Vision Substitution with Object Detection and Vibrotactile Stimulus

Ricardo Ribani and Mauricio Marengoni
Universidade Presbiteriana Mackenzie, São Paulo, Brazil

Keywords: Vibrotactile, Vision Substitution, Deep Learning, Object Detection.

Abstract: The present work proposes the creation of a system that implements sensory substitution of vision through a wearable item with vibration motors positioned on the back of the user. In addition to the developed hardware, the proposal consists in the construction of a system that uses deep learning techniques to detect and classify objects in controlled environments. The hardware comprise of a simple HD camera, a pair of Arduinos, 9 cylindrical DC motors and a Raspberry Pi (responsible for the image processing and to translate the signal to the Arduinos). In the first trial of image classification and localization, the ResNet-50 model pre-trained with the ImageNet database was tried. Then we implemented a Single Shot Detector with a MobileNetV2 to perform real-time detection on the Raspberry Pi, sending the detected object class and location as defined patterns to the motors.

1 INTRODUCTION

The World Health Organization estimates 285 million impaired people in the world, being 246 million with low vision and 39 million completely blind (World Health Organization, 2012). These people can increase their environment perception by using technologies that converts visual information to different sensors like hear or touch.

The brain plasticity (Bach-y-Rita and Kercel, 2003) has shown the capability of the brain to adapt to different patterns, no matter its source. With some limitation, the human brain can replace a lost sensor by using information from other existing sensors in the body. Some researchers discuss about this phenomena and the sensory substitution (Novich, 2015; Bach-y-Rita and Kercel, 2003; Visell, 2009) and they cite the capability of the brain to change his organizational structure to recognize different patterns. A common example of this structural change is the easiness of visually impaired people with the braille writing system.

Novich (2015) has presented a method to allow deeply deaf users to recognize words from a limited vocabulary, using encoded audio information captured with a cell phone and sending it to the a wearable vest that activates a series of vibration motors distributed along the user's back.

The present work shows a prototype that uses some portable hardware in conjunction with a simple camera and some vibration motors to send encoded visual information to the users. The information comes from an object detection model and is encoded using a fixed dictionary. The goal of this device is not to completely substitute the vision or other assistance tools (like the white cane). Since the device is wearable, the user can use the vest and the cane in conjunction adding more information about the environment. A limitation of the white cane is the detection of objects that are at the top of the user's field of view, like a tree branch that can be detected by a camera.

Additionally to the developed hardware, the system has an image processing module to perform inference of deep learning models to detect and classify objects in the scene. In our first trial to encode visual information to the vest we performed some tests with classification models that uses Global Average Pooling layers, like the ResNet-50 model (He et al., 2015). The idea behind trying a classification model was to have a lighter model in a portable device like a Raspberry Pi, but the limitations on encoding the information made us to use an object detection model. The chosen model for this task was the Single Shot Detector (Liu et al., 2015) with a MobileNet V2 (Sandler et al., 2018) and the performance of this model on the Raspberry Pi exceeded our expectations.

The paper is organized as follows: in section 2 we present the related work; in section 3, we describe the architecture of our prototype, including the hardware and software; finally in section 4, results and future work are discussed.

Ribani, R. and Marengoni, M.
Vision Substitution with Object Detection and Vibrotactile Stimulus.
DOI: 10.5220/0007577205840590
In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2019)*, pages 584-590
ISBN: 978-989-758-354-4
Copyright © 2019 by SCITEPRESS – Science and Technology Publications, Lda. All rights reserved

2 RELATED WORK

In the field of sensory substitution just a few use image capture and tactile stimulus (Cancar et al., 2013; Dakopoulos and Bourbakis, 2008; Pereira, 2006; Johnson and Higgins, 2006; Bach-y-Rita et al., 1969). The existing trials to encode image to touch consists in simple methods, like downscale the image to the size of a matrix of actuators or downscale the edge image. In some other works that has a focus on navigation (Cancar et al., 2013; Dakopoulos and Bourbakis, 2008; Cardin et al., 2007; Johnson and Higgins, 2006; Meers and Ward, 2005) the signal is a basic function that tells the user the distance from objects in the scene and they normally use depth sensors. In this cases each motor is activated according to a point of depth giving the user a spatial perception of the environment to avoid obstacles. Considering the brain capability of adapt, these methods are very modest ones.

