-160647, https://pmc.ncbi.nlm.nih.gov/articles/PMC5044782/, https://ndpr.nd.edu/reviews/sensory-substitution-and-augmentation/)* – users need hands-free wearable solutions that don't interfere with mobility or self-perception, has SSDs are not yet "hedonically pleasing" and don't feel natural, attractive, or enjoyable.
- **Cognitive overload and attention limits** *(https://pmc.ncbi.nlm.nih.gov/articles/PMC5044782/, https://cognitiveresearchjournal.springeropen.com/articles/10.1186/s41235-020-00240-7)* – Adding another extra sensory input to an already demanding environment is maybe too much information even if useful and can even reduce real-world situational awareness has humans have finite attentional capacity
- **Training and intuitiveness** *(https://dl.acm.org/doi/10.1145/2700648.2811324, https://journals.sagepub.com/doi/full/10.3233/RNN-160647)* – Most SSDs require significant training to use so adoption drops even if long-term proficiency is possible
- **Processing constraints of sensory systems** *(https://pmc.ncbi.nlm.nih.gov/articles/PMC5044782/, https://journals.sagepub.com/doi/full/10.3233/RNN-160647)* – The possible bandwidth and qualities that could be offered by auditory and haptic channels are lower compared to vision thus conveying visual or spatial details through these channels is inherently difficult.

What we had taken from this and our solutions

- **Device is in a form of a Hat** – Since this is a prototype of a sensory substitution device and is bound to be huge or clunky, to prevent it interfering with mobility or minimize its sense of contact we decided to use a hat has a hat can store heaver circuit components without it putting strain on the eyes or ears, also prevent it interfering with mobility or sense of contact has it is not directly touching the user's sensitive areas such as around the eyes and ears while only minimizing mobility to a person wearing a hat. It also prevents ergonomical or aesthetical issues has a engineered heavy eye glasses brings more attention then a Hat.
- **Sensory targeting** – The human mind has finite attention capacity thus if the user uses our SSDs it should controllable by the user either by hand gestures captured by hand gestures detection or the user rotating their heads to see specific areas. This solves the cognitive overloading issues has the user doesn't need to know their direct environment if the user is just lying down, in this case the only information provided are objects moving closer to the user. This also solves the processing constraints has the user doesn't need all the of the visual information, only parts the user needs similar to how the brain only focuses on visual areas the person is focusing on. Thus our device has to implement some features of the brain's automatic visual processing for the users such has change between focusing on the general environment (low focus) or a specific location (high focus, determined by head direction pointer at area of interest), or the eyes more easily can detect changes with respect to area of focus, or the human's logarithmic senses where objects with constant movement are automatically removed over time, and much more.
- **Modularity** – Every one is unique and people have specific priorities, so with a device like this would need to be customized to what specific method of visual transition the user need based on their specific visual imparities like text recognition for people who can see but have difficulty reading text, or completely blind person wanting to go from place to place or message people. This is where we take advance that our device uses a raspberry pi which is just a mini-computer. By allowing support for applications for their specific needs we make our device more easier to update, customize, and improve. Fixing issues like having limited generalizability, or being easier to train on has different components of visual translation (first volume for depth than timbre for texture, etc.) can be introduced in order and if actually needed by the user. This allows more possibilities from using a language model to have a portable general assistant, hand gesture recognition to navigate the virtual operating system, image recognition for more specific interest, or even a browser interface to allow the user to message and use networks like maps or social media.

3. Methodology

The main process of our device is to translate visual and environmental information into audio or haptic information, secondary process is the system of users in a network to allow them to interact with each other and allow sensors (GPS, IMU, Microphone, etc.), actuators (Speakers, etc.), and APIs(social media, maps, etc.) to interact allowing more advance processes like world mapping, weather tracking, and much more.

- **Materials we are to use**

No.	Component	Purpose
1	Raspberry PI 4	Main microcontroller
1	Earphones	To input and output(Binaural) audio information, if possible could be replaced with bone-conductive headset
1	Powerbank	Powers the raspberry PI
2	OV5640 USB camera	The two cameras are used to stereo project a 3D environment from two 2D images
1	GPS Module USB	Provides GPS information to allow tracking and map features
1	IMU	Detects head movements of the user (Inertial Measurement Unit – accelerometer + gyroscope)

- **Framework**
  We will be using C++ OpenCV to execute the main process of Sensory Substitution from Visual and Environmental information to audio information, but for the secondary process of available applications the main system would be running Linux and the main process that initializes the main process and interface to allow applications is DenoJS, this processes connects to the main server giving sensor data and receiving information such as API responses, local language model responses, etc.
- **Program's primary data flow**
  The device captures two images from the horizontally displaced cameras; they are then compared to generate the Depth map through Stereo projection (Cameras are labeled C with Q being the depth)

$$C_1 = \begin{bmatrix} f & 0 & cx_1 & 0 \\ 0 & f & cy & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, C_2 = \begin{bmatrix} f & 0 & cx_2 & T_x f \\ 0 & f & cy & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \therefore Q = \begin{bmatrix} 1 & 0 & 0 & -cx_1 \\ 0 & 1 & 0 & -cy \\ 0 & 0 & 0 & f \\ 0 & 0 & -T_x^{-1} & T_x^{-1}c(x_1 - x_2) \end{bmatrix}$$

