# Vidi: Artificial Intelligence and Vision Device for the Visually Impaired

R. L. A. Pinheiro and F. B. Vilela

**Abstract**

Independence and autonomy represent the capability to make your choices and act the way you want, without needing help. The visually impaired in their daily lives face several difficulties and often need help. To make users more independent, a prototype that uses artificial intelligence and narrates images from a camera attached to the glasses was developed. Some of the functions are facial, object and color recognition, all through machine learning and image recognition. Practical tests were carried out by the authors placing different objects in front of the device to verify if they would be recognized. The results were satisfactory for proof of concept about the device.

**Keywords**

Artificial intelligence • Autonomy • Machine learning • Visually impaired

## 1 Introduction

A person becomes truly independent when is able to make his own decisions and carry out his basic activities autonomously. Grimley Evans (1997), defined autonomy as "the ability of individuals to live as they want" [1]. The concept of the expression "autonomy" can be represented as: "the ability to act for oneself, to be responsible for one's own actions, without depending on others or on conditions of the environment". Adding, "it is necessary to take into account the relativity of this definition, consequently, to speak of degrees and levels of people's autonomy, since, to a greater or lesser extent, depend on others in different situations of daily life" [2]. Therefore, autonomy means, for the author, the ability to conduct his behavior independently, in certain situations and frequent events of daily life [3].

In general, autonomy is the ability to perform simple everyday tasks, which can sometimes be extremely difficult for people who are blind or have low vision, such as the act of identifying bank notes [4]. In a paper carried out at the Federal University of Paranã, blind people were interviewed. They also pointed out that changed the pattern of Brazilian banknotes, and the two reais banknote is smaller, but using only the tact, they cannot identify [4]. One of the people interviewed, an elderly woman, reporting the same difficulty for identifying cash and mentioned her strategy of separating the notes in the wallet with the help of her daughter, just because she memorizes the order of the notes and their values respectively [4]. That same paper identified some difficulties when the blind person or person with low vision goes to the supermarket. In a reported case, a person with a visual disability feels many difficulties in identifying products in the supermarket and stores [4]. Not all products have Braille identification, which are bumped dots on the packaging. So, they take the product and tries to use touch, but often they are very similar and confuse them, like canned foods. Thus, they reported that often buys the wrong products, highlighting the fact that the help of someone to make purchases in a market is almost indispensable, as the identification of the products is practically impossible [4]. Another important point to highlight is the predisposition of people around the disabled to help them. Many have a certain resistance and fear, or even help once or twice, but soon give up. Facts observed by Araújo in her paper, who interviewed some people about their daily lives, reported that people often offer help but not in the right way. For example, they want to hold the arm themselves instead of letting the blind support on the shoulder. At other times, they were abandoned in the middle of the street while crossing with the help of a person, putting they life at risk [5].

The state of the art points out some solutions that can improve this scenario. Scientific solutions led to commercial devices that seek to bring more autonomy for the visually impaired. However, the most sophisticated and complete

R. L. A. Pinheiro (✉) · F. B. Vilela
National Institute of Telecomunication, Santa Rita do Sapucaí, Brazil
e-mail: robertopinheiro@gea.inatel.br

© Springer Nature Switzerland AG 2022

T. F. Bastos-Filho et al. (eds.), *XXVII Brazilian Congress on Biomedical Engineering*,
IFMBE Proceedings 83, https://doi.org/10.1007/978-3-030-70601-2_200

being imported, they have a high cost, reaching the range of three thousand dollars. These devices are capable of transcribing static texts, some colors and product labels [6]. There are also applications for smartphones with features that assist the disabled people in some activities. Pay Voice for example, is an application to check the value to be paid written on the card machine [7]. For notes, there is the Cash Reader, an application identifies money bills [8]. By pointing the phone towards the money, the application speaks aloud what the banknotes are worth. Eye-D makes texts convert to speech [9]. But the most used today is Be My Eyes. It works as a support network between people who see completely and visually impaired, promoting video calls so that a user with perfect vision can describe to the other drawings on the screen and read texts [10].