Some other projects also uses image capture but to generate a different type of output, like audio (Sainarayanan et al., 2007; Hub et al., 2004; González-Mora et al., 1999; Meijer, 1992) or touch using gloves (Lin et al., 2012; Meers and Ward, 2005) sending signals to the the fingertips. All these works that uses different sensors to encode image are evidence that doesn't matter the origin of the signal, the brain learns how to deal with the information.

As the goal of this work is to evaluate the brain capability to classify detected objects using touch, below we describe the most relevant projects to our research.

2.1 Tactile Vision Substitution System (TVSS)

Developed by Bach-y-Rita et al. (1969), this work is an early example in the field of tactile sensory substitution. Mounted in a dentist chair, it has a 20x20 vibration motor matrix that project the image captured by an analogical TV camera. No preprocessing is performed on the image.

The TVSS experiments was performed with 6 users, a totally blind person since 4 years old and other 5 person totally blind since birth. The first experiments included identification of vertical lines, horizontal lines, diagonal lines and curves. After 20h-40h of training, the users started to recognize the line orientation with 100% of accuracy. A next experiment comprised recognition of combined lines and objects from 25 classes, like telephone, chair and cup. The authors didn't published the detailed accuracy for object recognition but explained that the users were

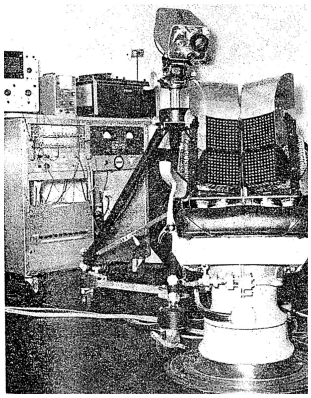


Figure 1: Tactile Vision Substitution System (Bach-y-Rita et al., 1969).

capable of recognize perspective variation and distance variation in function of the size of the object.

2.2 PhD Thesis of Mauro Pereira Conti

This research was presented by Pereira (2006) from São Paulo University and resulted in a prototype that performs vision substitution using a camera and an electro-tactile matrix positioned in the user's abdomen. The image is captured using a simple camera and a customized hardware was developed to activate the electrodes. The input image is processed using a PC for edge detection, the edges image is downscaled to the size of the matrix and send to the electrodes matrix.

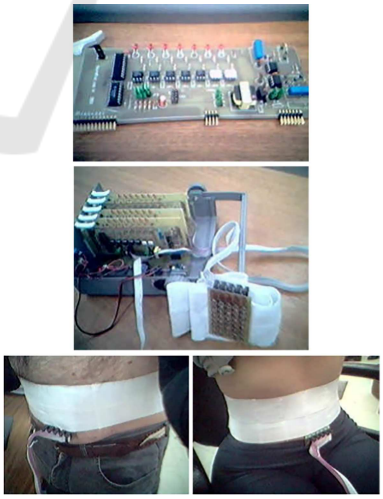


Figure 2: System developed by Pereira (2006). Top: Controlled board that activates the electrodes. Middle: Main board with controller boards connected. Bottom: Electrodes matrix worn by the user.

The author cites the limitations of real-time processing and the limitation regarding to the hardware

size. In the Figure 2 is possible to see big controller boards that makes the system unfeasible to use in everyday activities. The use electrodes also requires the use of a conductive gel to improve conductivity.

In the first experiment, users were subjected to recognize lines. The group of users with people blind since birth scored 88% of right answers and the group with normal vision people scored 70%. In the second experiment the users were presented with complex geometric symbols, like letter L, square, triangle and circle. The group of users with people blind since birth scored 80% of right answers and the group with normal vision people scored only 44%. The last experiment was intended to show different objects to the users and all the users scored less than 30% correct.

2.3 BrainPort

The BrainPort V100 Vision Aid (Stronks et al., 2016) is a commercially available device that captures the image using a simple HD camera and outputs the signal of the downscaled image into a 20x20 electrotactile tongue display. According to the product specification, after some training the user is capable of identify light variation, detect simple objects, recognize small words and detect movements. This system is an evolution of the TDU (Tongue Display Unit), developed by Bach-y-Rita et al. (1998). The price of this system is around US$10,000.00 and is available at USA, Europe and Hong Kong.

