Effatá: Obstacle Identification System to help the Blind in Urban Displacement

Lukas Bergengruen
Facultad Ingeniería
Universidad de Montevideo
Montevideo, Uruguay
lbergengruen@correo.um.edu.uy

Diego Duran
Facultad Ingeniería
Universidad de Montevideo
Montevideo, Uruguay
dduran@correo.um.edu.uy

Rafael Sotelo
Facultad Ingeniería
Universidad de Montevideo
Montevideo, Uruguay
rsotelo@um.edu.uy

2021 IEEE Global Humanitarian Technology Conference (GHTC) | 978-1-6654-3372-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/GHTC53159.2021.9612497

***Abstract*— This paper presents Effatá, a product developed to help blind people walk safer and with greater ease on urban roads. The system uses a Deep Learning algorithm to detect obstacles in the user's vicinity through image detection on a Raspberry Pi 4 connected to two cameras. Once detected, the obstacle's relative position and distance to the user is estimated and informed through 3D sound via bone conduction (open-ear) headphones. The detection cycle takes 0.7 secs. Initial feedback from users, indicates that this solution has proven to be of great aid for blind people when walking through unknown sidewalks. The result product provides the user with a better understanding of surrounding objects, providing a greater sense of safety and independence.**

Keywords— blind, visual disability, deep learning, raspberry pi, 3d sound, glasses

I. Introduction

According to World Health Organization (WHO) reports in 2015 there are 217 million people with low vision in the world and 36 million are blind [1]. This number is expected to keep growing. Walking sticks and guide dogs have offered people with visual disability the independence to move relatively freely on normal sidewalks, but they don't achieve complete insertion into society. Effatá seeks to redesign the way blind people circulate daily. It works towards equality of opportunity for all human beings, regardless of their biological or health condition.

Even though some technological tools do help on blind people's inclusion in society, these have not changed much in the last 100 years.

These tools are very useful for detecting obstacles below a person's waist but are not able to recognize objects above this height and are not efficient at detecting holes on sidewalks either. Branches, street signs, and holes on the street are always dangerous and a cause of preoccupation when using walking sticks or guide dogs.

Two background papers that helped design this project are Let Blind People See [2] and Real-Time Object Detection for Visually Challenged People [3]. But unlike these state-of-the-art projects which focus on exploration assistance, Effatá focuses on day-to-day circulation of a blind person on the streets. Because of this Effata's design focuses on quickly transmitting the essential information for decision making, which include informing in a smart way both the position and distance of possible obstacles.

The objectives and design of this system were achieved with the experience and validation of four experts from the Uruguayan National Center for Visual Disability, which belongs to the Ministry of Social Development (MIDES). After a couple of meetings, a final design for the product was obtained in which the most challenging problems for blind people were addressed.

II. Objectives

The objective of this project is to develop a system capable of detecting, locating, and notifying a set of obstacles that are identified as potentially dangerous in an 8-month-period project. The specific objectives defined at the beginning of the project are:

- Build an optimized software program that detects a set of objects and street anomalies and calculates their relative position relative to the user in real time.
- Report this data back through 3D sounds indicating their direction and distance in an efficient and noninvasive way.
- Integrate the whole solution in a Raspberry Pi 4 focusing on the usability and aesthetics of the system, as well as having an energy independent and affordable device.

III. Methods and Technologies

A. Raspberry Pi 4

A Raspberry Pi is a single board computer developed by the Raspberry Pi foundation [4] which seeks to make computing accessible to all parts of the world.

The Raspberry Pi 4 model is the latest offered by the foundation to date. It has an ARM Cortex-172 processor with four cores at 1.5 GHz, 4 GB of memory and Bluetooth 5.0. The Linux operating system and all the necessary information is stored on an SD card.

B. Object Detection

This Computer Vision technique consists in identifying and locating objects in an image or video. It is often used to count objects in a scene, determine their position, or to label them.

These kinds of algorithms usually result in a set of rectangles that enclose all objects in an image. These rectangles are usually called Bounding Boxes. The efficiency of an object detector is measured by how closely the bounding box fits the object in the image with the correct label.

Single Shot Multibox Detector (also known as SSD) is an object detector which consists of a convolutional neural network. Other Convolutional Networks were considered and evaluated for the final solution. These include Yolo V3, Yolo V3 Tiny, DETR and RetinaNet. Figure 1 shows the comparison of mean inference time for each algorithms considered.

In the end, SSD was selected due to its small inference time and good performance on close objects. This allowed the final product to run in real time.