Based on the information obtained, the present paper aims to create a proof of concept of an artificial intelligence and vision support device, which can be attached to regular eyeglasses. It is able to transcribing the objects, faces, colors and texts for the user. Therefore, it converts the images captured by the camera into audio by an artificial intelligence system. The images are processed by an embedded algorithm, using OpenCv and developed in Python within a Raspberry Pi 3B+, that would be fixed in the belt or stored in a backpack. The main objective of this project is the development of a light, versatile and efficient device that can give greater autonomy and facilitate non-dependent relationships for people with visual disability.

## 2 Materials and Methods

The project started with a search for existing technologies, seeking all solutions focused on helping people with any level of visual disability. A bibliographic search was made to identify the main technologies available in computer vision (CV), where similar works were found, which proposes the use of CV with the return in 3D sounds, and also works with a focus on mobile applications (APP), which also with CV provides assistance for cross the street at the crosswalk, or study applications that help in the academic scope [11–13]. Tests were also carried out on three APPs available for download highlighted in the article "Apps for different visuals" by Mellina, visually impaired who uses them frequently and portrayed them as being the best, they are Aipoly Vision, Seeing Ai and Tap Tap See [14].

Based on the information obtained, five main functionalities were designed to meet the minimum requirements identified as important:

- Voice recognition: the user interacts with the device through speech, requesting the task to be performed;
- Natural language reproduction: the device can reproduce responses to user requests;
- Object identification: when an object already saved in memory is placed in front of the disabled person, an audio with the name of the object will be played;
- Face recognition: when a person already stored in the system is in front of the user, her name will be reproduced;
- Color detection: the device returns in audio the predominant color in front of the disabled person.

To validate these functionalities, some basics tests had been developed. Initially, they were performed only by the authors. With the device positioned in the glasses, the head was directed to a surface, initially empty, and requested by the voice for the device to start the recognition. Then a water bottle was placed on the surface and wait for processing. The next step was to put an apple next to the bottle, to check the recognition of two objects together. The device then was directed to a person to check facial recognition. After these steps, a sheet of colored paper was placed in front of the user, and they requested the color identification. A moving test was also performed. The person walks in a hallway that is one meter wide. Objects were along the way, to verify that the device is able to inform them to the user.

### 2.1 Hardware

With the functionalities defined, the development of the prototype began. As the main hardware, a Raspberry Pi B3+ was used, which has a Broadcom BCM2837B0 64-bit ARM Cortex-A53 Quad-Core processor and 1GB of RAM. Figure 1(a) highlights the device, which is a compact minicomputer that has all the main components of a computer on a small card sized [15,16]. Other Raspberry models will be tested for comparison purposes only.

Together, a Microsoft Lifecam Cinema webcam—H5D-00013, shown in Fig. 1b, was used in order to capture images in real time and with a 720p definition, which exceeds the needs of the project. This camera was selected not only for having high resolution, but also for having a built-in micro-

phone, facilitating the development and reducing the costs of the prototype [17].

For the audio feedback, a mini 30 mm speaker, with 32 Ω and 100 dB, illustrated in Fig. 1c, was used.

(a) Photo of Raspberry Pi B3 +

(b) Microsoft Lifecam Cinema-H5D-00013

(c) Views of the used speaker

**Fig. 1** Hardware parts

## 2.2 Software

Development started with speech recognition and natural language processing. Application programming interfaces (APIs) of Google, Text to Speech (TTS) and Speech to Text (STT), were selected, which make it possible to transform audio into text and vice versa [18,19].

