Robust Object Detection and Recognition for the Visually Impaired

Naman Patel[2], Yubo Pan[1,2], Yuanda Li[1,2], Farshad Khorrami[2],
John-Ross Rizzo[4], Todd Hudson[4], Edward K. Wong[1,2], Yi Fang[1,2,3,*]
[1] NYU Multimedia and Visual Computing Lab
[2] NYU Tandon School of Engineering
[3] NYU Abu Dhabi
[4] NYU Langone Medical Center

***Abstract*—An Electronic Travel Aid (ETA) capable of providing real-time situational awareness is developed as a potential tool for the visually impaired. A robust ETA may enable those with visual impairment the ability to navigate more safely in three-dimensional space. A unified deep neural network-based approach is proposed for detecting and localizing objects in the scene-views of a stereoscopic camera system deployed in the ETA, subserving the goal of spatial comprehension for the immediate environment of the visually impaired end-user. Object detections by the deep neural networks are concurrently combined with their corresponding depths to localize the potential hazards in three-dimensions. The network architecture for object recognition is learned end-to-end by optimizing a multi-task loss comprised of four object proposal sub-networks at multiple scales and an object detection network; the depth is evaluated from the disparity map obtained from the stereoscopic cameras. Our device is capable of detecting multiple obstacles in real time in the vicinity of an end-user by modality augmentation in the RGB and depth domains. This ETA has significant potential in its application to those afflicted with sight loss, supplementing existing tools it its core functionality or enhancing meta-modal sensor fusion approaches.**

I. Introduction

The World Health Organization (WHO) estimates there are 285 million suffering from visual impairment worldwide (39 million blind, 246 million low vision) [1]. Blindness and low vision result in a host of social, emotional and health-related problems, often due to antecedent difficulties with mobility, as it becomes progressively more difficulty to maintain pre-morbid activities [2], [3]. This decreased mobility engenders reduced rates of employment and overall sedentary lives, generating quality of life compromises. Alarmingly, visual deficiency throughout the world significantly increases morbidity and mortality [4]. The reduced mobility and reciprocal rises in illness not only engenders productivity losses, but it also leads to insurmountable healthcare costs.

While guide dogs, white canes and adaptive mobility devices [5]–[12] have been used for decades as primary mobility tools, representing the foundation of orientation and mobility training for low vision navigation, electronic travel aids (ETAs) have also been in use for years, but as secondary devices [13]–[21]. These aids have been used in conjunction with primary devices to provide supplementary information, above and beyond what the primary tool would be able to survey. ETAs may provide pertinent threat information to an end user at greater distances and within a greater field of view than a standard white cane or guide dog. While there is intrinsic value in their stated properties, previous devices have been met with limited end-user adoption. This may have stemmed from a combination of factors, but included the technological limitations at the time of device conception and fabrication.

Here we propose to develop an electronic travel aid (ETA) capable of providing real-time situational awareness. The device has the potential to remedy a myriad of the canes shortcomings, in addition to further augmenting the ability of visually impaired persons to localize and recognize objects in their environment. The ETA is comprised of a GPU-embedded computer (Jetson TX1), a stereoscopic camera system (Stereolabs ZED™ camera), and a rechargeable battery. The system architecture for the aid consists of two major tasks, obstacle detection and obstacle localization in three dimensions. Our approach is centered on a novel deep learning-based paradigm composed of an efficient real time object recognition method combined with depth computations. This ETA can operate in unstructured environments, indoors as well as outdoors and is able to detect, recognize and localize obstacles in real time.

II. Related Work

Real-time object detection is a major task in our development pipeline for the electronic travel aid, creating robust environmental awareness for the visually impaired. There have been numerous methods proposed to achieve real-time detection; more robust methods have been based on cascade detection [22]. The cascade based detection architecture has been widely used to implement sliding window detectors for cars, pedestrian and faces [23]. In order to improve the speed of the architecture methods for fast feature extraction based on histogram of gradient features and aggregate feature channels based fast detections have been proposed [24]–[26]. Also, soft cascade methods have been proposed which make an optimal tradeoff between accuracy and run-time [27]. The primary

* Corresponding Author: Yi Fang (yfang@nyu.edu), Electrical and Computer Engineering, NYU Abu Dhabi and NYU.

difficulty with using cascade based detection is that its time complexity increases with the number of classes for detection.

