International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT2173130

Machine Learning Based Computer Vision Application for Visually Disabled People

Shubhada Mone[1], Nihar Salunke[2], Omkar Jadhav[2], Arjun Barge[2], Nikhil Magar[2]
[1]Faculty at Department of Computer Engineering, SPPU, India
[2]Department of Computer Engineering, SPPU, India

ABSTRACT

Article Info
Volume 7, Issue 3
Page Number: 488-494

Publication Issue :
May-June-2021

Article History
Accepted : 12 June 2021
Published : 22 June 2021

With the easy availability of technology, smartphones are playing an important role in every person's life. Also, with the advancements in computer vision based research, Automatic Driving cars, Object Recognition, Depth Map Prediction, Object Distance Estimation, have reached commendable levels of intelligence and accuracy. Combining the research and technological advancements, we can be hopeful in creating a computer vision based mobile-application which will help guide visually disabled people in performing their day to day tasks with easily available mobile applications. With our study, the visually disabled can perform simple tasks like outdoor/indoor navigation without encountering obstacles, also they can avoid accidental collisions with objects in their surroundings. Currently, there are very few applications which provide the same assistance to the visually impaired. Using physical tools like sticks is a very common practice when it comes to avoiding obstacles in a visually disabled person's path. Our study will be focused on object detection and depth estimation techniques- two of the most popular and advanced fields in Intelligent Computer vision studies. We have explored more on the traditional challenges and future hopes of incorporating these techniques on embedded devices.

Keywords : Machine Learning, Computer Vision, Mobile Application Development, Cloud Computing

I. INTRODUCTION

The growth of machine learning has proven to be vital in solving complex problems which otherwise could have taken a lot of effort and thousands of lines of code to be solved. The application of ML is solving computer vision based problems is extremely obliging as perception of vision by computers is very different. The application which we are planning to build relies completely on two important steps- determining the object if any in the path of the user and secondly, determining the distance of the object from the user(the blind person). Object detection will help us determine the position and identify the object. [1] Depth estimation will be crucial in determining the distance of the object from the user.

One of the key problems in computer-vision was to predict the depth of a given image. Traditional approaches [16], highly relied on masking algorithms to solve this issue with reasonable accuracy but the key trade-off is that these techniques have a lot of hardware requirement. Our motive is to provide an application which will be readily available and usable.

Monocular depth estimation models can be trained with the help of standard state of the art datasets [17], [18]. But a common observation persists that machine learning based approaches always incur a lot of prediction even on powerful computing devices. Hence our primary priority is to use light-weight computation models such as the MobileNet.

SYSTEM ARCHITECTURE

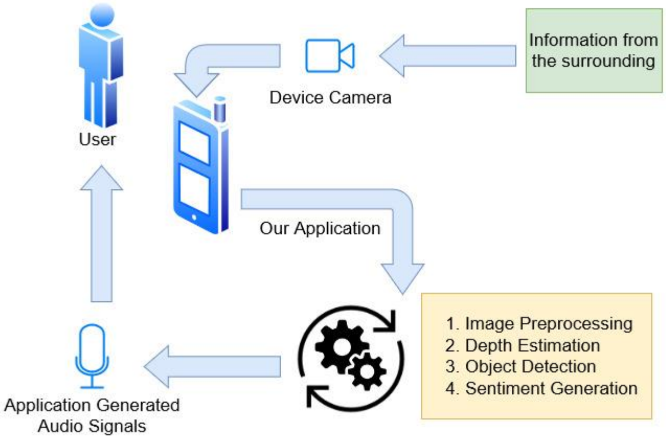

Fig.1 System Architecture

1. Information Gathering : The information is captured using the device's camera. With the advancements in camera quality in mobile phones, we can expect satisfactory results in the image quality.
2. Depth Estimation; is crucial in order to obtain the depth value associated with each object in the image. Depth values will eventually give us the distance of each object from the user/viewer.
3. Object Detection: In the aiding application, object detection will play a crucial role in obtaining information about an object and its exact position with reference to the user.
3. Sentiment Generation and Audio Output: Sentiment generation will be used to summarize the information in the image so that the user can understand the view or objects present before him. The summary will be generated similar to table 3 and then narrated to the user by using the device speakers.

