Wireless Personal Communications
https://doi.org/10.1007/s11277-019-06294-1

Real Time Multi Object Detection for Blind Using Single Shot Multibox Detector

Adwitiya Arora[1] · **Atul Grover**[1] · **Raksha Chugh**[1] · **S. Sofana Reka**[1]

Abstract
According to world health statistics 285 million out of 7.6 billion population suffers visual impairment; hence 4 out of 100 people are blind. Absence of vision restricts the mobility of a person to pronounced extent and hence there is a need to build an explicit device to conquer guiding aid to the prospect. This paper proposes to build a prototype that performs real time object detection using image segmentation and deep neural network. Further the object, its position with respect to the person and accuracy of detection is prompted through speech stimulus to the blind person. The accuracy of detection is also prompted to the device holder. This work uses a combination of single-shot multibox detection framework with mobileNet architecture to build rapid real time multi object detection for a compact, portable and minimal response time device construction.

Keywords Deep neural network · Single shot multibox detector · Text to speech · Mean average precision · Rectified linear unit

1 Introduction

As we are advancing with generations, technology is accelerating and reaching out to pervade the globe. A large portion of the population suffers the disability of visual impairment which restricts their mobility to a great extent. Hence there is grave need for a device that aids their mobility. There are devices existing in the category of ETA (electronic travel aid) but shortcomings of each of these primitives have paved way for progressive and precise automation.

Traditionally blind person used a guide dog or a walking stick for moving around. During 1920s white cane was designed for aid of blind as this stick was foldable and color coded as white to distinguish it from other sticks and indicate their presence [1]. The canes helped the person to measure and scan the circumference for obstacles. Smart Canes were introduced with ultrasonic sensors which measures the distance of obstacle by measuring the time difference between the dispatched signal and received signal to alert the stick bearer [2]. In [3] authors have modified the white cane by combining

✉ S. Sofana Reka
chocos.sofana@gmail.com; sofanareka.s@vit.ac.in

[1] School of Electronics Engineering, VIT University, Chennai Campus, Chennai, Tamilnadu, India

Published online: 30 March 2019

vibration sensors to the handle of the cane and ultrasonic sensors for detection of low hanging obstacles. In 2008, authors of [4] used RFID Radio Frequency Identification Technology for indoor navigation with all major objects having RFID tags and reader held by the white cane [5]. In 2010 smart electronic white cane was introduced for indoor and outdoor navigation which used RFID—Radio Frequency identification and GIS—geographic information with RFID reader fed in the white cane and further guide's blind person [5].

In 2013, Zigbee based Bus alert system was introduced for blind aid in which zigbee units are installed in the buses and held by the passenger and by forming a huge wireless sensor network the information is conveyed to the passenger through a voice synthesizer [6]. In 2013 device with vibrotactile technology was experimented that resembled the one on mobile devices; which was used to alert the bearer about the obstacles through the stimulation of skin receptors [7]. One of the ETAs that was extensively used was Laser cane that uses acoustic signals to alert the person, here acoustic signal produced is corresponding to the distance between the object detected and the user [8].

With increased use of smart phones a combination of obstacle detector cane, GPS in mobile phone for navigation and earphones for speech signal transmission to the blind was used [8]. The major problem with devices using cane was the field view range encountered; obstacles above chest could not be detected and the angle of field view was also close to 62°. In [9], the authors have built a smart walker which resembles the structure of an original walker and has an extension for computational purpose. It uses two vibrotactile motors on the handles and two lasers for distance and height measurement of the object and both are connected through a vibrotactile feedback generator. In [10] the author has compared all the technology on basis of 14 versatile features of a wearable ETA and exhibited drawbacks of all technologies till date.

Vibrotactile sensing lacks accuracy of detection as it can detect but not identify an object for the person. Also this technology needs training for the user to use it duly. Ultrasonic and laser sensing have helped object detection but it is more desirable to advance to a technology that is close to a virtual eye that can capture and process the images in real time. With the age of computer vision, artificial intelligence and machine translation where neural networks has played an extensive role, image processing is also one of the fields that is advancing by means of deep neural networks in far reaching applications like robotics, image compression, character recognition etc.