In order to deal with this problem, deep neural network which had succeeded in image classification [28] were utilized in object detection frameworks. The proposed networks' computational complexity didn't increase with the number of classes and had similar or better accuracies than cascade detection networks. One of the first networks to exploit deep neural networks for object detection were R-CNN networks which combined an object proposal mechanism and a CNN classifier [29]. While the R-CNN surpassed previous detectors by a large margin, its speed was limited by the need for object proposal generation and repeated CNN evaluation. Spatial pyramid pooling (SPP) based network [30] were used to alleviate this problem. Spatial pyramid pooling based network compute CNN features only once per image, thereby increasing the detection speeds considerably. The Fast-RCNN based network [31] further improved on SPPNet by introducing ROI pooling layers and multi-task loss (bounding box regression and object classification) which made all modules except object proposal network differentiable. In order to make the system learn end to end, the object proposal network was also made differentiable in faster RCNN architecture [32] which significantly increased the detection speeds. In order to make detections faster by just using the convolution features, YOLO network was proposed [33], which outputs object detections within a predefined grid. PVANet [34] were also proposed for real time object detection. This makes the network run faster and real-time at the cost of accuracy. One of the problems with detection networks are that the objects to detect are in a broad range of scales, so they have to be localized precisely. In order to alleviate this problem multipath networks were proposed [35] which consists of skip connections that give the detector access to features at multiple network layers, a foveal structure to exploit object context at multiple object resolutions, and an integral loss function and corresponding network adjustment that improve localization. We solve this problem by optimizing multiple losses at intermediate layers of a single network which produce robust object detector at each intermediate output layer [36]. The network architecture is based on MS-CNN framework [37] which has four object proposal sub-networks and a detection network.

Despite incredible scientific and academic effort, no electronic travel aid has gained widespread acceptance. A travelling aide's intrinsic value can be evaluated by its ability to help individuals restore or achieve optimized mobility, the state in which one travels with safety, comfort, grace and independence. In addition, the interface should be intuitive and it should also be easy to train the end user of the device. Adding overly complicated electronic travel aid that may require extensive and complementary training periods to procure proficiency does not represent a tenable solution, and is certainly not an approach that would be logistically sound, given the already constrained instructors who are overwhelmed with lesson plans [13], [20], [38]. Acknowledging these facts, we design an intuitive interface which allows the end user to navigate through his/her immediate surroundings without significant training using state of the art computer vision techniques for object detection and depth estimation. The device will be able to detect and recognise obstacle in the vicinity of the end user and in addition localize obstacles in three dimension by obtaining depth from disparity of stereo images in real time.

III. Method

This section presents the system architecture for object detection and localization in three dimension for aiding visually impaired. The system architecture, as shown in figure 1, broadly consists of two sub-networks, namely the object detection architecture and 3D object localization architecture.

A. Object Detection

Real-time object detection is one of the major task in our pipeline for the electronic aid for robust environment awareness for visually impaired. We use a multi-scale convolutional neural network [37] based architecture because of its relatively low time complexity and good accuracy. The architecture performs well because object proposal is done at four different scale by the network efficiently. The other advantage of this network is that everything including features, proposals and detections are learnt end to end which makes the network relatively fast. The time complexity of the network is not constrained by separate object proposals and detection networks.

1) Architecture: Our architecture is a single unified network consisting of a main trunk and four different proposal subnetworks connected to it at different scale in addition to the detection network. The network detects objects at four different scales. The network consists of two layers of 64 convolutional filters with height and width of 3 and stride of 1 followed by ReLU non-linearity, followed by max pooling of pixels in the window of height and width of 2 and strides of 2, followed by two layers of 128 convolutional filters of the same size and stride as before with non-linearity, followed by maxpooling of pixels with same size and stride as before. The size and stride of the convolutional filter and maxpooling window remains constant throughout the main trunk. The maxpooling layer is followed by three layers of 256 convolutional filters with ReLU non-linearity, followed by maxpooling layer, followed by three layers of 512 convolutional filters and ReLU nonlinearity. The last layer consisting of 512 filters are used as feature maps for the detection network and the first proposal sub-network. Also the three layers of 512 convolutional filters are followed by max pooling layer followed by three more convolutional layers with 512 filters with stride of 1. The output feature maps of the last convolutional layer are used as feature maps for the second proposal sub network. The last layer of 512 convolutional filters are maxpooled and passed through a layer of 512 convolutional filters with ReLU nonlinearity. These convolutional maps are used as features for the third proposal sub-network. These feature maps are also maxpooled and passed through a layer of 512 convolutional

