Embedded Selforganizing Systems

Issue Topic: "Artificial Intelligence & Machine Learning in Computer Vision Applications "

Collision Avoidance by Identifying Risks for Detected Objects in Autonomous Vehicles

Hydar Hasan
Department Computer and Automatic Control Engineering
University Tishreen
Latakia, Syria
hydar.h.hasan@tishreen.edu.sy

Majed Ali
Department Computer and Automatic Control Engineering
University Tishreen
Latakia, Syria
mjed.ali.1976@tishreen.edu

Abstract[1]— The past decades have witnessed an increasing percentage of traffic accidents resulting from errors in the perception of the human driver. as confirmed by the World Health Organization in 2018, according to a report issued by it, which shows traffic accidents as the ninth cause of death in the world. Today with the rapid development of computer vision systems, and what deep learning has achieved in the field of artificial intelligence in solving many complex problems. In this study, we have proposed a system based on computer vision and deep learning that detects objects in front of the car in our road environment and the risk factor for it according to the estimate of its distance from the vehicle, as when the distance of the detected object is less than the imposed minimum (20 meters), the system alerts the driver as soon as Possible time to take appropriate action to prevent collision and reduce its occurrence. The main requirement of the proposed system is to provide a technical solution that is easy to implement, provides high accuracy, works in real time and at reasonable prices, which reduces the increased accident rates. Therefore, our proposed system uses monocular camera images and the latest object recognition model, the YOLO model, which achieved a recognition rate of 61% and estimated distance using the boundaries of the detected object and depth maps accurately and quickly to help the driver alert him and avoid collision. Our proposed system provides a starting point for an integrated system that avoids collision, achieves safety and saves lives in our street environment.

Keywords— deep learning; distance estimation; collision avoidance; depth map; machine learning.

I. INTRODUCTION

in the field of developing autonomous driving cars computer vision has become a key component, due the advances in visual environment perception like object classification, detection, segmentation and distance estimation Techniques .Researchers mainly focus on object classification, detection and segmentation [1, 2, 3], beside a lot of efforts on improving the accuracy of visual perception crucial information for cars to avoid collisions, adjust its speed for safety driving can be provided through distances estimating between camera sensors [25,26,27] and recognized objects (e.g. cyclists , pedestrians ,cars) and not only by recognizing the objects on the way.

Today with the rapid growth of technology, internet applications which led to the availability of huge quantities of data, the creation of robust and highly performant computers ,the emergence of the convolutional neural networks, researchers have achieved remarkable progress on traditional 2D computer vision tasks using deep learning techniques, such as object detection, semantic segmentation, instance segmentation, scene reconstruction [4, 6, 7, 5],but we have failed to find any deep learning application on object-specific distance estimation

In this paper, we will focus on how deep learning and computer vision technique can be used to estimate the distance of the detected object (s) and raise an alarm if this distance is within the danger range, then this can help boost the driver to take action in order to avoid the crash by reducing the speed significantly.

II. MOTIVATION

This paper is mainly motivated by a research report released in 2018 by the World Health Organization (WHO) which showed how alarming was the number of deaths on our roads each year[8].This same report says that if nothing is done road accident will be the 5th cause of death among the youths by 2030. there is no barrier preventing us to take advantage of these to propose a solution. To find a solution, we suggest a system which will recognize objects on our streets, valuation the distance of these object from the camera and warning the driver if this distance is equal or less than the limit value(20meters).

The main motivation of this paper is to formulate an "easy to implement" and affordable technology solution for the ever increasing accident rates during driving.

III. LEARNING-BASED DETECTION TECHNIQUES

It's a computer vision technique whose aim is to enable the computer to detect surrounding objects even better than humans and even in poor lightening and weather conditions. Object detection is made up of classification and drawing of the bounding box around the object. Object detection is used in many activities such as driver assisted and self-driving (automobile industry), Surveillance systems (smart security).

A. Deep Learning in Detection

computer vision researches are widely using deep learning, in general object recognition using deep learning have two approaches the first one which is known as a two stages approach (R-CNN) comparing to previously published sliding window-based techniques provides a remarkable improvement [9], an unattended algorithm for feature generating extraction by CNN's separately is used in R-CNN known as selective search . to estimate the classes of objects SVM classifier is applied by R-CNN in the last step , and to fine-tune the positions and sizes of detection boxes a linear regression is applied to get a better performance. After the spectacular effect of RCNN, many new ideas have been introduced on CNN, such as Fast R-CNN [10], Faster R-CNN [11] and the spatial pyramid pooling network (SPP-Net) [12]. As for the object detection techniques mentioned above, they have significantly improved the accuracy and speed of object recognition, while there are techniques that adopt a one-stage approach such as YOLO (you only look once) which is considered one of the fastest object detection algorithms which makes use of Convolutional Neural Network. It's not the most accurate but it's one of the best choices for real time object detection tasks while maintaining an accepted level of accuracy. Some researchers have done a good job to convert the YOLO implementation from DarkNet to a TensorFlow using python programming language. The latest version (YOLO v3) has codename DarkNet53[15] which indicate that it contains 53 convolution layers each separated by a normalization layer and a Leaky ReLU activation. It's a fully convolutional Network. another technique similar to the YOLO approach is The SSD (Single Shot multi box Detector) [14], which to increase the accuracy uses multi-scale feature mapping layers and standard boxes .