The system comprises a portable processing unit (Figure 3) that receives images from a camera mounted in a pair of glasses and allows the user to adjust zoom and contrast. The system converts the image to grayscale and reduces it size to 20x20, mapping each pixel to a point in the electrodes matrix. The brighter values are responsible to activate the electrodes in a higher voltage and the darker values will generate lower voltage. The Figure 4 shows the captured image and the output image.

Different kinds of tests were made to evaluate the BrainPort system, including object identification, text recognition, light variation, contrast and mobility. We highlight here the results obtained for object identification and text recognition. The experiments were made with 18 visually impaired people and the goal was to identify if the can identify objects in 4 classes: ball, banana, text marker and cup. After 15–20 hours of general device training, the subjects had an average correct rate of 75% in the object recognition task. For the word recognition task, 10 words were used having 3-5 letters and the subjects had an average correct rate of 15%. The author relates the poor results to the low resolution images.

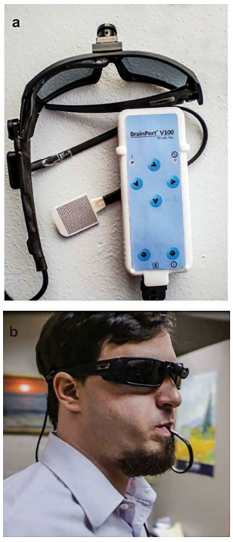


Figure 3: (a) The BrainPort V100 system. (b) BrainPort in use by a user. (Stronks et al., 2016).

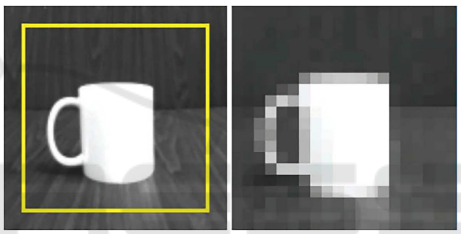


Figure 4: Example of image captured using the BrainPort system and the 20x20 output image. (Stronks et al., 2016).

In addition to the research experiments, clinical trials were conducted to allow commercial use approval. The same tasks were applied to 75 totally blind individuals. After 10 hours of training, the average correct rate for object identification was 91% and for word recognition was 58%.

2.4 VEST

Our prototype is very similar to the VEST (Figure 5). Although it's a system developed for hearing-impaired and not visually impaired, the work of Novich (2015) stands out by using a more elaborate method to encode the audio information in a vest with 26 vibro-tactile actuators and explains that the information encoded in these actuators must be the result of a function that the brain is capable of decoding without generating confusion about the input signal. The results obtained with the VEST proved the brain's ability to understand complex patterns.

To encode the information from sound to touch the author uses the k-means clustering method trained with english sentences. The number of centroids

Figure 5: VEST System Novich (2015).

is equal to the number of vibration motors in the vest and the algorithm is applied in certain intervals of the input audio data, resulting in 26 frequency bands mapped to the motors.

After 12 days of training with a 48 words dictionary and after achieving a correct rate of 75% in the training set, the subjects were submitted to a test set with different words. The performance ranged approximately 35% to 65%, considered a very good result since the individuals haven't never seen the test set patterns before.

3 PROTOTYPE

Considering the tasks of identify objects in a unknown environment without any assistance and based on the theory of the brain's plasticity, our work proposes a system similar to the one developed by Novich (2015) but focused on visually impaired people. The system proposes vision substitution using a vest with vibration motors positioned in the back. One of the reasons for choosing to work with a vest was due to the possibility of the user performing day-to-day tasks without obstruction of any other sense.

Tapu et al. (2014) describes some requirements needed for an electronic travel aid. These requirements can be applied to our solution and they are presented in the Table 1.

3.1 Hardware

In the development of the first prototype, we tried to meet the maximum possible of requirements quoted by Tapu et al. (2014). The system created is wearable, portable, reliable, inexpensive, user friendly and does not use cables. The robustness will not be evaluated at first, because the tests will be done in controlled environments. Since it is a prototype, unexpected situations can occur even in these environments and user safety will be prioritized.

The first prototype consists of a vest with 10 cylindrical vibration motors positioned in the back, a pair of Arduinos model Uno R3, a Raspberry Pi 3 model B+ and a simple HD camera.

Table 1: Requirements needed for an electronic travel aid. (Tapu et al., 2014).