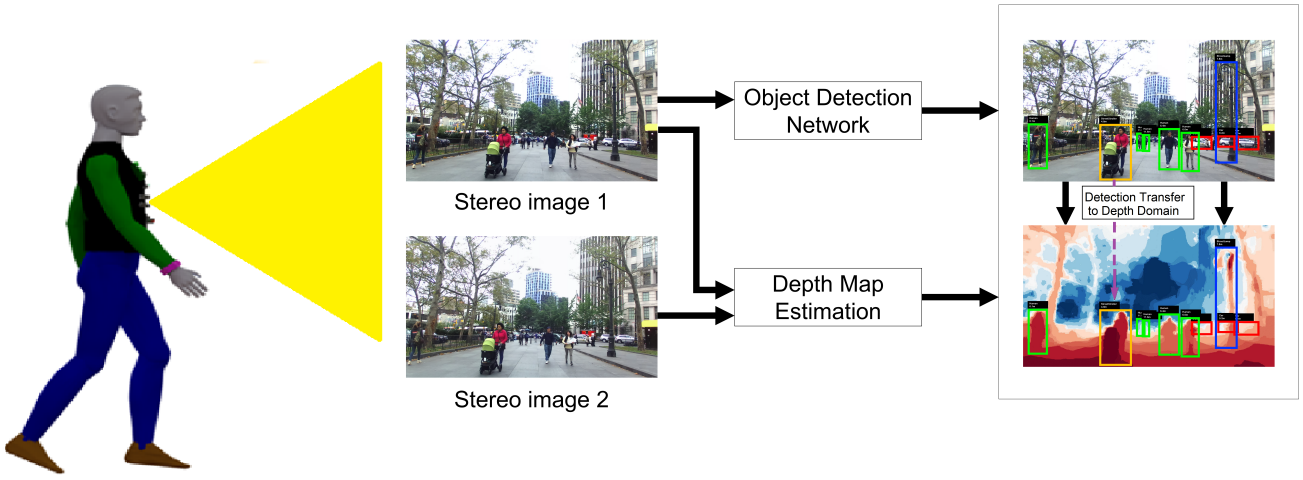

Fig. 1. Stereo images obtained from the stereo camera mounted on the end user's vest are used to estimate the depth and detect objects in the vicinity of the end user. Detection is transferred from RGB domain to depth domain to localize the detected object in three dimensions.

maps. These convolutional maps are used as features for the fourth proposal sub-network.

The region proposal network outputs a set of rectangular object proposals with their objectness score (class probability).The idea behind using multiple proposal network at different scale is that object proposals at multiple receptive fields are obtained by the network. The first proposal sub-network consists of a convolution layer (since the sub-network is close to lower layers of convolution of the main trunk which may cause instability due to the sub-network being closest to the main trunk), followed by a detection layer which maps the convolution map into a fixed feature vector. These feature vector are fed to two fully connected networks which outputs class of the detection and the proposal bounding box coordinates. The three other proposal sub-network have similar detection layers which output the class of the detection and its corresponding bounding box coordinates. The proposals are given by the sliding window paradigm. The proposal network itself can be used as a detection network but does not perform as well. Thus, a separate detection network is introduced.

The object detection sub-network consists of a deconvolution layer in order to increase the resolution of the feature map. The deconvolution layer upsamples the features from the Conv4-3 layer as shown in Figure 2. Feature upsampling does not increase the memory and time complexity of the network significantly but rather boosts detection performance of small objects. The outputs from the proposal sub-networks are used to perform region of interest pooling with context for the detection subnetwork. Context has been shown useful for object detection. We take the context nearby the proposed object into account and evaluate the features of the anchors obtained from deconvolved feature maps. The features are obtained from anchor at two different scale and concatenated as shown in the figure 2. The context anchor is 1.5 times larger that the proposal network anchor. These concatenated features are passed through a convolution network to halve the number of feature maps and reduce the number of parameters. These feature maps are passed through a fully connected network which outputs the class probability and bounding box coordinates for the corresponding region of interest.