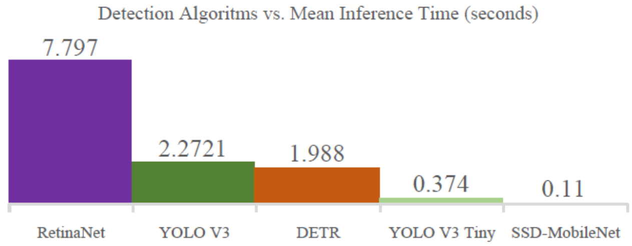

Fig. 1. Mean inference time for the detection algoritms considered.

C. *Model Training: Transfer Learning*

Transfer Learning is a Deep Learning training method in which an already trained model is used as starting point to obtain a new model for a different task. [5]. Essentially, the early layers of this model are frozen and only latter layers are re-trained for a new task. The method relies on the fact that neural networks usually try to detect edges and shapes in the early and middle layers to then work on some task-specific features in the later ones.

Transfer Learning allows developers to avoid the need of large amounts of data. Also, this often speeds up the training process and results in more accurate and effective models.

A SSD-MobileNetV2 320x320 model pre-trained with COCO Dataset was selected to be re-trained with a custom dataset of 18 classes with a total of 8178 images. This model was taken from a Git repository called TensorFlow Model Garden [6] which provides an API for object detection. This API facilitates the construction and training of detection models.

This set of classes (detectable objects) were selected with MIDES, to include only the most essential obstacles. It is of great importance that only valuable information is reported to the user, to avoid any possible overload and annoyance. This list includes possible hazards and obstacles that require the user to change path. Minor obstacles (e.g., bottles) were excluded.

The list of obstacles consists of barrels, beacons, bicycles, buses, cars, chairs, cones, dogs, fire hydrants, street holes, horses, lamp posts, palm trees, people, street signs, tables, traffic lights and trees. Most of these images were obtained from Google's Open Images Dataset V6 (also known as OID).[1] The rest of the dataset was built manually since OID does not contain all the required classes.

D. *Estimation of an Object's relative position*

In order to help blind people navigate the streets by using a device there are a couple of decisions to be made. These variants respond to three unknowns: How will we obtain information from the user's environment? How will we recognize obstacles and their relative position to the user? How will we translate that information back to the person?

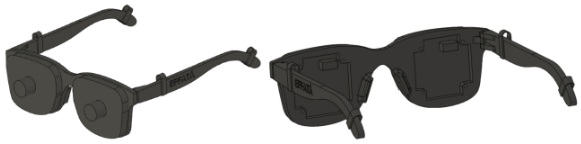

Fig. 2. Glasses design for 3D printing that holds both cameras in place.

Different solutions to each of these questions were considered. Discarded options include using a Microsoft Kinect camera, distance sensors and depth maps. These were discarded for not providing enough information or being too expensive. The final product is shown at Figure 2. It consists of a pair of lenses that have two cameras and two bone conduction speakers incorporated. After every iteration, one picture from each camera is processed by the object detection model.

By having two cameras it is possible to estimate an object's distance by taking the different positions of an object in each image. This is called the Parallax method and simulates the way our eyes estimate distance.

Once obtained the detections from each camera, a process called Stereo Match [7] is used. This process identifies which objects in one image are present in a second one. It is works by comparing features from each detected object and evaluating how similar they are. Figure 3 illustrates this method.

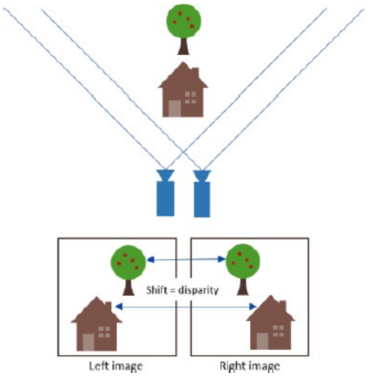

Fig. 3. Stereo Match between two offset cameras. Image taken from [7].

Once this is done, the object's distance can be estimated. The closer the object, the bigger the disparity between the objects position in both images. Equation (1) details the relationship between the object's distance, the cameras' offset and the disparity of the object between images:

$$objects\_disparity \propto camera\_offset \, / \, distance \qquad (1)$$

[1] More information in: https://storage.googleapis.com/openimages/web/factsfigures.html

In the system, all objects located more than 8 meters from the user are filtered out so as to only inform the important information. At first one may think more information is always better, but when working with a person which relies on a particular sense, the last thing we want is to block it with useless information.