A. Depth Estimation

Depth sensing is a critical function for robotic tasks such as localization, mapping and obstacle detection. There has been a significant and growing interest in depth estimation from a single RGB image, due to the relatively low cost and size of monocular cameras. However, state-of-the-art single-view depth estimation algorithms are based on fairly complex deep neural networks that are too slow for real-time inference on an embedded platform, for instance, mounted on a micro aerial vehicle. With our application, we have encorporated the use of fast depth estimation model- the MobileNet[10].

Since we are building an application that highly depends on real-time prediction, we need to make sure that the accuracy is high as well as the prediction time of our model is as low as possible. State of the art depth estimation models which are trained on heavy datasets, rely on complex architectures for training their models and thus end up with higher prediction time as proposed in [6].

Due to the performance constraints, we selected one of the best performing model with the least complexity to perform the depth estimation task in our application[10].

In the findings in [10], the depth estimation neural network uses a Mobile Net Architecture which is a very light weight deep learning architecture with very few training parameters as compared to the other famous architectures. The point to notice here is that, the efficiency comes at the cost of reduced accuracy with MobileNet Architecture.

How did MobileNet architecture for Depth estimation [10] achieve such high performance? - Along with using a light weight architecture, other key factors such as Depthwise separable convolutions, network prunning, skip connections, using addition of channels instead of concatenation have helped the model to performing exceptionally well with embedded devices.

Traditional convolutions: In traditional convolutions, a 3-dimensional kernel is used to perform convolutions on all the channels (R, G and B) of a given image. This process is computationally expensive during training since the number of multiplications per convolution are 3*3*Si*Si*Sk*Sk, where the given image is a 3-channel image, Sk is the height and width of the kernel and Si is the height and width of the given image.

Depth-wise separable convolutions: It is a technique where we perform convolutions with kernels of size 1*1*N considering that the image has N channels. This method is very computationally efficient and has provided multiple order magnitude of speedup, hence widely used in the MobileNet architecture [10]. **Pruning;** is the process of removing weight connections in a network to increase inference speed and decrease model storage size. In general, neural networks are very over parameterized. Pruning a network can be thought of as removing unused parameters from the over parameterized network.

In our application, the distance of a detected object from the viewer will be estimated with the help of depth estimation.

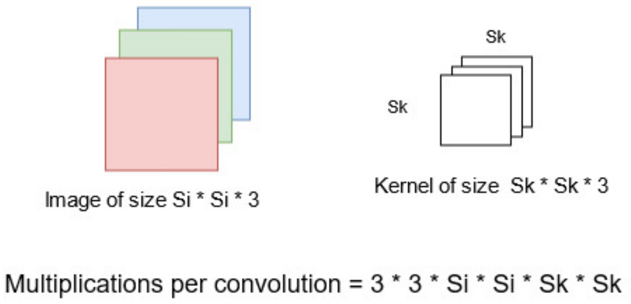
Image of size Si * Si * 3
Sk
Sk
Kernel of size Sk * Sk * 3
Multiplications per convolution = 3 * 3 * Si * Si * Sk * Sk

Fig.2 - Traditional Convolution

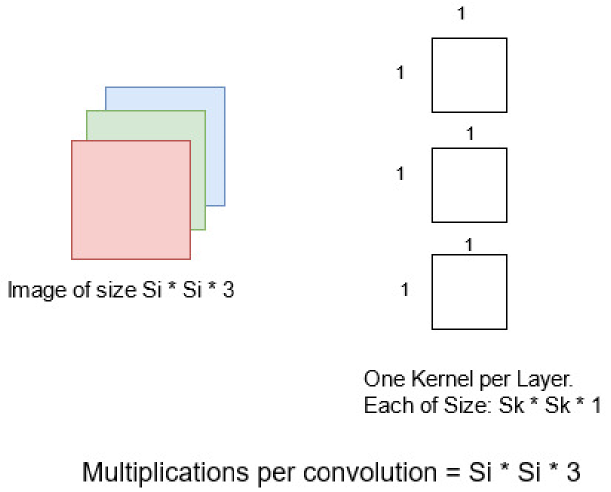
1
1
1
1
1
1
Image of size Si * Si * 3
One Kernel per Layer.
Each of Size: Sk * Sk * 1
Multiplications per convolution = Si * Si * 3

Fig.3 - Depth-wise Separable Convolution

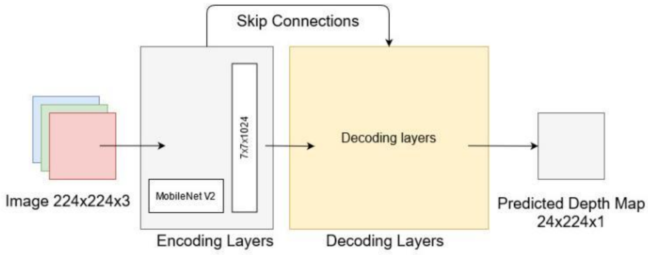

Fig.4 - MobileNet Architecture for Depth Estimation

Experimentally Observed Average Prediction Time with MobileNetV3 Object Detection: **10ms.**
Image Size: 240x240x3.