Having the user interface ready and totally non-visual, image processing was the next step, for this the Open Source Computer Vision Library (OpenCV) was selected. It is an open source computer vision and machine learning software library that was built to provide a common infrastructure for computer vision applications and to accelerate the use of machine perception in the commercial products. Being a Berkeley Software Distribution (BSD)-licensed product, OpenCV contributes for the utilization and modification of codes [20]. To expand the possibilities of OpenCV, the TensorFlow (TF) library, from Google, was used. TF is an end-to-end open source platform for machine learning (ML). It abstracts the pattern recognition to classify images, using the Convolutional Neural Networks (CNN) architecture. It has a comprehensive, flexible ecosystem of tools, libraries and community resources that lets researchers push the state-of-the-art in ML and developers easily build and deploy ML powered applications [21].

There are three points that need to be shown to train a neural network: the anchor, positive images and negative images. The anchor is the image of the reference face or object that will be identified. The positive image is an image that also contains the item, and the negative image does not have the item to be identified, in other words, does not have the same identity. The point is that the anchor and positive image both belong to the same face or object while the negative image does not. The neural network computes the 128-d embeddings for each item and then tweaks the weights of the network (via the triplet loss function) such that the 128-d embeddings of the anchor and positive image lie closer together. While at the same time, moving away from the embeddings for the negative images. In this manner, the network is able to learn to quantify and return highly robust and discriminating embeddings suitable for the detection [22].

Some features require an internet connection to use. It is necessary because part of the data processing is done remotely. This is the case with the TTS and STT interface features. For this, a Wi-Fi connection was used.

## 2.3 Physical Structure

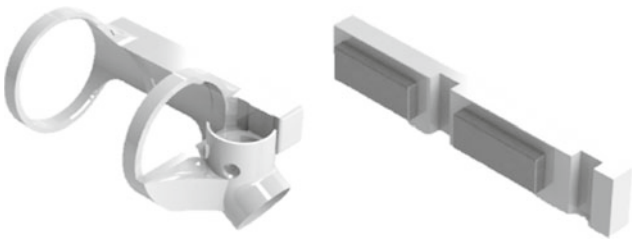
With the electrical part ready and the virtual analysis algorithm completed, the project physical supports were developed. It was used the computer program SolidWorks and a 3D printer filled with acrylonitrile-butadiene-styrene (ABS) to print part of the structure. The structure has a support for the camera to be fixed (this is shown as part A in Fig. 2a), and a second support to be attached to glasses rods by means of plastic seals (this can be seen as part B in Fig. 2b). It was tested a very thin and simple structure, to allow any glasses to be attached to the rods.

**Part A**: in it, the camera, the activation button and the speaker were coupled. It is the part where all the technology and the two magnets were attached, allowing the user to use it in his hand if he feels comfortable or taking it out to store or change glasses without losing any functionality. The speaker is positioned to direct the sound to the ear of the user, so as not to obstruct the hearing of the disabled, as a headset would. This is shown in Fig. 2a.

**Part B**: this part is the cheapest one and it can be replaced if necessary, allowing the user to have one for each pair of glasses. There are only two magnets in it, which have the right polarity for fitting Part A in the exact position.

# 3 Results

The finished support, with the camera and its respective magnets can be seen in Fig. 3. The component connection model can be seen in Fig. 4.



(a) Part A (b) Part B

**Fig. 2** Separate parts

Furthermore, the logical operation is illustrated in Fig. 5.

The final prototype is shown in Fig. 6, and has reached the expected goal at this stage. It was only observed that the size is very large, and the part attached to the glasses is very heavy.

The tests mentioned in Sect. 2 obtained the following results: The device understood all the voice commands and recognized the two objects used. It pronounced them correctly. The recognition of the face and color was also successful. The device was able to identify the authors' faces and speak their names. In the movement test, while the authors were walking, the device repeated the objects in front of them, until they left his field of vision.