Requirement	Description
Real-time	The system should promptly send messages to the user as soon as they're processed.
Wearable	It should be worn by the user. The ears and hands should be free.
Portable	It should be lightweight and easy to mount, which can be carried over long distances, small and ergonomically shaped.
Reliable	Must have a good correct rate and recall evaluation. However, it must also have correction functions for unexpected situations.
Low cost	It should be commercially accessible to users.
Friendly	Simple to use, easy to learn, no long and expensive workouts.
Robust	The device must resist to difficult environments and misuse.
No cables	TThere shouldn't be wires that limit the user's mobility.

The Raspberry Pi board was chosen because it is a low-cost processing unit and portable enough to be carried by the user. This card has a ARM Cortex-A53 64-bit processor with 4 cores of 1.2 GHz, 1 Gb of RAM and is capable of running different operating systems. The Raspbian distribution of the Linux operating system was specifically developed for this hardware. The Raspbian system is a lightweight and capable of running the TensorFlow software, which was used for the inference process of the deep learning models.

The choice of the Arduino boards was mainly due to the ease of prototyping and the number of PWM ports available to activate the motors. The Raspberry Pi card has only 4 PWM ports, which is little for the proposed design. Therefore, the task of activate the motors was centered on the Arduino boards and the inference of the models on the Raspberry Pi.

The cylindrical motors were used due to the fast response between fully active and fully static, also it's easy to control the frequency using a standard voltage of 3V to 5V in the model used. Each motor can be activated or deactivated to obtain the desired number of actuators in operation according to the experiment.

In addition, it is also possible to change the positions of the motors, since they are fixed with velcro strips, which makes it possible to carry out tests with different configurations, as can be observed in Figure 6.

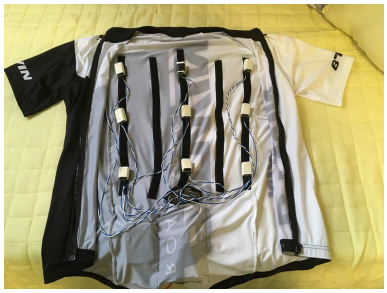
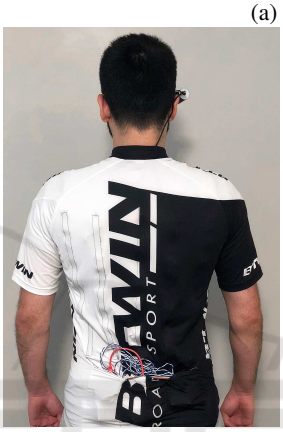


Figure 6: Front facing image of the prototype showing the internal part. In this image, the motors are configured as a 3x3 array.

The motors were mounted in the internal back of a cycling vest. The vest is made of a an elastic material, so that they stay close to the body and give the user the sensation of vibrations and differentiation when the motor is on or off . For powering the motors an auxiliary battery was necessary, since the current supplied by the Arduino was not enough. The Figure 7 shows how the vest is when it is worn. The Arduino boards, the battery and the Raspberry Pi are placed in the back pockets. The camera is attached to a regular eyewear and connected to the Raspberry Pi.

The camera is connected to the Raspberry Pi through USB. The captured image is processed using Tensorflow to perform inference of a deep learning model. The output of the model is post-processed and sent to the Arduinos. The signal sent indicates which motors should be activated. Each Arduino board has only 5 PWM ports in the Uno R3 model, so we needed two boards connected through the I2C protocol. The figure 8 presents a diagram of the prototype that was developed.

(a) Front



(b) Back (c) Camera detail

Figure 7: Our vest prototype. The boards and batteries are stored inside the back pockets of the vest.

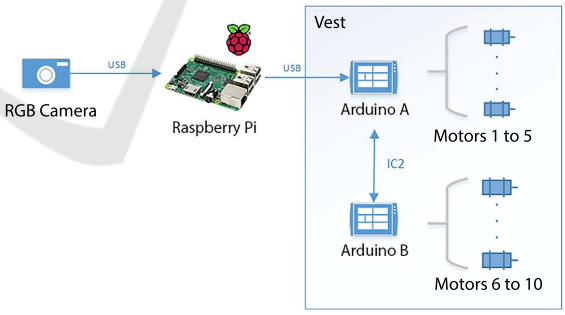


Figure 8: System architecture.