2) Implementation of Detection in RGB Domain: In order to optimize the network and learn the parameters of the unified object detection network, the network is trained in two stages. In the first stage only the object proposal network is trained and in the second stage object proposal network and the detection network are jointly trained. The reason for multi-stage training is that training the unified network end to end will make it unstable in the early iterations.

In the first stage, we use a network [39] pre-trained on Imagenet dataset [40] data and fine tune it by optimizing the parameters using the joint multi-task classification and regression loss of the object proposal network. The network is fine-tuned for object proposals using the sliding window paradigm. In the sliding window paradigm, an anchor is centered at the sliding window on the corresponding layer at the given scale. The anchor is scaled according to the corresponding filter height and width. The anchor bounding box in the sliding window is labeled as positive if the intersection of union with the ground truth bounding box for the object is greater than equal to 0.5 and put in the negative pool if its less than 0.2. In order to deal with the asymmetry of positive and negative samples, the negative samples are randomly sampled according to a uniform distribution such that there are ten

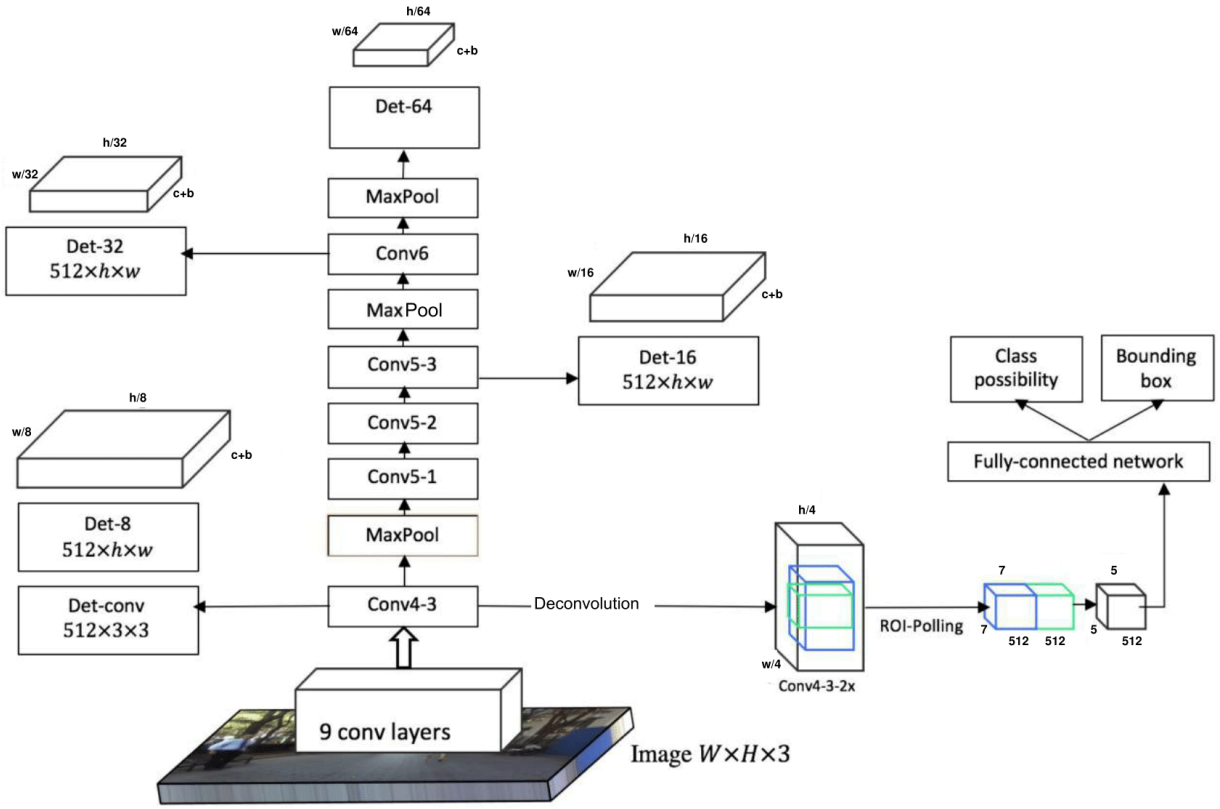

Fig. 2. Network architecture for object detection in RGB domain. W and H are the width and height of the input image. There are four outputs(cubes) for the four object proposal sub-network where c are the number of classes and b the number of bounding box coordinates, h and w are the filter height and width. The object detection network outputs the class probability and the bounding box coordinates for each detected object. The green cube represents object ROI pooling and the blue cube represents context ROI pooling.

negative samples for one positive sample. Other than random sampling strategy, we using a bootstrapping sampling strategy where only the hard negatives are samples and given priority. There are cases where there are no positive samples available for a detection layer, which can make the network unstable. In order to deal with this problem, the cross entropies of positive and negative samples are weighted as according to a ratio of maximum negative samples to maximum positive samples with logistic regression loss. Non maximum suppression is applied to reduce the number of proposal regions for detection.