In tests carried out, it was possible to identify personal objects on a table, some obstacles on the way and people ahead, including their faces when the person had already been registered, as well as interacting with the system in a practical and simple way. The lack of some simple functions was identified, like knowing what time it is, or even identifying the weather. So it was implemented in the project, using Google Weather tools for the weather, and Google Time to identify the

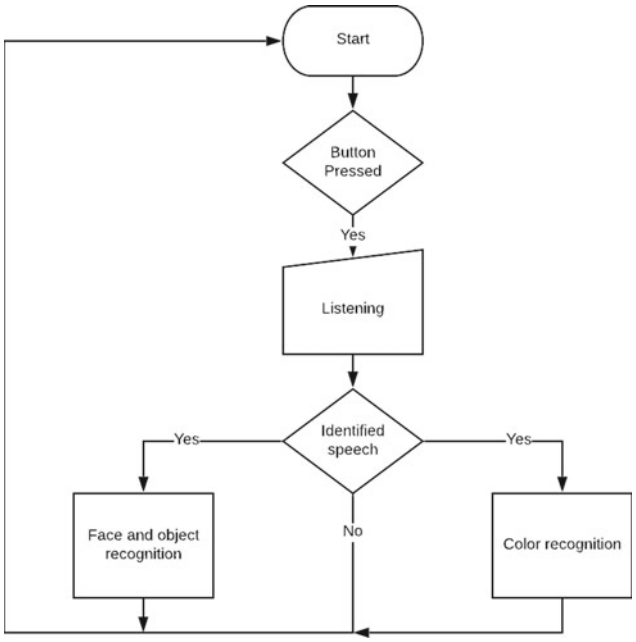



**Fig. 5** Simplified flowchart of the system

**Fig. 6** Finalized prototype

**Fig. 3** Parts A and B together

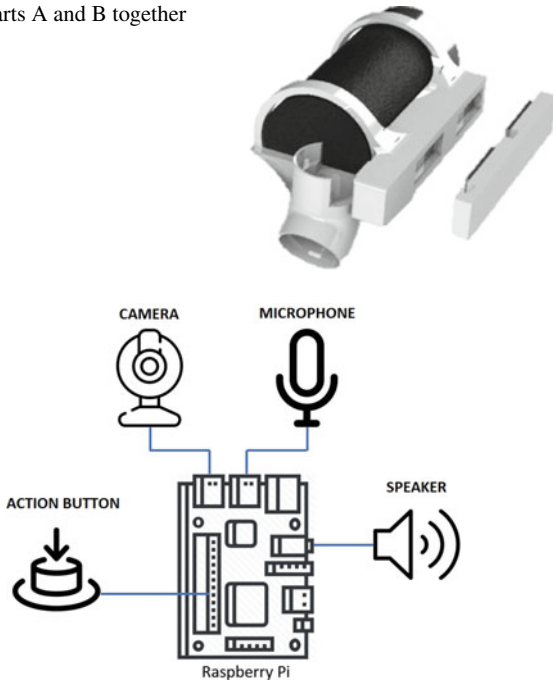



**Fig. 4** Block diagram of the system

hours. Then, when the device is asked for any of its functions, they check on the internet and it will speak.

In the first tests, the Raspberry Pi Zero W was used, which is much smaller than that mentioned in Materials and Methods (Pi B3+). However, due to its Broadcom BCM2835 ARM11 1 GHz Single-core processor and its 512 MB of RAM, not being enough to execute the code perfectly, overheating and occupying all its processing causing latency [23]. Figure 7 presents a photo from the internal thermometer of the Raspberry where the problem can be observed.

The tests with object recognition were performed too and can be seen in Fig. 8, with that some objects that had already been processed before, would be recognized by the device and TensorFlow generates the rectangle automatically.

It is possible to add any objects, the process is as follows. First approximately three hundred pictures of an object are taken, in different positions and angles, as in the example with playing cards in Fig. 9.

Then, with all the images properly marked, using a TensorFlow tool, the images are analyzed and processed by the neural network, resulting in a xml file with a transcribed image mapping that represents the pattern playing card. Importing this pattern in the algorithm, a playing card is identified [24].

