Visual Computing Magazine

Object Recognition System on a Tactile Device for Visually Impaired

Souayah Abdelkader, MOKRETAR KRAROUBI Abderrahmene, Slimane LARABI, USTHB

Abstract

People with visual impairments face numerous challenges when interacting with their environment. Our objective is to develop a device that facilitates communication between individuals with visual impairments and their surroundings. The device will convert visual information into auditory feedback, enabling users to understand their environment in a way that suits their sensory needs.
Initially, an object detection model is selected from existing machine learning models based on its accuracy and cost considerations, including time and power consumption. The chosen model is then implemented on a Raspberry Pi, which is connected to a specifically designed tactile device. When the device is touched at a specific position, it provides an audio signal that communicates the identification of the object present in the scene at that corresponding position to the visually impaired individual.
Conducted tests have demonstrated the effectiveness of this device in scene understanding, encompassing static or dynamic objects, as well as screen contents such as TVs, computers, and mobile phones.

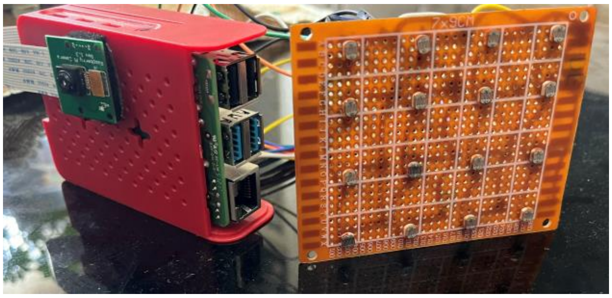


Figure 1. The developed prototype

1- Introduction

People with visual impairments face numerous challenges in their daily lives. They are unable to perceive the world in the same way as those with sight and encounter multiple difficulties, including orientation, obstacle detection and avoidance, limited mobility, and an inability to recognize shapes and colors of objects in their surroundings. In addition to these challenges, they are completely excluded from understanding and interacting with the real world scene.
Numerous technological advancements have been made to assist people with visual impairments. Among the different technological solutions deployed to address this specific need, computer vision-based solutions appear as one of the most promising options due to their affordability and accessibility.
Systems with human-scene interaction generate outputs after processing the captured scene. They consist of a set of computer vision and machine learning techniques aimed at improving the user's life in various activities such as content interpretation, navigation, etc. Generally, these systems process the data received from the real world using depth or RGB sensors and transform them into instructions and signals [1,2].
The goal of this work is to assist individuals with visual impairments in perceiving the information contained in an image by displaying the coded scene on a tactile device. They can explore the image by touching the pins on the device, with each pin representing a corresponding object in the scene.
The developed prototype is illustrated in Figure 1.

Object Recognition System on a Tactile Device for Visually Impaired

Souayah Abdelkader, MOKRETAR KRAROUBI Abderrahmene, Slimane LARABI, USTHB

2- Method

Our system aims to assist visually impaired individuals in identifying objects and their locations from images. A tactile device has been developed to provide auditory feedback corresponding to the identity of the detected object, thereby helping these individuals obtain information about the scene (see Figure 1). The proposed system is capable of identifying 17 types of objects in the observed scene.

This system is divided into three processes as illustrated in Figure 2.

The first two processes cooperate in interpreting and detecting objects in the observed scene. Since the system is embedded on a Raspberry Pi, it has limited resources (low RAM and processing power). Therefore, we have developed three object detection models, each responsible for detecting objects in a specific environment: Office, Kitchen, and Bedroom.

The main process is responsible for recognizing the appropriate environment in order to load the corresponding model. This process is reactivated at each new camera location and when the detection rate falls below a predefined threshold.

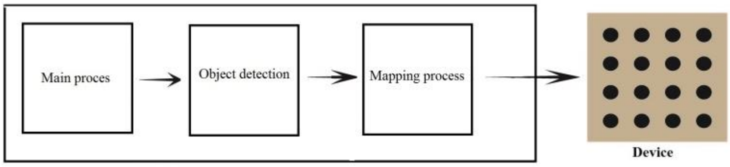


Figure 2. The processes of the proposed system

The second process is responsible for detecting objects and their locations in the image, and it transfers the coordinates of the object locations to the next process.

The third process involves associating the detected objects with a location on the tactile device and interacting with the user to produce corresponding sound feedback for the detected object.

2.1-The Detection Processes

The acquired image is input into the main process, which determines the appropriate environment or scene category (e.g., office, kitchen, bedroom) based on the visual cues and characteristics present in the image.

These three models are based on YOLOv5 and have been retrained on a dataset consisting of seven specific object classes. The goal of each model is to detect and recognize objects belonging to these seven classes in their corresponding environment.