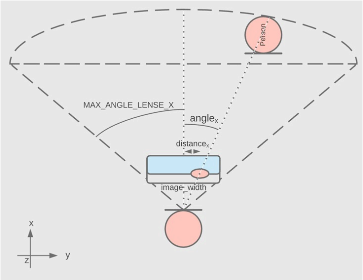

Fig. 4. Estimation of the spatial coordinates of an person relative to the user.

The next step is to obtain the spatial coordinates of the object relative to the user by taking its position in one image and its distance. With the maximum angle of the lens, it is possible to estimate the angle formed by the object. Some lenses tend to have some distortion at the edges of their images. In this project, all types of distortion were ignored, and the angle aperture of the object is directly proportional to the distance from the object to the center axis. Figure 4 illustrates the way this method works.

The formula in Equation (2) is used to calculate the horizontal angle (as for the vertical angle) of an object from its distance from the center of the image. After calculating the spherical coordinates of each object, these are transformed to Cartesian coordinates for their later use.

$$sin(angle_x) = (\frac{2*distance_x}{image_{width}}) * sin(MAX\_ANGLE\_LENSE\_X) \qquad (2)$$

E. 3D Sound Feedback

After identifying all objects and locating them spatially the user is reported of their position. In one of the interviews with experts from the National Center for Visual Disability, some concern was expressed about the danger of interfering with ambient sound. This is a crucial matter for blind people. Ambient sound is one of the main means by which they obtain information from the outside world. It was imperative that this medium was not hampered by the device.

Because of this, the solution's design, validated by MIDES, uses bone conduction headphones. These not only complement ambient sound without obstructing it but are a viable solution for people with conductive hearing loss. Bone conduction headphones transmit sound to the inner ear through vibrations in the bones of the persons face.

For each of the objects detected and located the OpenAL[2] library is used to create a 3D sound simple tone which will report the object's direction and distance. Based on the expert's opinions received at the design stage, this device intentionally hides the class of the obstacle and only informs its existence and location. Again, this is due to our desire to only transmit essential information. Throughout a user's displacement on a sidewalk, it generally doesn't matter what of obstacles are presented, if it is a person or lamp post, only its presence is important. If the solution used a text-to-speech engine for each object the solution is no longer real time and due to the fast-changing environment, some false information can be conveyed It may also be annoying if used for long periods of time.

The only exception to this rule are objects that need to be identified, like street holes. Due to its nature, holes in the ground must be distinguished from other type of obstacles. Therefore, the final product produces a higher pitch tone for these detections.

The mechanic used to inform direction is similar to how our ears work. When hearing a sound from either side of our body, sound waves first reach the closer ear creating an auditory shadow over our head. At the same time, the sound is louder on the closer ear. This comparison allows us to estimate the direction of the sound. Similarly, distance is determined by the tone's general volume. The louder the sound, the closer the object. Each object is informed individually, with a simple tone that lasts for a whole second. The pitch and enveloping curve of the tone was selected experimentally with a volunteer in order to balance comfort and ease of identification.

To provide a smarter solution, a couple of performance improvements were also applied to this stage of the process.

The detection region of the device, depicted in Figure 5, is around 60º with a maximum distance of 8 meters. It is a desired feature of the system that the angle of detection is small enough to exclude objects that do not impact the user's decision.

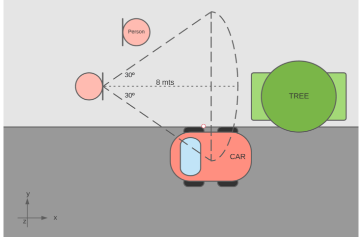

Fig. 5. Detection Region of the final product

One problem that then arises is that all predictions are relatively close to the center. Many objects tend to be notified in front of the user. Because of this it may be difficult to distinguish their position.

To deliver better information to the user, the y-axis coordinate, which determines the horizontal distance from the center axis, is exaggerated to better distinguish its variations. The reports may be a bit less intuitive at first because of this, but no major inconvenience are identified after a sensitization period. This highly improves the performance of the device.

The second performance improvement include notifying detected obstacles in a smart way. Only three objects are to be notified per iteration since too much information can be detrimental for the user's experience. If there are more than three objects detected in an image, this number should be reduced. For this two functions were implemented: detection merge and detection prioritization.