The resulting dept map is then converted into a point cloud which transform a single-channel disparity(depth) map into a 3-channel image representing a 3D surface where pixels (x,y) are projected to (X,Y,Z)

$$\begin{bmatrix} X \\ Y \\ Z \\ W \end{bmatrix} = Q \begin{bmatrix} x \\ y \\ disparity(x, y) \\ 1 \end{bmatrix}$$

The data is than scaled and fitted into the virtual world environment with help from the oscilloscope's measurement or through error correction (comparing previous point cloud to the current to determine head movement). This helps normalize the relative distances between captures. What comes after is implementing the human aspect of image recognition to ensure that the final audio will not be overwhelming or unnecessary. Based on the user's situation (walking, resting, standing, looking around) a priority list is assembled including environment (low focus), specific object (high focus), environmental logarithmic changes (staring at one area), or approaching objects (rest mode), basically attempt to minimize has much information as possible and not frequently, only taking into account necessary data by taking inspiration on how the brain normally processes visual information. From which the resulting needed information is collected. Here some post processing steps are included like the Equal-loudness contour to prevent incorrect volume sensing and minimizing high frequency noises, or the Fourier transformation to generate the resulting audio signal live for both speakers.

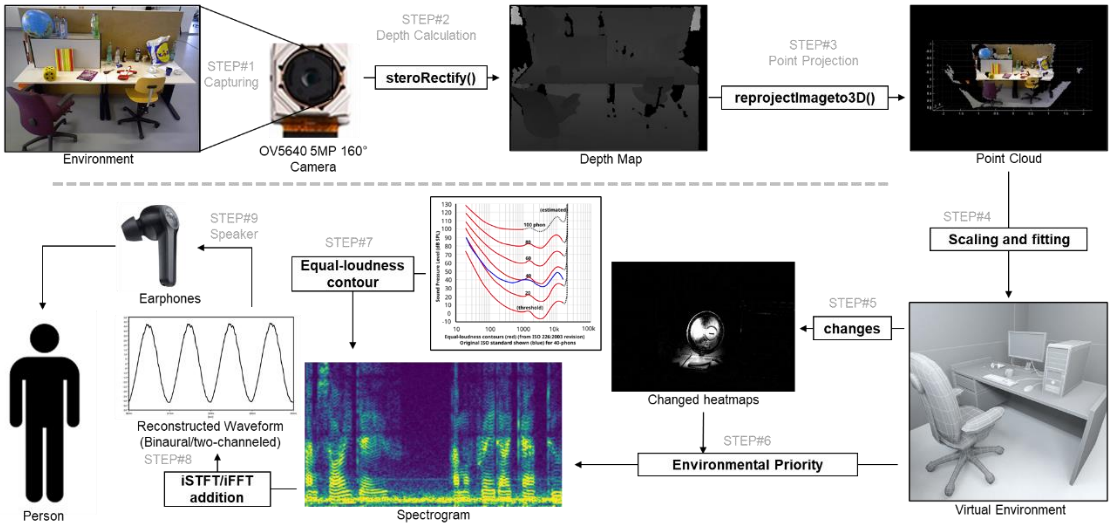

Figure 1: The device's main process's data flow

- **Testing and Experimenting**
  Our method of testing should ensure that we achieve our main three solutions, which mainly relies on qualitative information. Thus, our research will be a qualitative research which aims to achieve our three main solutions:
  - The Hat is partially comfortable and doesn't not greatly interfere with the mobility, ergonomics, and aesthetics of the user. However, since the hat is just on the prototype and can be improved to even possibly a light eyewear, this priority is less important than the next two.
  - The user is not overwhelmed by the auditory information given an understandable training process
  - The device can allow applications on which the user's specific interest works completely with minimal issues
- **Future Plans**
  The device is not only limited to the visually impaired has its initial primary goal of being a sensory substitution device is something to prove itself that it can be used to argument people has technology develops. In the tech field there are several fields like AI, Web3, Argument Reality, IoT, etc. one field we prioritized is Argument Reality because the distance between advance tech and human control grows wider has in the methods of communication towards the AI and the Web are stagnant (Phones, and Laptop) so this devices offers a possibility to argument people has technology becomes more advanced to be limited by devices like phones and laptops (e.g. Having an AI see what you do and become your assistant integrated to your life, or connecting to the internet without holding any devices, etc.). The current design being a hat is of the time being has the prototype would be too heaving to put near the eyes and ears so we keep the hat until better design arises. This also solves the distribution and demand issues, has although this device is initially mainly targeted towards the visually impaired, it wont do much if there isn't much demand for it as not everyone is severely visually impaired meaning that those with severe visual impairments are at a disadvantage because of this. If this device offers something new to the user (Augmented Reality) over modular operating system) then the demand can support it to distribute to most areas including the visually impaired and more people with specific interest, which can grow the

server making the embedded IoT network/server larger building a foundation sustainable from revenue.

4. Results

[Prototype not finished]

5. Discussion/Analysis

[Prototype not finished]

6. Conclusion

[Prototype not finished]

7. References