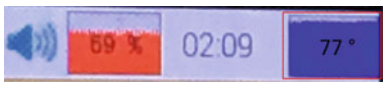



**Fig. 7** Photo of the internal thermometer showing high temperature

**Fig. 8** First test performed, the identification of a chair can be observed

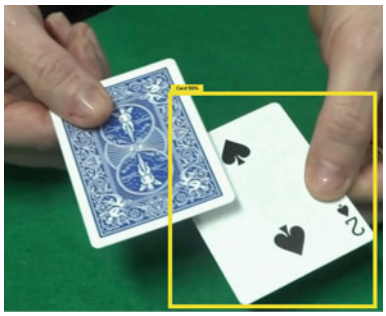


**Fig. 10** Similarity percentage

When an object is identified by the camera, the software provides a percentage of similarity compared to the object saved in memory (anchor), if this percentage is greater than the set point established, which in this case was 85%. It states that the object seen by the camera is actually the saved object, as can be seen in Fig. 10.

For facial recognition, the principle is the same as object recognition. However, it should be applied in two key steps using OpenCV resources: detects the presence and location of a face in an image, but does not identify it, and extract the 128-d feature vectors (called "embeddings") that quantify each face in an image. Thus, after having a face identified and cut out of the video frame, it will be processed in the same way as object detection. The neural network computes the 128-d embeddings for each face and then tweaks the weights of the network, via the triplet loss function [22]. The process is depicted in Fig. 11.

In color recognition, the process is a little different. When the device is asked to make the identification, a photo is taken by the camera. Therefore, the color pattern that is most prevalent in the image is defined with the color to be identified.

Three filters are applied to this photo, being defined by red, green and blue (RGB) characteristics and through the color balance and set points defined for the percentage of each color present. The concepts related to the hue circle in the hue, saturation and value system (HSV), highlighted in Fig. 12, are used to make the color definition [25].

In Fig. 13 it is possible to observe the filters applied to an image with a "mixed" color, to identify the color percentages.

**Fig. 9** Photos taken from playing cards for machine learning [24]

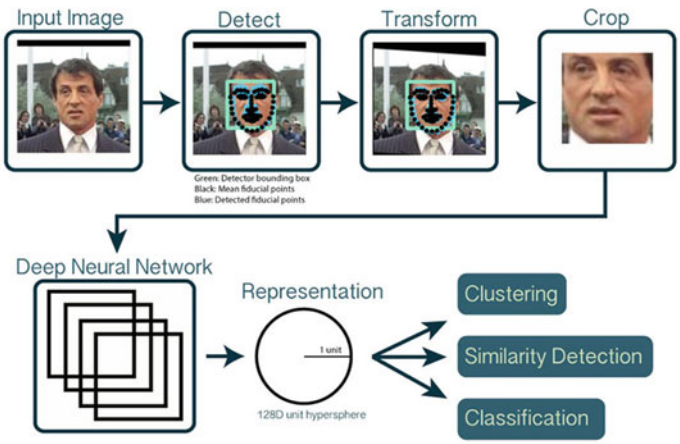



**Fig. 11** An overview of the OpenCV face recognition pipeline. The key step is a CNN feature extractor that generates 128-d facial embeddings [22]

## 4 Discussion

When tests were realized using the prototype, by the authors, due to its high weight, discomfort was observed after a short interval of use, explaining the need to reduce the total mass of the device. The next step is to send a request to the ethics committee to carry out tests on humans in a larger sample group.

Despite the satisfactory results, after the adjustments in the hardware, high latency was still observed for the application of the device, being possible that when the user was warned of an object in the way, it was already too close for some action to be taken. The need to look for a new processor should be analyzed.