Artificial Intelligence also known as Machine Intelligence can be understood as the effort of humans to design a machine that works as human brain to resemble intelligent human nature. AI is achieved by studying the way human brain thinks, learns, works, recognizes objects and solves problems in day to day life. It has proved to be useful in field of marketing, hospital and medicines, finance, education etc. Machine learning can be considered as the part of AI which makes computers capable of learning without being programmed explicitly [11–13]. Computer algorithms improve automatically through experience. There are mainly three types of learning in ML: Supervised Learning, Unsupervised Learning and Reinforcement Learning. Supervised learning involves a teacher that supervises the algorithm while the system is learning. It basically involves input–output mapping. Teacher feeds in a certain set of input and defines its output.

Algorithm gets trained with the input and the output data provided [14, 15]. Unsupervised learning algorithm learns by finding similarities and differences in the input data. It categorizes the data into different classes depending upon the dataset [16]. Reinforcement learning [17, 18] on the other hand can be considered as a mixture of above two learning methodologies. Analogy to this type of learning is training a pet. A dog is supposed

to bring back a thrown ball. If it brings back the ball, owner rewards it. But if dog brings something else instead of ball, owner punishes it. Dog understands the output it is expected to produce.

By the above explanation, it can be concluded that ML requires a large amount to data to train the model. Also ML has proved to be inefficient in solving certain critical AI problems. The problems connected with Machine Learning have paved way for Deep learning. Deep learning algorithms are capable of selecting right features for themselves without much guidance from the programmer. Deep Learning is a part of Machine Learning that is focused on programming the computer to function as the complex and large number of parallel interconnections of neurons present in human brain using artificial neural networks [19, 20]. There are about 10 billion neurons and 60 trillion interconnections. Making such complex computer is not possible; researchers are working on the most fundamental part which is neuron.

ANN has been successful in mimicking biological neuron in number of ways. Nonlinear nature is amongst the most important feature. As stated above, it has been able to learn through different ways of input–output mapping. After the training is done, output of the testing might vary from the actual data but ANN allows self-learning enabling different parameters to adapt to the environment. We tend to produce evidences about the statements we make based on past experiences. Similarly to humans, ANN makes decisions and produces confidence level for the same [21, 22]. Confidence level reflects how confidently the algorithm produces certain result. There can be loss in connection between neurons, or a neuron can become weak to transmit proper electrical signal through axon, biological neurons have ability to overcome such faults. This property is known as fault tolerance and can be seen in ANN also. All these things have been realized only because Artificial Neural Networks are VLSI implementable. Deep learning has led to many inventions and research is going on for further inventions. To name a few significant applications self-driving car, automatic machine translation, and automatic image caption generation can be considered.

There is a part in our brain called Visual Cortex. This part is sensitive to specific regions of visual field and became the motivation behind Convolutional Neural Network [23]. CNN is a feed-forward Artificial Neural network [24, 25]. When image is input to a computer, different channels are responsible for detecting different information. There are 3 channels: Red, Green and Blue having their own pixel values. This is the case for colored images. CNN basically has the following layers: Convolution, ReLU, Pooling and then a fully connected layer. CNN extracts some features through the training data and compares these features with the input image at the time of testing for its classification.

There are three object detection methods in deep learning based object detection—RCNN, YOLO and SSD. RCNN is a region based convolutional neural network which uses Region Proposal algorithm for detecting object location. Region Proposal Network (RPN) is a convolutional network that predicts boundary of the object and detects the object with confidence scores. Faster RCNN combines this module with Fast RCNN detector that uses these proposed regions [26]. RPNs can efficiently and accurately generate region proposal. It involves four basic steps of object detection: generation of potential bounding boxes, feature extraction, classification, and post processing to refine the bounding boxes.

You Only Look Once (YOLO) is an algorithm in which we process the image only once to predict what objects are present and their location. Prediction of multiple bounding boxes and class probabilities for those boxes is done using single convolutional network [27]. This works by dividing the image into grids. The grid cell which is at the center of object is responsible for its detection. SSD is simpler as compared to methods that require object proposals. It eliminates the stages of proposal generation and subsequent pixel or

feature resampling. In this method, the output space of bounding boxes is discretized into a set of default boxes over different aspect ratios and scales per feature map location. It encapsulates all computation in a single network [28, 29].

As compared to YOLO, SSD is much faster and also as accurate as slower techniques that perform region proposals are pooling. So it is suitable for resource constrained devices regarding size, memory and graphical processing speed. Here dual parameters are considered—Precision and time latency.