Some previous work has been done on the techniques of distance estimation from an object given in an image[8]. But this technique uses the ResNet50 [9] architecture in Distance Regressor model which is very heavy for computations.

B. Object Detection

Object detection algorithms such as YOLO[4], SSD[3], HOG[2], Fast R-CNN are used in several applications today. However, it is observed that Single Shot Detector(SSD) has outperforms the other algorithms in terms of prediction time and fast computation. The YOLO algorithm is observed to be ineffective in cases where target objects are closely placed in the image and cases where the target object is very large [5].

Many famous architectures used for object detection such as ResNet and VGG16 are computation-heavy and hence cannot be easily used on embedded devices. In our application, we use the MobileNet SSDv2 model [11] because it is a light weight model and it can deliver expected performance even on embedded devices. [11] describes an efficient network architecture and a set of two hyper-parameters in order to build very small, low latency models that can be easily matched to the design requirements for mobile and embedded vision applications. The MobileNet structure is built on depth-wise separable convolutions [7] as mentioned in the previous section except for the first layer which is a full convolution. By defining the network in such simple terms we are able to easily explore network topologies to find a good network.

MobileNet can also be deployed as an effective base net-work in modern object detection systems. [12] reported results for MobileNet trained for object detection on COCO database which won the 2016 COCO challenge. MobileNet Architecture used for Multiple Object detection:

Table.1 MobileNet Architecture

Type / Stride	**Filter Shape**	**Input Size**
Conv / s2	3 x 3 x 3 x 32	224 x 224 x 3
Conv dw / s1	3 x 3 x 32 dw	112 x 112 x 32
Conv / s1	1 x 1 x 32 x 64	112 x 112 x 32
Conv dw / s2	3 x 3 x 64 dw	112 x 112 x 64
Conv / s1	1 x 1 x 64 x 128	56 x 56 x 64
Conv dw /s1	3 x 3 x 128 dw	56 x 56 x 128
Conv / s1	1 x 1 x 128 x 128	56 x 56 x 128
Conv dw / s1	3 x 3 x 128 dw	56 x 56 x 128
Conv / S1	1 x 1x 256 x 256	28 x 28 x 128
Conv dw / s2	3 x 3 x 256 dw	28 x 28 x 256
Conv / s1	1 x 1 x 256 x 256	28 x 28 x 256
5X Conv dw / s1 Conv / s1	3 x 3 x 512 dw 1 x 1x 512 x 512	14 x 14 x 512 14 x 14 x 512
Conv dw / s2	3 x 3 x 512 dw	14 x 14 x 512
Conv / s1	1 x 1 x 512 x 1024	7 x 7 x 512
Conv dw / s2	3 x 3 x 1024 dw	7 x 7 x 1024
Conv / s1/	1 x 1 x 1024 x 1024	7 x 7 x 1024
Avg Pool / s1	pool 7 x 7	7 x 7 x 1024
FC / s1	1024 x 1000	1 x 1 x 1024
Spftmax / s1	Classifier	1 x 1 x 1000

Table.2 Comparison of MobileNet with other Deep Learning Architectures

Framework Resolution	Model	mAP	Billion Mult-Adds	Million Parameters
SSD 300	Deeplab-VGG Inception V2 MobileNet	21.1% 22.0% 19.3%	34.9 3.8 1.2	33.1 13.7 6.8
Faster-RCNN 300	VGG Ineception V2 MobileNet	22.9% 15.4% 16.4%	64.3 118.2 25.2	138.5 13.3 6.1
Faster-RCNN 600	VGG Ineception V2 MobileNet	25.7% 21.9% 19.8%	149.6 129.6 30.5	138.5 13.3 6.1

Experimental Average Prediction Time with MobileNet SSDV2 Object Detection: **110ms**
Image Size: **320x320x3**

C. Combined Inference

Combined inference is our algorithm wherein, based on the detected object, we will find the approximate distance between our user and the object by referring to the object's depth colour. Darker the colour, closer is the object.