The system operates using an object detection model that is responsible for detecting characteristic objects in each environment. Then, the k-nearest neighbors algorithm is executed to recognize the observed environment. In this algorithm, objects represent the features, and environments represent the target classes. Once the appropriate environment is determined, the second process takes over for object detection and location, while the first process is paused. It transfers the coordinates of each detected object to the final process. This process also has the responsibility of reactivating the main process when the detection rate falls below a predefined threshold (e.g., when the object detection model fails to detect more than 20% of the objects in the observed environment).

2.2-Mapping process

This process is responsible for converting the coordinates of objects in the image into relative coordinates on the tactile device. In cases where there is overlap between two objects, where both objects may appear in the same grid cell of the tactile device, or when a relatively large object occupies multiple grid cells, we have developed an algorithm to determine the order of objects belonging to the same grid cell. Additionally, this process is responsible for interacting with the user through the tactile device. It produces sound feedback corresponding to each detected object. When the user touches or interacts with a specific pin on the tactile device, a specific sound is emitted to provide feedback to the user.

2.3-Model selection

In order to select the appropriate model for the specific task of integrating an object detection model into a Raspberry Pi, we conducted a comparative study of object detection algorithms based on Convolutional Neural Networks (CNNs). Considering our objective of achieving acceptable precision and recall values while working with embedded systems, we conducted the comparison while considering the following constraints:

We focused on using the latest reduced versions of each model, commonly referred to as "tiny models," such as YOLOv5 [4], Faster R-CNN [5], and SSD [6].

The object detection models used in the comparison were trained on the same image dataset and shared the same backbone architecture. This ensured that we could make meaningful observations regarding the advantages and disadvantages of these methods.

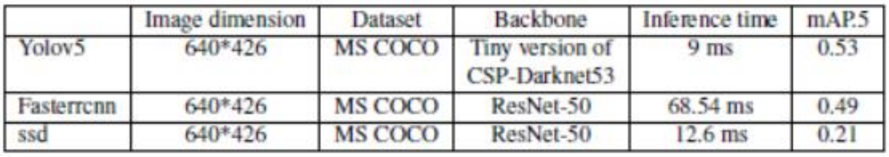


	Image dimension	Dataset	Backbone	Inference time	mAP.5
Yolov5	640*426	MS COCO	Tiny version of CSP-Darknet53	9 ms	0.53
Fasterrcnn	640*426	MS COCO	ResNet-50	68.54 ms	0.49
ssd	640*426	MS COCO	ResNet-50	12.6 ms	0.21

Table 1. Comparison results on public database

Tables 1 and 2 present the results obtained on the MS-COCO benchmark and a collected image dataset in terms of mAP0.5 (mean Average Precision at IoU threshold of 0.5). These results validate the effectiveness of YOLOv5 in comparison to Faster R-CNN and SSD. The tables demonstrate that the YOLOv5 structure is better suited for real-time applications due to its faster processing speed compared to the other structures.

The selected model, determined by the main process, utilizes a YOLOv5-based object detection method to identify and locate objects in the image. It generates bounding boxes that enclose each detected object, accompanied by confidence scores that must exceed 0.5 to be deemed valid.

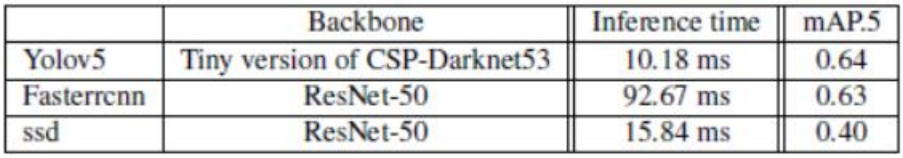
The coordinates of the detected objects, represented by the bounding boxes, are extracted from the object detection model and transmitted to the final process for encoding them on the tactile device.

Object Recognition System on a Tactile Device for Visually Impaired

Souayah Abdelkader, MOKRETAR KRAROUBI Abderrahmene, Slimane LARABI, USTHB

3- Experimental Results

In Figure 3, we present the components of our interactive device designed for visually impaired individuals.
This compact and portable device is a Raspberry Pi equipped with a high-definition camera, a 2GB CPU, and RAM. The device analyses the user's environment and detects objects in real-time. The gathered information is then transmitted to the user through a haptic feedback system.
The haptic feedback system utilizes a device with 16 photoresistor sensors, enabling visually impaired individuals to comprehend their environment using their fingers. These sensors detect the presence of fingers and convert this information into audio feedback.



	Backbone	Inference time	mAP.5
Yolov5	Tiny version of CSP-Darknet53	10.18 ms	0.64
Fasterrcnn	ResNet-50	92.67 ms	0.63
ssd	ResNet-50	15.84 ms	0.40

Table 2. Comparison results on our collected images.