In this paper we propose a method for object detection using aforementioned Technology. More specifically, convolutional neural network is used which captures the context information from feature maps for improved object detection accuracy [30]. Figure 1 shows the prototype design.

2 Proposed Methodology

This project combines Single Shot Detection and MobileNet for detection. SSD is much faster and accurate as compared to other methods. MobileNet is taken as the base network. Base network can be considered as the standard layers of predesigned architecture used at the early stages of Network model for improvised image classification. MobileNet is suitable for devices like Raspberry pi, Smartphones, etc. COCO dataset was used to train MobileNet SSD which was further improved using PASCAL VOC. Combination of SSD and MobileNet has successfully predicted the bounding boxes around objects, category of objects and their scores or the confidence.

When an image is input to a computer, the 3 channels (Red, Green, Blue) undergo various stages before producing the output. CNN is used in contrast to fully connected network because it causes less computation. In fully connected network all the nodes are connected

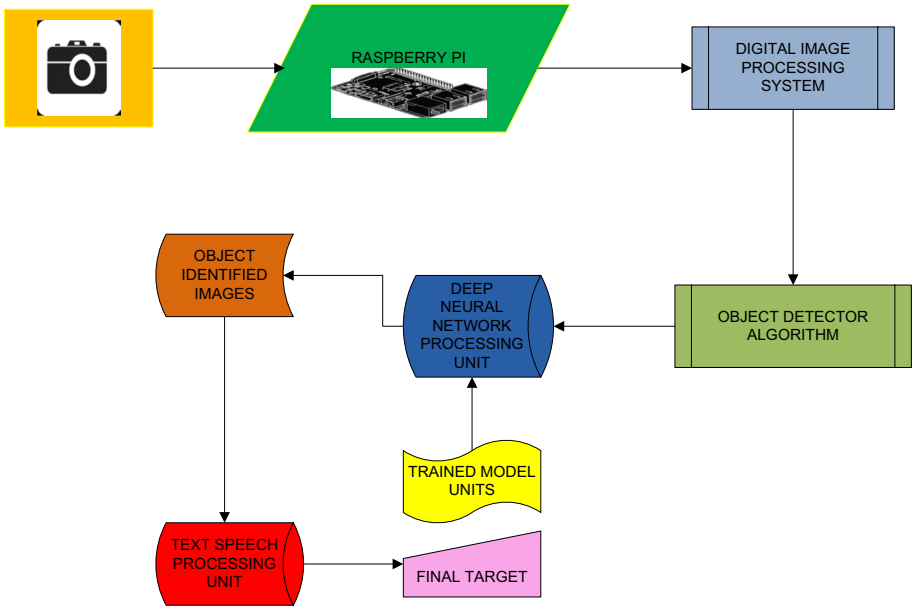


Fig. 1 Prototype design

to each other which is not the case with CNN. Computer understands an image by the value at each pixel. CNN compares the input image in patches or in pieces and compares it with features, also called filters at similar positions for classification.

At the end of base network, convolutional feature layers are added. Features are then put on the input images and if it matches, image is classified into one of the classes. At the convolution layer, feature or filters are lined up on the input image and pixel by pixel multiplication is done. All these pixel values are then added and divided by the total number of pixels in the feature. A map is created and value obtained above is put in the map. This filter is moved throughout the image and values at each pixel are obtained and recorded in the matrix form. These steps are followed for all the feature maps or filter used in the network. Therefore after passing through the convolution layer, there will be feature maps equal to the number of filters used.

At the ReLU layer, which can be considered as the activation function. These functions produce an output which is linear to the input only when the input value is above a certain predefined threshold, otherwise the output is zero. So ReLU layer will basically replaces the values that are negative or below the threshold with zero as per the output of convolution layer. This is done for all the features. Pooling layer is used to reduce or shrink the size of image. Maxpool takes the maximum value from a matrix depending upon the window size and stride. Window is moved throughout the image and a new matrix which is smaller in size is obtained. After the pooling layer, matrix equal to the number of filters is passed onto the next stage. This reduced size data is passed through the same combination of convolution, ReLU and pooling layer again and again to further reduce the size of images. These images after passing through different stages of above mentioned three layers are now stacked together in the form of a list. This step takes place in fully connected layer. Depending upon the values in the list, fully connected layer classifies input image.