3.2 Software

3.2.1 Classification and Localization using ResNet-50

In our first attempt to encode image to touch we tried the ResNet-50 (He et al., 2015) model pre-trained with ImageNet. Our initial idea was to use a lightweight classification model that can also output the spatial location of the object. The ResNet-50 model uses global average pooling layers allowing us to estimate the position of the classified object using the process described by Zhou et al. (2015). Each activation map of the layer before the global average pooling works as a pattern detector in the image and the weights that connect the last two layers of the model represent the contribution of each of these patterns. In order to obtain the activation map that contains the location of the classified object, we sum the outputs in the activation layer weighted with the contributions in the last layer.

For testing purposes we selected two objects and defined a fixed signal that activates each object. The first object was a guitar, that activates the vibration motors horizontally from left to right. The second object was a laptop, that activates from right to left. We

got a pre-trained model with the ImageNet dataset.

The output of the global average pooling layer in the ResNet-50 model has a 7x7 shape. As we have a limit of 10 motors in our prototype we needed to reduce the size of the global average pooling to 3x3 to send the information regarding the location of the object. A new route was created after the existing global average pooling layer with one more global average pooling layer, downsizing the activation map to 3x3. The vibration motors in the vest were configured as the same, allowing us to map the signal directly to it.

3.2.2 Object Detection with SSD

Even estimating the location of the object through global average pooling layers, the use of a classification model has limitations for the proposed application. There is no information regarding the number of detected objects and it's always predicting the class with the greatest probability. This is critical in a system that sends the signal to vibration motors, because the signals are sent to the motors all the time causing confusion to the user. To work around this problem, we started using an object detection model. The chosen model was the Single Shot Detector (Liu et al., 2015) with MobileNetV2 (Sandler et al., 2018) as backend that has a good balance between performance and accuracy, 22 mAP in the Microsoft COCO dataset (Lin et al., 2014) and average 1 FPS running inference on the Raspberry Pi.

The Microsoft COCO database has 90 classes. However, in order to attend this experiment we used objects of only 4 classes (cup, remote control, scissors and bottle) and the layout of the motors was modified in relation to the previous experiment, from 3x3 rows and columns to 4x2. The activation of the vibration motors is done per row, where each pair of motors in each row is activated when detecting an object among these 4 classes.

Since the setup of the motors was modified, we also changed the output signal indicating the position of the object. To map the location with respect to the input image, it was divided into 8 quadrants according to the configuration of the vest, in 4 rows and 2 columns. Since the model already provides the position and size of the bounding box, we calculated the area of it in each quadrant obtaining the intensity value for each motor, as can be seen in Figure 9.

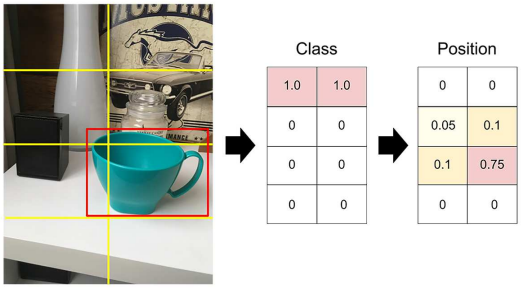


Figure 9: Detected object (cup) and the values of activation for the vibration motors for class and position. **Left:** Input image with the 4x2 grid and the detected object. **Center:** Values send to the vibration motors to indicate the cup class. This value is hardcoded from a defined dictionary. **Right:** Values send to the vibration motors to indicate the position of the object.

4 DISCUSSION AND FUTURE WORK

In the first attempt we used a classification model in which it was possible to perform a simple classification in the image and take advantage of the global average pooling layers to locate the object. However, this model has a limitation for the proposed use case, mainly by constantly activating the vibration motors that can confuse the user. In order to solve this problem, we decided to use the SSD object detection model.

From the tests performed in our laboratory using the vest with the SSD object detection model, it was possible to send signals to the vest that represents the objects and the position of this objects in sequence. By using the vest with a real-time detection, it was possible to clearly perceive the difference between signals sent that represents different objects. The signal regarding the location of the object seemed a little more confusing at first, but after a few minutes of training the signal started to make more sense. This phase was intended to check the system operation, including the communication between all the parts and validation of the signal send to the vest, that will enable the next steps of our research. No evaluation were made with visually impaired individuals yet.

Novich (2015) explains that the encoded signal must be a function for the brain to be able to learn to decode the input. Considering this, adjustments can be made in the object detection model to create outputs that are functions of the learned features in the hidden layers of these models. That is, a possible option for a next version is to use the features learned by the network to generate a pattern that is similar when the input images are similar. To achieve this goal we

will perform experiments with clustering algorithms (such as k-means), dimensionality reduction layers to match the number of vibration motors, among other techniques that will be investigated.