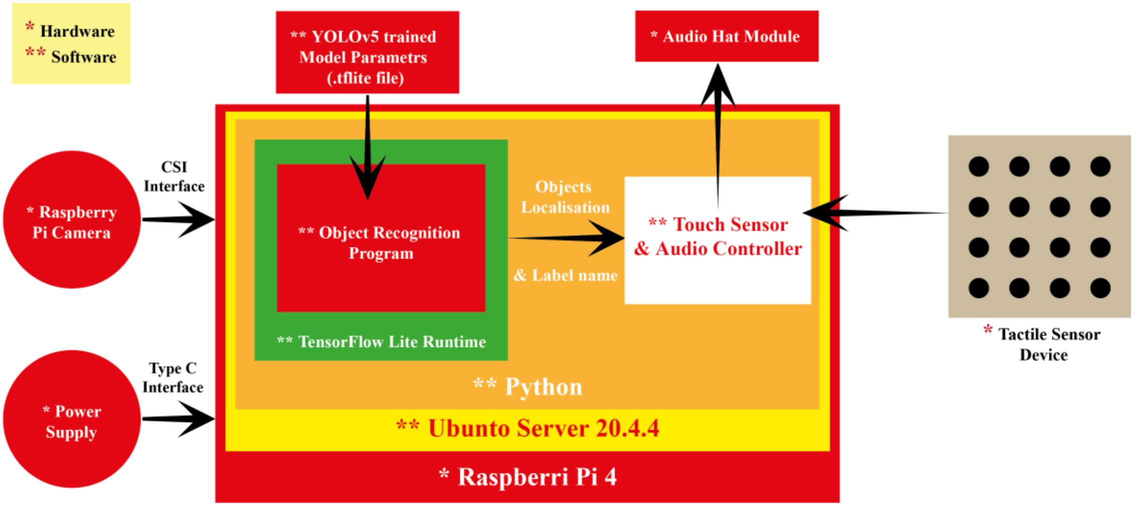


Figure 3. Components of the system

Object Recognition System on a Tactile Device for Visually Impaired

Souayah Abdelkader, MOKRETAR KRAROUBI Abderrahmene, Slimane LARABI, USTHB

3.1 Making the device

Our main goal is to enable tactile-audio interaction with visually impaired users. To accomplish this, we have opted to utilize photoresistor technology, an electronic component that exhibits varying electrical resistance in response to incident light. The resistance of a photoresistor changes inversely proportional to the intensity of light it receives. In tactile interaction, this technology can be employed to detect changes in light caused by the user's touch on a light-sensitive surface.

By arranging multiple photoresistors as pins on a surface, we can detect which resistors are touched by the user. This enables tactile interaction where the user can interact with different pins and trigger actions, such as audio feedback through an audio output module connected to the Raspberry Pi. Each touched resistor corresponds to a specific sound based on the detected object's position in the image relative to the pin.

Figure 4 illustrates the circuit connecting 16 photoresistors, with each photoresistor connected to one of the Raspberry Pi's pins. The associated code utilizes the RPi.GPIO library to manage GPIO pins on the Raspberry Pi. It configures the port for the photoresistor as an input. In the main loop, it checks the state of the photoresistor. If it is triggered (HIGH), it displays a message indicating that the user has touched the photoresistor.

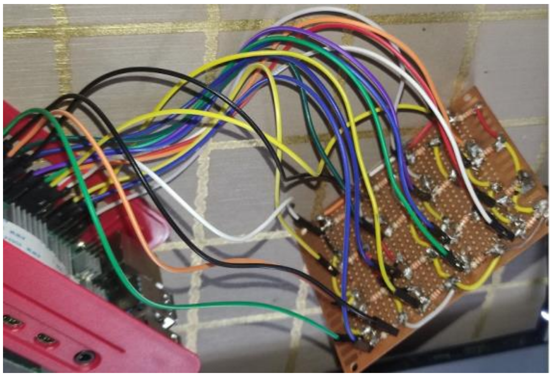


Figure 4. Connecting Raspberry Pi and the tactile device.

3.2 Object Detection

The primary objective of this project is to identify and detect 17 different classes distributed across three categories: office, kitchen, and bedroom. During the data collection phase, we obtained a total of 2677 images specifically for the office environment.

The dataset consists of a total of 2677 samples, which are divided into 7 classes representing office environments. The smallest class contains approximately 290 samples. Each class has an adequate number of images distributed across the training, validation, and test sets. The image dataset is organized into three files: train (70%), validation (20%), and test (10%).

Object Recognition System on a Tactile Device for Visually Impaired

Souayah Abdelkader, MOKRETAR KRAROUBI Abderrahmene, Slimane LARABI, USTHB

3.3 Transfer Learning

We fine-tuned and configured the YOLOv5 architecture specifically for our dataset. To achieve this, we employed transfer learning, adapting the YOLOv5 framework to be compatible with our dataset. We utilized pre-trained weights from a different model that had been trained on the extensive COCO dataset.