MobileNet uses depth wise separable convolution which was introduced in [7]. This is used in cases where computational time has to be minimized like for self-driving cars, real-time object detection, etc. It divides the convolution into 3×3 depth wise convolution followed by 1×1 point wise convolutions. Filtering and combining is done by two separate layers. In the first layer, that is depthwise convolution, single filter is applied per each input channel and 1×1 convolution is done on the output from first layer to create linear combination of that output. 1×1 convolution is computationally cheap as it mainly involves matrix multiplication. Both these layers are followed by Batch norm and ReLU nonlinearity as mentioned in Fig. 2.

3 Algorithm Framework

This assistant is an alert system that captures the surrounding view of the blind person and processes it in real time with frame rate of 60 FPS (frames per second) to detect the objects and guide the subject accordingly. The project uses CaffeNet image model for extracting pre-trained object classes. Pre trained models are available as open source links that are trained for around 1 million objects and can be modified and trained further by the user. First snapshots are captures from the video streaming and path is defined to the snapshot and the pre trained caffe model. Initialize a threshold level of confidence of detection to discard erroneous and weak detected objects. Import the object classes for recognition in mobilenet SSD. From RGB scale of 0–255, random bounding boxes are defined uniformly

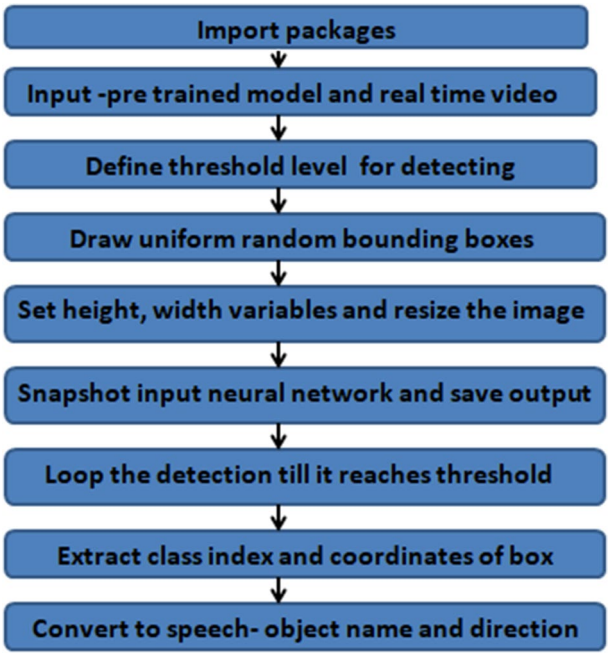


Fig. 2 Algorithm flow and analysis

between the scale according to the values in the input image. Width, height and boxes are taken as input variables. The image is resized according to computational capability of the hardware used. For example, this project used resizing standard as 250 * 250 pixels.

Before the image goes into the neural network it has to be pre processed to meet some standards and to reduce computational complexity. This pre processing includes regulating the illumination changes in the snapshots extracted from the video streaming. For this we need to do three important operations

1. Mean subtraction

The mean of Red, Green and Blue pixel values each are subtracted from the original image to obtain a mean subtracted image. This normalizes and combats the brightness changes in the image.

2. Scaling factor

To obtain a image that is suitable for detection we need to scale the pixel values by a scaling factor. This step is done to compensate for reduced pixel values obtained in the previous step.

$$R = (R_o - \mu_R)/\delta$$

$$G = (G_o - \mu_G)/\delta$$

$$B = (B_o - \mu_B)/\delta$$

Here R, G and B are the resultant red, blue and green pixel values respectively. R_o, $G_{o\ and}$ B_o are original image red, green and blue pixel values respectively. δ is the scaling

factor here, this scaling factor is calculated from the standard deviation of all the RGB values in the pre trained images of the extracted model.

3. Channel swapping (optional)

This step is taken to compensate the uneven Red blue green values obtained from the first two steps hence we can swap red and blue or red and green or blue or green channels to obtain a better image for further steps. We generally swap red and blue channels for optimum results.

The defined packages for deep neural networks are loaded as they are available in python, the working of these algorithm have been discussed in the previous section. The image is fed forward through the neural network and the forward image is saved as a variable. A loop is created while passing image through the network in iterations till the bounding box completely encloses a detected object with the defined threshold level, for example—92% accuracy. When the image is fed to the network, probabilities are defined for each class trained and extracted; the object class with the largest probability is taken. When the bounding box meets the threshold level, the object class label is extracted from the model and the dimensions of the bounding box are calculated and defined as output variables.