[2] More information in: https://openal.org/

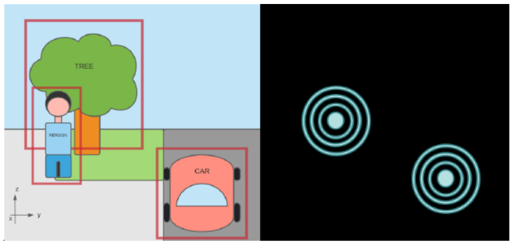

Fig. 6. Union of detections

Detection merge unifies two objects that are close to each other so that they are notified with a single tone. This can be seen in Figure 6. If after merging detections the amount of them is still more than three, then only the top priority objects are reported. The criteria used is based on the object's distance. Only the three closest detections to the user are notified.

IV. Results

A. Final Product

The final system accomplishes the initial objectives efficiently detecting obstacles in real time and reporting back to the user in a noninvasive way. The inference time of the detection algorithm for both images is of 0.7 seconds.

The different electronic components are integrated into a comfortable and aesthetic solution for everyday use, being at the same time energetically independent. The cost of the product was of around 355 USD, excluding labor work. This makes it a very affordable solution compared to other technologies in the market (which are usually over 1000 USD) and could be significantly cheaper if produced on a larger scale. This cost includes mainly the materials needed to integrate the device, which takes around 3 hours of labor work to initially assemble.

B. Test Validation

At the end of the product development, two test instances were carried out with two volunteers, both blind. In both cases a good feedback was received validating the product. The first instance was carried out at the National Center for Visual Disability. The second instance was carried out with a volunteer who, after hearing about the project, expressed his interest in participating. In each instance, four tests were carried out: Usability Validation, Distance Perception, Directionality Perception and Ability to Avoid Obstacles. A summary of the results can be seen in Table I.

TABLE I. Test Sheet Summary

Test ID	**Tests and Results**			
	Usability	***Distance***	***Directionality***	***Obstacle Avoidance***
1	User found the product comfortable but needed help to put it on.	Great results. Good identification of relative distance. Suggested reducing 8 meters to 5.	Good results. 60º of detection range. Found lag issues when moving head around.	User managed to identify the exact position of a person in its way and avoided him successfully.
2	User found the product comfortable but needed help to put it on. Commented it could be a heavy for for long periods of time.	Great results. Good identification of relative distance. Suggested reducing reachness to 8 meters.	Good results. 60º of detection range. Depending on its color, some objects were harder to detect.	User managed to identify the exact position of a person in its way and avoided him successfully.

In the future, this project would benefit of a larger test sample and test scenarios including, for example, moving objects. Ethical approval was not required since all work was motivated and advised by the highest technical authorities in the government area that deals with these issues.

V. Conclusions and Recommendations

The system developed responds effectively to the identified need and meets the requirements defined at the beginning of the project while maintaining a surprisingly affordable budget.

Some actionable recommendations would be to improve the detection model by retraining it with a new dataset that better explains reality eliminating some bias issues in some of the classes.

From a hardware perspective, a recommendation from one of the experts that helped on the project would be to try and move the cameras to a more stable location like the chest of the users avoiding confusion when moving their face to any side.

Another big improvement of this solution would be to use smaller and compact cameras, like the ones used in laptops. This would allow the device to clip to the user's glasses instead of replacing them and would permit partially sight impaired people to use the device on their glasses since the current glasses block all incoming light.

Both test subjects suggested limiting even more the range of detection of the device from eight meters to only five.

A possible next step would be adding new functionalities to this same setup. A GPS module that reports the name of streets when coming to a crossing has been identified as a possible and desirable feature to incorporate in the future.

References

[1] "Blindness and vision impairment," World Health Organization, 26 February 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment.

[2] R. Jiang, Q. Lin and S. Qu, "Let Blind People See: Real-Time Visual Recognition with Results Converted to 3D Audio," 2016.

[3] S. Vaidya, N. Shah, N. Shah and R. Shankarmani, "Real-Time Object Detection for Visually Challenged People," 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), 2020.

[4] "Raspberry Pi," Raspberry Pi Foundation, [Online]. Available: https://www.raspberrypi.org/products/raspberry-pi-4-model-b/.

[5] J. Brownlee, "machinelearningmastery," 20 December 2017. [Online]. Available: https://machinelearningmastery.com/transfer-learning-for-deep-learning/.

[6] H. Yu, C. Chen, X. Du, Y. Li, A. Rashwan, L. Hou, P. Jin, F. Yang, F. Liu, J. Kim and J. Li, "TensorFlow Model Garden," GitHub, 2020. [Online]. Available: https://github.com/tensorflow/models.

[7] J. Menant, G. Gautier, M. Pressigout, L. Morin and J.-F. Nezan, "An automatized method to parameterize embedded stereo matching algorithms," Journal of Systems Architecture, Rennes, France, 2017.