For the object learning process, it is also important to take images with various other non-desired objects in the pictures and pictures with multiple objects. To be able to detect the objects when they are overlapping, make sure to have it the overlapped in many images. The size of the images cannot be very large (maximum 720 × 1280 pixels). With the Label

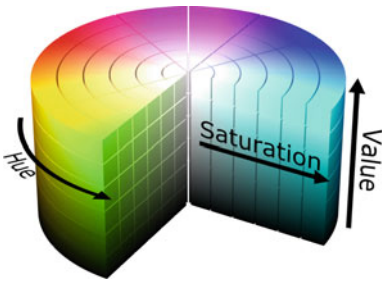



**Fig. 12** Demonstration of Hue, Saturation and Value

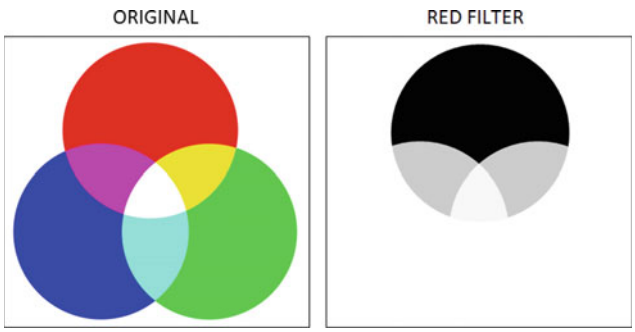



**Fig. 13** Red filter applied to an image with mixed colors

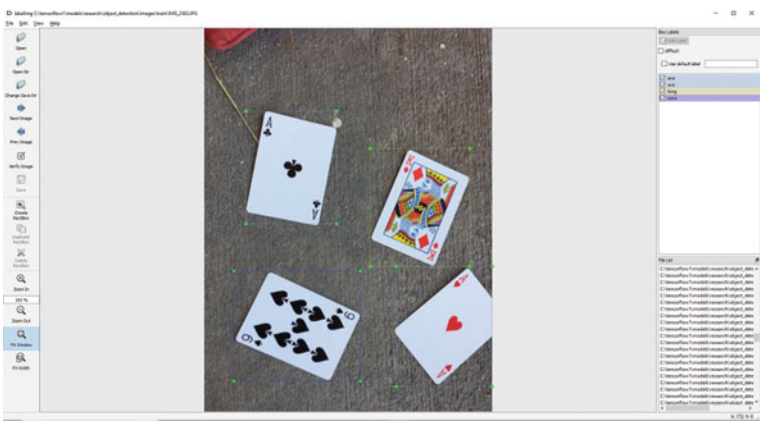


**Fig. 14** Marking the playing cards, image used as anchor [24]

Pictures software, at least twenty percent of the photos mark the object to be identified, as shown in Fig. 14, even with the example of playing cards [26].

To improve usability, the mobile network can be implemented, to make it independent from a Wi-Fi network. Another point to highlight is the voice recognition. The voice was well recognized and interpreted, but in noisier places, the device can present a little difficulties for processing the audio due to the location of the microphone (built into the camera). Thus, changes in the hardware for the microphone and its position must be considered.

## 5 Conclusion

The project started with the concern that despite different solutions designed for people with visual disability, these individuals still have difficulties in their daily lives. In addition, there is not always someone to help or a person willing to help correctly. Then, a previous review was made seeking for existing technologies applicable and a survey of the minimum necessary requirements.

The development was initiated in order to use tools that, in addition to simplifying and speeding up the design of the solution itself, were freely available and license-less for the use of developers. During this stage, some difficulties were encountered, with hardware and software, but they were overcome with good tools such as code classes, reliable documentation and changes in the development board.

During the tests, some points to be reworked were observed, such as the weight and size of the device, as well as the latency, which is an extremely crucial point for the safety of the user. Hardware issues were observed in the speech recognition, which failed when in noisy places.

Future steps for this project are to reformulate the project based on the problems already highlighted to develop hardware solutions that meet the mentioned needs. In terms of software, no irregularities or unreliability were observed in this first prototype.

**Acknowledgements** The authors would like to thank the Center for Development and Transference of Assistive Technology (CDTTA) for the availability of resources, equipment and infrastructure.