The extracted xy coordinates and the label class name are used to define the direction of the object with respect to the person. After the direction and the object name is defined. The information is sent through a text to speech conversion and the output is fed through speech signals to the earphones connected. This algorithm has to work for at least a 60 FPS accurate speed to be accepted for a real time use. Figure 2 shows the algorithm in the form of flowchart.

4 Results and Discussion

Figure 2 shows the device and its components, consisted of a camera, cap (to hold camera), earphones and Raspberry pi 3(small sized computer). Device is light weight and wearable. The idea is that blind person focuses on objects that are in front and close to him/her. This project gives the pleasure of not only object detection but also tells the category of object through voice (Fig. 3).

1. Camera module fixed in cap
2. Earphones
3. Raspberry pi

In the example below, a person is detected and recognized with 98.95% confidence. In second example, a dog is detected and recognized with 98.75% confidence. Device works fine even in dim light (Fig. 4).

Figure 5 shows the ability of deep learning to detect and localize multiple objects which is generally complicated. In this example two persons are detected with 97.80% and 97.04% accuracy. This technology paves way for a novel use of fields like deep learning, image processing and text to speech. It improves accuracy and latency of detection and hence can be used for real time operations. If manufactured as a product, this technology can assist blind people to aid their mobility at an affordable price and can match with future technology as this field is still in view of research and evolving.

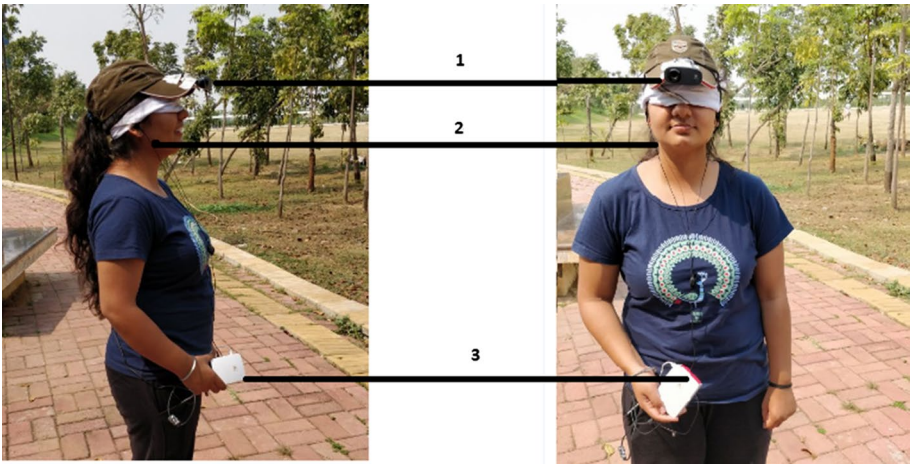
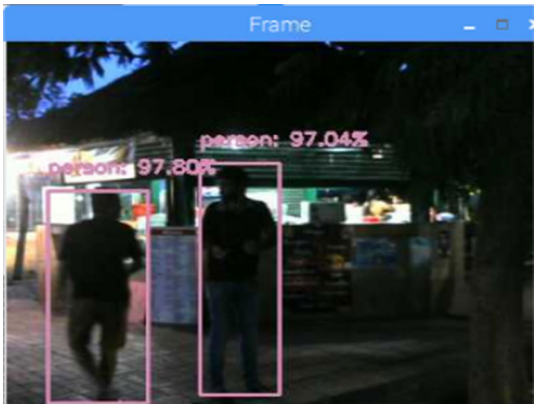


Fig. 3 Device and its components

Fig. 4 Practical analysis of the prototype developed



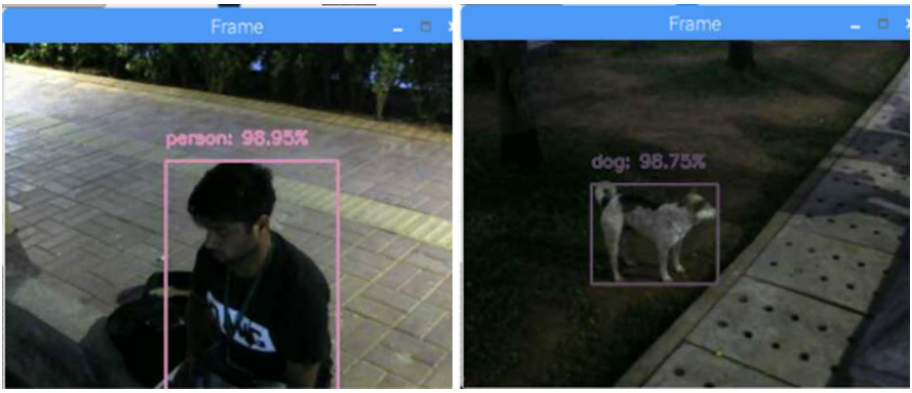


Fig. 5 Practical analysis of the prototype developed with human and animal detection