A bounding box will be constructed around the object and an average of all the pixel values within the bounding box will be taken to calculate the object distance.

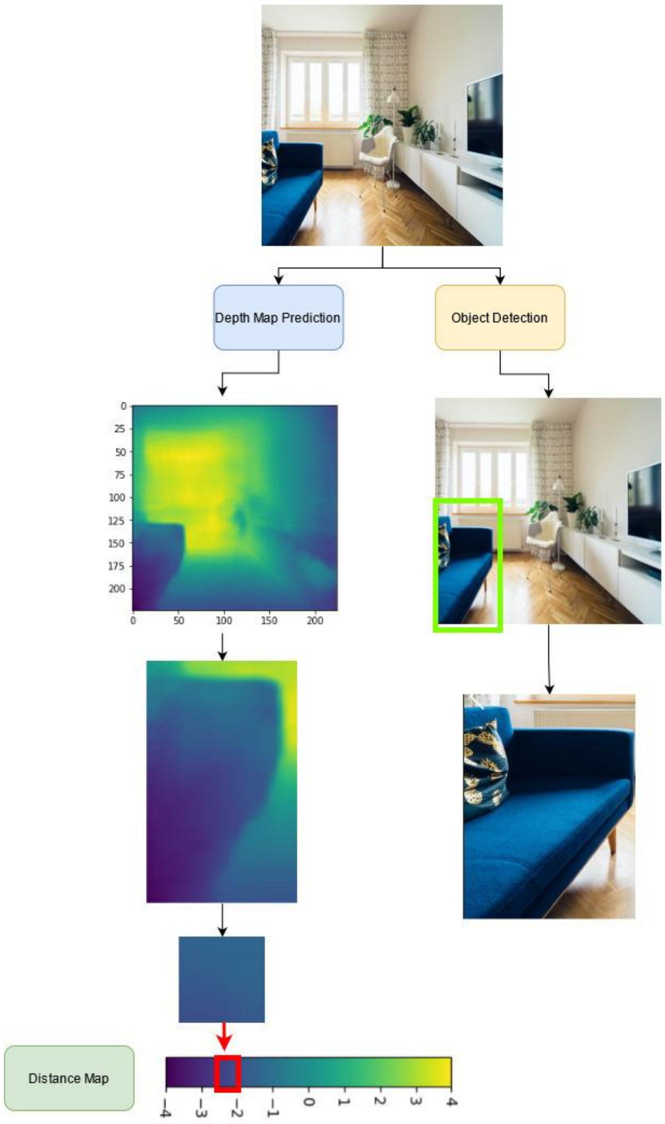

Fig.5 Combined Inference and Distance estimation

In figure 5, the object distance from the viewer is obtained from a single object which is the 'sofa'. But, if the same approach is used for all the detected objects in the image, we infer a better results about the surroundings.

Table.3 Inference generation

Object	**Range**	**Relative Position**
Sofa	Close	Left
Television	Close	Right
Chair	Far	Center
…	…	…

For simplicity in the results, we have divided the Distance Map into three segments viz. Close-Medium-Far. By doing this, we have a chance to compensate for the errors caused in average pixel values because of inaccurate bounding boxes.

After this step, the application will narrate the outputs of the table to the user. Audio signal output or outputs generated in the form of physical signals like vibrations could also prove to be vital in this case.

II. CONCLUSION

We have studied and applied the state of the art techniques of Depth Estimation and Object Detection - the 2 most important concepts when it comes to designing a software for the visually disabled. With a comparative study and analysis of MobileNet based machine Learning architectures, we have thus explored the wide possibilities of designing machine learning models for embedded devices.

III. REFERENCES

[1]. C. Godard, O. M. Aodha, and G. J. Brostow. “Digging into self-supervised monocular depth estimation 2018, arXiv:1806.01260. Online]. Available: https://arxiv.org/abs/1806.01260

[2]. G. Lian, "Pedestrian detection using quaternion histograms of oriented gradients," 2020 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS), 2020, pp. 415-419, doi: 10.1109/ICPICS50287.2020.9202071.

[3]. A. Womg, M. J. Shafiee, F. Li and B. Chwyl, "Tiny SSD: A Tiny Single-Shot Detection Deep Convolutional Neural Network for Real-Time Embedded Object Detection," 2018 15th Conference on Computer and Robot Vision (CRV), 2018, pp. 95-101, doi: 10.1109/CRV.2018.00023.