For training our model (yolov5s.pt), we utilized the standard Colab VM with 12GB of GPU memory. To enhance the robustness of the trained model and better utilize the available GPU resources, we set the batch size to 4. Additionally, we conducted training for a total of 100 epochs, observing that the trained model reached stability.

Throughout the experiments, we incorporated various hyperparameters. Some of these included weight decay = 0.0005, initial learning rate = 0.0042, final learning rate = 0.1, and momentum = 0.937. These parameters were maintained at their default values. Ultimately, we trained and tested YOLOv5 on the Colab VM using our dataset.

3.4 Results of objects detection

To provide a more detailed analysis of the model's training process and performance, Figure 5 (left) displays a plot showcasing the precision and recall mapping for detecting the seven classes during training. From the figure, it is evident that the model achieved a mean Average Precision (mAP) of 86.3%. This mAP value represents the area under the curve, indicating the trained model's ability to accurately detect objects with high precision and recall values.

To highlight the superiority of our selected object detection model for the desk environment (comprising the previously mentioned 7 classes), we conducted a comparison with the detection model prior to transfer learning, namely yolov5s. Figure 5 showcases the mAP results obtained by both models on a training image set.

It is evident from the results that our model surpasses yolov5s in terms of average precision for the 7 classes. It is important to highlight that the object labeled as dining table in yolov5 is distinct from the desk object. Based on the conducted experiments, our transfer learning model derived from yolov5s exhibits superior performance. Therefore, we can confidently utilize our model for the project.

3.5 Mapping

The pin grid provides an organized structure and spatial reference for each object based on its position and size. This facilitates further processing or interaction with the detected objects within the project's context.

The detected objects in the image are associated to their corresponding cells. Each object's bounding box is associated to the grid cell as long as the majority of its surface is within that cell. A bounding box can be associated with multiple cells, and likewise, a cell can have multiple bounding boxes .

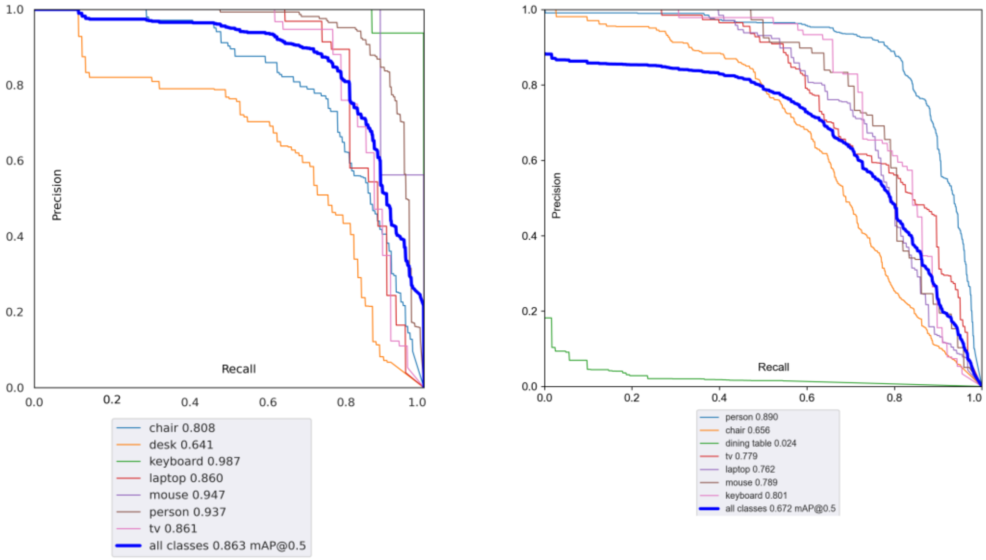


Figure 5. Precision, Recall for our model (left) and yolov5s (right) on train data.

References

[1] Zatout, Chayma and Larabi, Slimane and Mendili, Ilyes and Barnabé, Soedji Ablam Edoh.Ego-Semantic Labeling of Scene from Depth Image for Visually Impaired and Blind People. IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), 2019, pp. 4376-4384.

[2] Larabi S. Zatout C. Semantic scene synthesis: application to assistive systems. The Visual Computer, 38:2691-2705, 2022.

[3] Tsung-Yi et al. Lin. Microsoft coco: Common objects in context. In Computer Vision – ECCV 2014, pages 740–755, Cham, 2014. Springer International Publishing

[4] Chien-Yao Wang, Alexey Bochkovskiy, and Hong Yuan Mark Liao. Scaled-yolov4: Scaling cross stage partial network, 2021.

[5] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 28. Curran Associates,
Inc., 2015.

[6] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single shot MultiBox detector. In Computer Vision – ECCV 2016, pages 21–37. Springer International Publishing, 2016.