The loss in the first stage is a multi task loss consisting of losses at different scales. These losses at different scale are joint classification and regression losses. The parameters of the network are updated according to the way each loss affects each parameter. The classification loss used for the proposal network is the log softmax with negative log likelihood criterion and the box regression loss is the soft L1 loss. Data augmentation by rescaling the original image to multiple scales. Due to the large resolution of the image and limited memory, small batches are taken during training and a random crop of 448x448 around objects from the whole image is taken. We train the network with stochastic gradient descent with momentum and weight decay.

In the second stage, the model trained with proposal network multi-task loss is used. The sampling of negative examples are performed by bootstrapped sampling. The optimal parameters are learnt through back-propagating with the unified network including the detection sub-network. The detection sub-network also has a joint classification and bounding box regression loss similar to the proposal sub-networks. In the detection network the ROI pooling on the deconvolution layer is performed according to the regions proposed by the region proposal network. The network is trained with context embedding. Thus, this is our paradigm for training our object detection network.

B. Mapping from RGB Domain to Depth Domain

The positions of object detected by the deep learning based network are known in the camera's two dimensional frame of reference. These positions are transformed into three dimension by using the depth information from the disparity map of the stereo camera. The disparity map is obtained by doing dense stereo-matching of the two stereo images obtained from our stereo camera. The three dimensional positions are used by our aid to help navigate a visual impaired person through the obstacles. If an obstacle is in a close proximity of

the user, the aid conveys the information regarding the obstacle and the distance of the obstacle from the person which helps the person avoid it and safely navigate through his immediate surroundings. Thus, the electronic travel aid is helpful as it conveys more information with a greater field of view to the user in comparison to traditional methods like canes and dogs. The detections from the RGB based object detection network are transferred to the depth domain by finding the corresponding RGB and depth pixel. The depth of the pixel nearest to the end user in the bounding box is reported as the depth of the detected object and used by the aid for immediate surrounding navigation.

Detection and localization is just one of the aspects of the aid, the other important task of the aid is to convey information to the visually impaired in an intuitive manner. We propose to display the surroundings in the vicinity of the user through a belt via a haptic interface in a spatiotopically preserved, intuitive, body-centered fashion. The obstacles on the users immediate left are mapped, processed and re-displayed through vibrating actuators on the left aspect of the belt. Hazards that are shorter, as opposed to taller, vibrate fewer actuators in one column of actuators in the two-dimensional matrix and obstacles that are closer to the end user are communicated through a higher frequency of vibration of the actuator element, displaying depth or the z-axis, giving one then sense of tactile looming or approach. An important aspect of this interface is that the mapping, offering deconstructed views of the hazards in ones peri-personal space, is robust without significant computational and/or processing delays. Thus, pixelated, obstacle view of the immediate environment is conveyed in near real time to the end user.

IV. Experiment

We evaluate the performance of our architecture in indoor as well as outdoor environment using ZED^{TM} stereo camera and processing on Jetson TX1. The RGB images are captured from ZED^{TM} stereo camera and fed to the deep neural network to get the detections in real time. The depth map of the corresponding scene is also evaluated simultaneously using disparity map calculated by stereo matching of the two stereo images. The detected objects in the RGB domain are mapped in the depth domain to obtain the distance of the obstacle from the sensor in three dimensions. As shown in Fig. 3 and 4, the experiments are done in real time on outdoor environments like streets and indoor environments like a college campus. Accurate detection at different scales are obtained by the network and mapped in the depth domain as shown in the figures. The range of the depth map goes from 70cm to 20m. In the figure, the darker shades of red correspond to the objects nearer to the aid and darker shades of blue correspond to the objects which are far from the aid. Our object detection network as seen in the figures 3 and 4, is able to recognize objects like car, person, monitor, chair and trashcan.