[4]. W. Lan, J. Dang, Y. Wang and S. Wang, "Pedestrian Detection Based on YOLO Network Model," 2018 IEEE International Conference on Mechatronics and Automation (ICMA), 2018, pp. 1547-1551, doi: 10.1109/ICMA.2018.8484698.

[5]. Q. Zhao, T. Sheng, Y. Wang, F. Ni, and L. Cai, “CFENet: An accurate and efficient single-shot object detector for autonomous driving,” CoRR, arXiv:1806.09790, 2018.

[6]. Zhiqiang Long Dongbing Gu Ruiho Li, Sen Wang. Deepvo: Monocular visual odometry through unsupervised deep learning. In IEEE International Conference on Robotics and Automation (ICRA), 2018.

[7]. R. Zhang, F. Zhu, J. Liu and G. Liu, "Depth-Wise Separable Convolutions and Multi-Level Pooling for an Efficient Spatial CNN-Based Steganalysis," in IEEE Transactions on Information Forensics and Security, vol. 15, pp. 1138-1150, 2020, doi: 10.1109/TIFS.2019.2936913.

[8]. J. Zhu and Y. Fang, "Learning Object-Specific Distance From a Monocular Image," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3838-3847, doi: 10.1109/ICCV.2019.00394.

[9]. Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, “Deep Residual Learning for Image Recognition” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778.

[10]. D. Wofk, F. Ma, T. -J. Yang, S. Karaman and V. Sze, "FastDepth: Fast Monocular Depth Estimation on Embedded Systems," 2019 International Conference on Robotics and Automation (ICRA), 2019, pp. 6101-6108, doi: 10.1109/ICRA.2019.8794182.

[11]. Yu-Chen Chiu, Chi-Yi Tsai, Mind-Da Ruan;Guan-Yu Shen;Tsu-Tian Lee, “Mobilenet-SSDv2: An Improved Object Detection Model for Embedded Systems.(Object Detection)” 2020 International Conference on System Science and Engineering (ICSSE) .

[12]. Wang, B., Fremont, V., & Rodriguez, S. A. (2014). “Color-based road detection and its evaluation on the KITTI road benchmark.” 2014 IEEE Intelligent Vehicles Symposium Proceedings. doi:10.1109/ivs.2014.6856619 (kitti dataset)

[13]. Ming, A., Wu, T., Ma, J., Sun, F., & Zhou, Y. (2016). “Monocular Depth-Ordering Reasoning with Occlusion Edge Detection and Couple Layers Inference”. IEEE Intelligent Systems, 31(2), 54–65. doi:10.1109/mis.2015.94

[14]. Y. -C. Chiu, C. -Y. Tsai, M. -D. Ruan, G. -Y. Shen and T. -T. Lee, "Mobilenet-SSDv2: An Improved Object Detection Model for Embedded Systems," 2020 International Conference on System Science and Engineering (ICSSE), 2020, pp. 1-5, doi: 10.1109/ICSSE50014.2020.9219319.

[15]. S. Zhang, C. Wang and S. C. Chan, "A new high resolution depth map estimation system using stereo vision and depth sensing device," 2013 IEEE 9th International Colloquium on Signal Processing and its Applications, 2013, pp. 49-53, doi: 10.1109/CSPA.2013.6530012.

[16]. Silberman N., Hoiem D., Kohli P., Fergus R. (2012) Indoor Segmentation and Support Inference from RGBD Images. In: Fitzgibbon A., Lazebnik S., Perona P., Sato Y., Schmid C. (eds) Computer Vision – ECCV 2012. ECCV 2012. Lecture Notes in Computer Science, vol 7576. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-33715-4_54

[17]. Z. Xiao, B. Dai, T. Wu, L. Xiao and T. Chen, "Dense Scene Flow Based Coarse-to-Fine Rigid Moving Object Detection for Autonomous Vehicle," in IEEE Access, vol. 5, pp. 23492-23501, 2017, doi: 10.1109/ACCESS.2017.2764546.

Cite this article as :

Shubhada Mone, Nihar Salunke , Omkar Jadhav, Arjun Barge, Nikhil Magar, "Machine Learning Based Computer Vision Application for Visually Disabled People", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7, Issue 3, pp.488-494, May-June-2021. Available at doi : https://doi.org/10.32628/CSEIT2173130
Journal URL : https://ijsrcseit.com/CSEIT2173130