**Conflict of Interest** The authors declare that they have no conflict of interest.

## References

1. Johnstone D (2001) An introduction to disability studies, 2nd ed. David Fulton Publishers
2. Garcia RM (1999) Programa Básico para Favorecer a Autonomia Pessoal e a Vida Diária. Cooperativa de Educação e Reabilitação de Crianças Inadaptadas. 135 p. il
3. Santos MER, Rodrigues F, Rodrigues D (2006) Serviço Social e Deficiência Mental: a perspectiva subjectiva da qualidade de vida. ISMT
4. Carolina Aoki LA (2014) Análise de acessibilidade para pessoas cegas às embalagens. Trabalho de Conclusão de Curso (Graduação) - Universidade Tecnológica Federal do Paraná
5. Simone Moraes AS (2015) O encontro de pessoas cegas e não cegas pelas ruas do Recife. Revista Iluminuras - Publicação Eletrônica do Banco de Imagens e Efeitos Visuais - NUPECS/LAS/PPGAS/IFCH/UFRGS. 16:97–117
6. Mais Autonomia - OrCam MyEye® 2 at https://maisautonomia.com.br/produto/orcam-myeye-2-0/
7. Pay Voice by ABECS at https://appadvice.com/app/pay-voice/1344943724
8. Cash Reader—a money reading mobile app for blind and visually impaired at https://cashreader.app/en/
9. Eye-D—technology as a caring friend for blind and visually impaired at https://eye-d.in/
10. Be My Eyes—bringing sight to blind and low-vision people at https://www.bemyeyes.com/
11. Dias CM (2018) Alvisku: uso da visão computacional e sons 3D para auxílio a cegos. Engenharia de Computação JMV
12. Aparecida Oliveira SK (2013) Uso de visão computacional em dispositivos móveis para auxílio à travessia de pedestres com deficiência visual. Mestrado Mackenzie - Engenharia Elétrica e Computação
13. Cristina SJ, Jeferson Pezzuto DR, Cristina BJ (2015) Estudo de Aplicativos Móveis para Deficientes Visuais no Âmbito Académico. In: Brazilian symposium on computers in education (Simpósio Brasileiro de Informática na Educação - SBIE). CBIE-LACLO 2015 (Santo André, São Paulo, Brasil) SBIE
14. Quatro Patas Pelo Mundo—Aplicativos para Deficientes Visuais at https://www.4pataspelomundo.com/aplicativos-para-deficientes-visuais
15. Raspberry Pi 3 B+ at https://www.raspberrypi.org/products/raspberry-pi-3-model-b-plus
16. Olhar Digital—Raspberry Pi o que e para que serve e como comprar at https://olhardigital.com.br/noticia/raspberry-pi-o-que-e-para-que-serve-e-como-comprar/82921
17. Microsoft—Lifecam Cinema at https://www.microsoft.com/accessories/pt-br/products/webcams/lifecam-cinema/h5d-00013
18. Google—Text to Speech at https://cloud.google.com/text-to-speech
19. Google—Speech to Text at https://cloud.google.com/speech-to-text
20. Open CV at https://opencv.org
21. TensorFlow at https://www.tensorflow.org/
22. Py Image Search—OpenCv Face Recognition at https://www.pyimagesearch.com/2018/09/24/opencv-face-recognition

23. Raspberry Pi—aspberry Pi Zero W at https://www.raspberrypi.org/products/raspberry-pi-zero-w
24. Edje Eletronics—TensorFLow Object Detection API Tutorial Train Multiple Objects Windows 10 at https://github.com/EdjeElectronics/TensorFlow-Object-Detection-API-Tutorial-Train-Multiple-Objects-Windows-10
25. Py Image Search—OpenCv Python Color Detection at https://www.pyimagesearch.com/2014/08/04/opencv-python-color-detection
26. Tzutalin—Label Image at https://github.com/tzutalin/labelImg