Fig. 3. Outdoor object detection in RGB domain transferred to depth domain

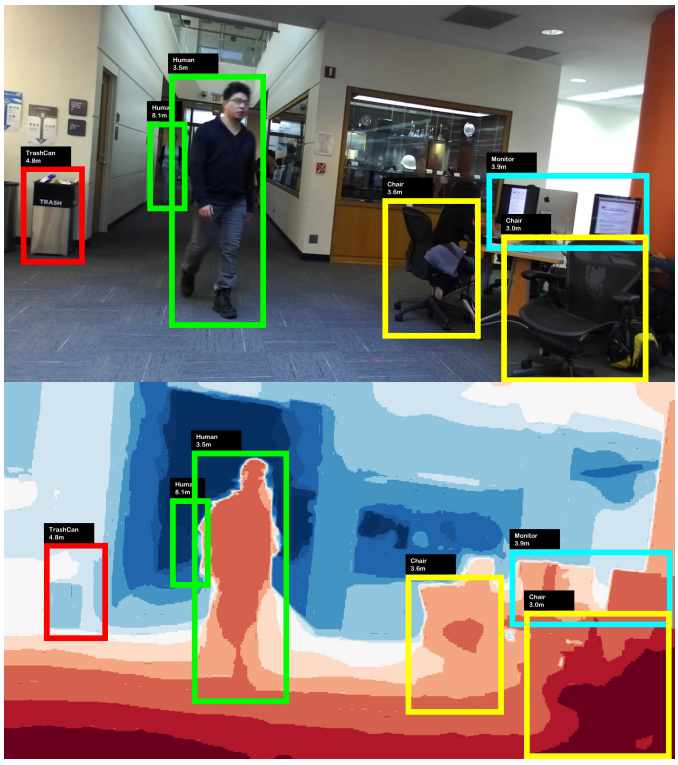

Fig. 4. Indoor object detection in RGB domain transferred to depth domain

V. Conclusion

We have presented a ETA-like system capable of performing accurate real time multi-obstacle detection and localization

in the immediate surroundings of an end user. This ETA has significant potential in its application to those afflicted with visual impairment, supplementing existing tools it its core functionality or enhancing meta-modal sensor fusion approaches.

Our future work will be focused on including the depth modality in our detection network to further improve our system. We also plan to integrate our real-time detection architecture with an intuitive interface that would assist end users with decision making during navigation, improving the safety and efficiency of travel in dynamic environments.

ACKNOWLEDGMENT

The work was supported by New York University Abu Dhabi Institute for providing the funding (AD131 and RE131) to support the research reported in this paper.

REFERENCES

[1] W. H. Organization *et al.*, "Visual impairment and blindness. who fact sheet, 282. geneva: Author," 2010.

[2] J. R. Evans, A. E. Fletcher, and R. P. Wormald, "Depression and anxiety in visually impaired older people," *Ophthalmology*, vol. 114, no. 2, pp. 283–288, 2007.

[3] J. M. Crewe, N. Morlet, W. H. Morgan, K. Spilsbury, A. S. Mukhtar, A. Clark, and J. B. Semmens, "Mortality and hospital morbidity of working-age blind," *British Journal of Ophthalmology*, vol. 97, no. 12, pp. 1579–1585, 2013.

[4] R. G. Siantar, C.-Y. Cheng, C. M. G. Cheung, E. L. Lamoureux, P. G. Ong, K. Y. Chow, P. Mitchell, T. Aung, T.-Y. Wong, and C. Y. Cheung, "Impact of visual impairment and eye diseases on mortality: the singapore malay eye study (simes)," *Scientific reports*, vol. 5, 2015.

[5] P. Putnam, *Love in the Lead: The Miracle of the Seeing Eye Dog*. University Press of Amer, 1997.

[6] A. Chur-Hansen, L.-K. Werner, C. E. McGuiness, and S. Hazel, "The experience of being a guide dog puppy raiser volunteer: a longitudinal qualitative collective case study," *Animals*, vol. 5, no. 1, pp. 1–12, 2014.

[7] B. Hoyle, "The batcane–mobility aid for the vision impaired and the blind," in *IEE Symposium on Assistive Technology*, 2003, pp. 18–22.

[8] L. Kay, "Ultrasonic spectacles for the blind," *Journal of the Acoustical Society of America*, vol. 40, p. 1564, 1966.