5 Conclusion

This technology paves way for a novel use of fields like deep learning, image processing and text to speech. It improves accuracy and latency of detection and hence can be used for real time operations. If manufactured as a product, this technology can assist blind people to aid their mobility at an affordable price and can match with future technology as this field is still in view of research and evolving. As going on with future work, given that we get high computing compatible computers with advanced graphic processors we can reduce the latency of training and run time to 10 times. We can use high speed complex algorithms for increasing accuracy. Also, we can use more number of cameras to increase field view. This work can further be evolved for face recognition to learn familiar faces encountered by the blind person.

References

1. http://www.acb.org/.
2. Shoval, S., Ulrich, I., & Borenstein, J. (2003). NavBelt and the Guide-Cane [obstacle-avoidance systems for the blind and visually impaired. *IEEE Robotics and Automation Magazine, 10*(1), 9–20.
3. Wang, Y., & Kuchenbecker, K. J. (2012). HALO: Haptic alerts for low-hanging obstacles in white cane navigation. In *2012 IEEE haptics symposium (HAPTICS), Vancouver* (pp. 527–532).
4. Chumkamon, S., Tuvaphanthaphiphat, P., & Keeratiwintakorn, P. (2008). A blind navigation system using RFID for indoor environments. In *2008 5th International conference on electrical engineering/ electronics, computer, telecommunications and information technology, Krabi* (pp. 765–768).
5. Faria, J., Lopes, S., Fernandes, H., Martins, P., & Barroso, J. (2010). Electronic white cane for blind people navigation assistance. In *2010 World automation congress, Kobe* (pp. 1–7).
6. Lavanya, G., Preethy, W., Shameem, A., & Sushmitha, R. (2013). Passenger BUS alert system for easy navigation of blind. In *2013 international conference on circuits, power and computing technologies (ICCPCT), Nagercoil* (pp. 798–802).
7. Adame, M. R., Yu, J., Moller, K., & Seemann, E. (2013). A wearable navigation aid for blind people using a vibrotactile information transfer system. In *2013 ICME international conference on complex medical engineering, Beijing* (pp. 13–18).
8. Ando, B. (2003). Electronic sensory systems for the visually impaired. *IEEE Instrumentation and Measurement Magazine, 6*(2), 62–67.
9. Wachaja, A., Agarwal, P., Zink, M., Adame, M. R., Möller, K., & Burgard, W. (2015). Navigating blind people with a smart walker. In *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS), Hamburg* (pp. 6014–6019).
10. Dakopoulos, D., & Bourbakis, N. G. (2010). Wearable obstacle avoidance electronic travel aids for blind: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40*(1), 25–35.
11. Balasuriya, B. K., Lokuhettiarachchi, N. P., Ranasinghe, A. R. M. D. N., Shiwantha, K. D. C., & Jayawardena, C. (2017). Learning platform for visually impaired children through artificial intelligence and computer vision. In *2017 11th International conference on software, knowledge, information management and applications (SKIMA), Malabe, Sri Lanka* (pp. 1–7).
12. Mancini, A, Frontoni, E, & Zingaretti, P. (2018). Mechatronic system to help visually impaired users during walking and running. *IEEE Transactions on Intelligent Transportation Systems, 19*, 649–660. ISSN 1524-9050.
13. Dunai, L. D., Lengua, I. L., Tortajada, I., & Simon, F. B. (2014) Obstacle detectors for visually impaired people. In *2014 International conference on optimization of electrical and electronic equipment (OPTIM), Bran* (pp. 809–816).
14. Xiong, J. (2018). Tutorial-1: Machine learning and deep learning. In *2018 23rd Asia and South Pacific design automation conference (ASP-DAC), Jeju, Korea (South)* (pp. 19–25).
15. Noble, F. K. (2017) A mobile robot platform for supervised machine learning applications. In *2017 24th International conference on mechatronics and machine vision in practice (M2VIP), Auckland* (pp. 1–6).
16. Barbosa, C., Santana, O., & Silva, B. (2017). An unsupervised machine learning algorithm for visual target identification in the context of a robotics competition. In *2017 Latin American robotics symposium (LARS) and 2017 Brazilian symposium on robotics (SBR), Curitiba* (pp. 1–6).