After having consistent results using the vest in the laboratory, we plan to perform experiments with visually impaired users.

ACKNOWLEDGEMENTS

The authors thank Fundo Mackenzie de Pesquisa (Mack-pesquisa) from the Universidade Presbiteriana Mackenzie for the financial support for this research.

REFERENCES

Bach-y-Rita, P., Collins, C. C., Saunders, F. A., White, B., and Scadden, L. (1969). Vision substitution by tactile image projection. 221:963–4.

Bach-y-Rita, P., Kaczmarek, K. A., Tyler, M. E., and Garcia-Lara, J. (1998). Form perception with a 49-point electrotactile stimulus array on the tongue: A technical note. *Journal of Rehabilitation Research and Development*, 35(4):427–430.

Bach-y-Rita, P. and Kercel, S. W. (2003). Sensory substitution and the human–machine interface. *Trends in Cognitive Sciences*, 7(12):541 – 546.

Cancar, L., Diaz, A., Barrientos, A., Travieso, D., and Jacobs, D. M. (2013). Tactile-Sight: A sensory substitution device based on distance-related vibrotactile flow regular paper. *International Journal of Advanced Robotic Systems*.

Cardin, S., Thalmann, D., and Vexo, F. (2007). A wearable system for mobility improvement of visually impaired people. *Visual Computer*.

Dakopoulos, D. and Bourbakis, N. (2008). Preserving visual information in low resolution images during navigation of visually impaired. In *Proceedings of the 1st International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '08, pages 27:1–27:6, New York, NY, USA. ACM.

González-Mora, J., Hernández, A. R., Ramos, L. F. R., Dfaz-Saco, L., and Sosa, N. (1999). Development of a new space perception system for blind people, based on the creation of a virtual acoustic space.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

Hub, A., Diepstraten, J., and Ertl, T. (2004). Design and development of an indoor navigation and object identification system for the blind. In *Proceedings of the 6th International ACM SIGACCESS Conference on Computers and Accessibility*, Assets '04, pages 147–152, New York, NY, USA. ACM.

Johnson, L. A. and Higgins, C. M. (2006). A navigation aid for the blind using tactile-visual sensory substitution. In *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 6289–6292.

Lin, K. W., Lau, T. K., Cheuk, C. M., and Liu, Y. (2012). A wearable stereo vision system for visually impaired. In *2012 IEEE International Conference on Mechatronics and Automation*, pages 1423–1428.

Lin, T., Maire, M., Belongie, S. J., Bourdev, L. D., Girshick, R. B., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S. E., Fu, C., and Berg, A. C. (2015). SSD: single shot multibox detector. *CoRR*, abs/1512.02325.

Meers, S. and Ward, K. (2005). A substitute vision system for providing 3 d perception and gps navigation via electro-tactile stimulation.

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

Novich, S. D. (2015). Sound-to-Touch Sensory Substitution and Beyond. Master's thesis, Rice University.

Pereira, M. C. (2006). *Sistema de substituição sensorial para auxílio a deficientes visuais via técnicas de processamento de imagens e estimulação cutânea.* PhD thesis.

Sainarayanan, G., Nagarajan, R., and Yaacob, S. (2007). Fuzzy image processing scheme for autonomous navigation of human blind. *Appl. Soft Comput.*, 7(1):257–264.

Sandler, M., Howard, A. G., Zhu, M., Zhmoginov, A., and Chen, L. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. *CoRR*, abs/1801.04381.

Stronks, H. C., Mitchell, E. B., Nau, A. C., and Barnes, N. (2016). Visual task performance in the blind with the BrainPort V100 Vision Aid.

Tapu, R., Mocanu, B., and Tapu, E. (2014). A survey on wearable devices used to assist the visual impaired user navigation in outdoor environments. In *2014 11th International Symposium on Electronics and Telecommunications (ISETC)*, pages 1–4.

Visell, Y. (2009). Tactile sensory substitution: Models for enaction in hci. *Interact. Comput.*, 21(1-2):38–53.

World Health Organization (2012).

Zhou, B., Khosla, A., Lapedriza, À., Oliva, A., and Torralba, A. (2015). Learning deep features for discriminative localization. *CoRR*, abs/1512.04150.