[9] R. Hoover, "The cane as a travel aid," *Blindness*, pp. 353–365, 1950.

[10] D. S. Kim and R. W. Emerson, "Effect of cane technique on obstacle detection with the long cane," *Journal of visual impairment & blindness*, vol. 108, p. 335, 2014.

[11] D. S. Kim, R. W. Emerson, and A. Curtis, "Drop-off detection with the long cane: Effects of different cane techniques on performance," *Journal of visual impairment & blindness*, vol. 103, no. 9, p. 519, 2009.

[12] J. Benjamin, N. Ali, and A. Schepis, "A laser cane for the blind," in *Proceedings of the San Diego Biomedical Symposium*, vol. 12, no. 53-57, 1973.

[13] R. Farcy, R. Leroux, A. Jucha, R. Damaschini, C. Grégoire, and A. Zogaghi, "Electronic travel aids and electronic orientation aids for blind people: technical, rehabilitation and everyday life points of view," in *Conference & Workshop on Assistive Technologies for People with Vision & Hearing Impairments Technology for Inclusion*, vol. 12, 2006.

[14] H. Petrie, V. Johnson, T. Strothotte, A. Raab, S. Fritz, and R. Michel, "Mobic: Designing a travel aid for blind and elderly people," *Journal of navigation*, vol. 49, no. 01, pp. 45–52, 1996.

[15] R. Pyun, Y. Kim, P. Wespe, R. Gassert, and S. Schneller, "Advanced augmented white cane with obstacle height and distance feedback," in *Rehabilitation Robotics (ICORR), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1–6.

[16] Y. Gao, R. Chandrawanshi, A. C. Nau, and Z. T. H. Tse, "Wearable virtual white cane network for navigating people with visual impairment," *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, vol. 229, no. 9, pp. 681–688, 2015.

[17] M. Choudhury and A. Barreto, "Design of a multi-sensor sonar system for indoor range measurement as a navigational aid for the blind." *Biomedical sciences instrumentation*, vol. 39, pp. 30–35, 2002.

[18] P.-H. Cheng, "Wearable ultrasonic guiding device with white cane for the visually impaired: A preliminary verisimilitude experiment," *Assistive Technology*, no. just-accepted, 2016.

[19] H.-H. Pham, T.-L. Le, and N. Vuillerme, "Real-time obstacle detection system in indoor environment for the visually impaired using microsoft kinect sensor," *Journal of Sensors*, vol. 2016, 2016.

[20] S. Bhatlawande, M. Mahadevappa, J. Mukherjee, M. Biswas, D. Das, and S. Gupta, "Design, development, and clinical evaluation of the electronic mobility cane for vision rehabilitation," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 22, no. 6, pp. 1148–1159, 2014.

[21] S. Bhatlawande, A. Sunkari, M. Mahadevappa, J. Mukhopadhyay, M. Biswas, D. Das, and S. Gupta, "Electronic bracelet and vision-enabled waist-belt for mobility of visually impaired people," *Assistive Technology*, vol. 26, no. 4, pp. 186–195, 2014.

[22] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[23] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1532–1545, 2014.

[24] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[25] M. J. Saberian and N. Vasconcelos, "Boosting algorithms for detector cascade learning." *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2569–2605, 2014.

[26] R. Benenson, M. Mathias, R. Timofte, and L. Van Gool, "Pedestrian detection at 100 frames per second," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2903–2910.

[27] L. Bourdev and J. Brandt, "Robust object detection via soft cascade," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2, June 2005, pp. 236–243 vol. 2.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[29] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *European Conference on Computer Vision*. Springer, 2014, pp. 346–361.

[31] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

[32] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[33] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[34] K.-H. Kim, Y. Cheon, S. Hong, B. Roh, and M. Park, "Pvanet: Deep but lightweight neural networks for real-time object detection," *arXiv preprint arXiv:1608.08021*, 2016.

[35] S. Zagoruyko, A. Lerer, T.-Y. Lin, P. O. Pinheiro, S. Gross, S. Chintala, and P. Dollár, "A multipath network for object detection," in *BMVC*, 2016.

[36] F. Yang, W. Choi, and Y. Lin, "Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2129–2137.

[37] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, 2016, pp. 354–370.

[38] G. Virgili and G. Rubin, "Orientation and mobility training for adults with low vision," *The Cochrane Library*, 2006.

[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[40] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large

scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.