17. DiStasio, M. M., Francis, J. T., & Boraud, T. (2013). Use of frontal lobe hemodynamics as reinforcement signals to an adaptive controller. *PLoS ONE, 8*, e69541. ISSN 1932-6203.
18. Chhatbar, P. Y., Francis, J. T., Fridman, E. A. (2013). Towards a naturalistic brain–machine interface: Hybrid torque and position control allows generalization to novel dynamics. *PLoS ONE, 8*, e52286. ISSN 1932-6203.
19. Moshovos, et al. (2018). Value-based deep-learning acceleration. *IEEE Micro, 38*(1), 41–55.
20. Ranganathan, H., Venkateswara, H., Chakraborty, S., & Panchanathan, S. (2017). Deep active learning for image classification. In *2017 IEEE international conference on image processing (ICIP), Beijing, China* (pp. 3934–3938).
21. da Silva, L. C. B., de Oliveira Rocha, H. R., Castellani, C. E. S., Segatto, M. E. V., & Pontes, M. J. (2017) Improving temperature resolution of distributee temperature sensor using Artificial Neural Network. In *Microwave and optoelectronics conference (IMOC) 2017 SBMO/IEEE MTT-S international* (pp. 1–5).
22. Han, W. S., & Han, I. S. (2017). Bio-inspired neuromorphic visual processing with neural networks for cyclist detection in vehicle's blind spot and segmentation in medical CT images. In *2017 Computing conference, London* (pp. 744–750).
23. Yang, H., Yuan, C., Xing, J., & Hu, W. (2017). SCNN: Sequential convolutional neural network for human action recognition in videos. In 2*017 IEEE international conference on image processing (ICIP), Beijing, China* (pp. 355–359).
24. Deng, Z., Fan, H., Xie, F., Cui, Y., & Liu, J. (2017). Segmentation of dermoscopy images based on fully convolutional neural network. In *2017 IEEE international conference on image processing (ICIP), Beijing, China* (pp. 1732–1736).
25. Cho, C., Lee, Y. H., & Lee, S., (2017). Prostate detection and segmentation based on convolutional neural network and topological derivative. In *2017 IEEE international conference on image processing (ICIP), Beijing, China* (pp 3071–3074).
26. Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(6), 1137–1149.
27. Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *2016 IEEE conference on computer vision and pattern recognition (CVPR), Las Vegas, NV* (pp. 779–788).
28. Liu, W., et al. (2016). SSD: Single shot MultiBox detector. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), Computer vision—ECCV 2016. ECCV 2016.
29. Ning, C., Zhou, H., Song, Y., & Tang, J. (2017). Inception single shot MultiBox detector for object detection. In *2017 IEEE international conference on multimedia & expo workshops (ICMEW), Hong Kong* (pp. 549–554).
30. Cengil, E., Çınar, A., & Özbay, E. (2017). Image classification with caffe deep learning framework. In *2017 International conference on computer science and engineering (UBMK), Antalya* (pp. 440–444).

Adwitiya Arora is pursuing her B.Tech. degree from VIT University, Chennai in the field of Electronics and Communication Engineering. She is from Delhi. Her area of interests are Image Processing, Computer Vision and Embedded systems.

Atul Grover is an under graduate student studying in VIT Chennai. He is from Ludhiana, Punjab. His field of interest is embedded systems, IoT and works in multidisplicanry projects.

Raksha Chugh is pursuing her B.Tech. degree in Electronics and Communication from VIT University, Chennai. She comes from Jaipur, Rajasthan. Her areas of interest are Networking, Neural networks and Data Analysis.

S. Sofana Reka completed her B.E. (Electrical and Electronics Engineering) from Vellore Institute of Technology (formerly Vellore Engineering College), Vellore, and her M. Tech. from SASTRA University, Thanjavur, India, Ph.D. from VIT University Vellore. She is currently working as Assitant Professor at School of Electronics, VIT University, Chennai campus. Areas of interest are Power Systems and Soft computing Techniques.