Regarding the deep learning approaches, the two step strategies have advantages in terms of recognition accuracy and localization accuracy. However, a large amounts of time and resources is needed and the efficiency of calculation is poor .unified network structures accelerate he single-stage methods ,although the accuracy of the process decreases .In addition , a major factor for deep learning-based methods is the amount of dataset

B. Distance Estimation

In computer vision this is one of the top problems which often comes forward after the object of interest is identified since object detection and depth estimation goes together. Many researchers have worked on distance estimation or depth estimation of an object from the source, using camera, A variety of techniques, methods and algorithms have been developed in the due course. This finds application in wide range of areas: intelligent transportation systems, robotics, the automobile industry with self-driving cars[17] etc. Distance estimation can be done either passively or actively.

1) The Active method : signals (usually radio & laser beam) are sent to the object(s) which we want to measure the distance. Such a system uses sensor such as Ultrasound, laser scanners or time-off light to search for surrounding objects. Active methods are usually very expensive methods.

2) The Passive method: The images provided from the camera and computer vison techniques to be able to estimate the object distance from the camera. Such a system usually has a low cost and can be used in variety of domains such as robotics, Virtual reality and industrial automation[18].

a) The Stereo vison: This method uses two cameras to calculate the distance of the object. The two cameras parallel to each other and observe the object from different locations which result in different image locations. It exploits the parallax phenomenon which refers to a displacement of the apparent position of the object view from different line of sight. The object distance is calculated from the relative distance of the object position on both cameras. The difference in image location is known as disparity while the distance between the two cameras is called baseline.

b) The Monocular camera: This method uses a monocular camera and computer vison technique to estimate the distance of the object in the image Even though many researchers have work on stereo vision, we are interested in the monocular method which sooth this paper context and turns to be cheaper. Many researches go from the principle of stereo vision and try to implement it using a single camera Several. where they used strategies like the temporal method for distance estimating between position of the monocular camera and the target , in temporal method for example the distance is computed based on the temporal sequence of object copies (e.g. visual odometry), while in The another way the prediction of distance is in the actual frame independently of previous frames (e.g. triangulation methods or depth estimation based on CNN techniques).

One important thing note in Many researches Which I was reading was suffering from limited by the depth used to the camera performance This is what we are more interested in and we went forward to look at what some researchers have done regarding this.

There are many methods to estimate the distance of objects using different sensors. The advantages and disadvantages of some of these methods are presented in the table 1 .

TABLE I. :COMPARISON OF DISTANCE ESTIMATION TECHNOLOGY.

technology	***advantages***	***Disadvantages***
TOF Camera	small aspect ratio, easy calibration	requires continuous active lighting, resolution poor
LIDAR	High resolution, precise perception in the dark	Heavy and large devices. high cost. Vulnerable to bad weather condition.
Stereo Camera	Image scale and depth information is easy to be retrieved Provide 3D vision	More cost and needs more calibration effort than monocular cameras, Complex multi-step refinement processes or Global processing requirements that demand huge memory size and bandwidth
Monocular	Low cost, good for small robotics Simple calibration	Suffer from image scale uncertainty, Very large training set and learning approach. computationally intensive

C. depth estimation

depth image is an image which contain information about the distance of each point in the image from the viewpoint. The intensity value of each pixel represents the distance of the that point from the camera. Depth map images are grayscale and brighter a point the further way it is and the darker a point, the closer it is. Now our days we have specialised camera devices (e.g. Kinect) and 3D software which are used to produce depth images. Such images are very important and used in augmented reality usually for special visual effect and scene reconstructions[21]. Researchers have also worked to develop this techniques. where there very interesting method is the depth estimation with monocular images is done using deep learning model trained on pair of images with their depth maps. This method provides decent accuracy[22]. This can be done by either using Supervised Learning or Unsupervised learning. The commonly used dataset which contain both label depth maps and raw data used to trained the model are; NYUv2(RGB-D depth maps for indoor images), Kitti(RGB-D depth maps for road images and Make3D (RGB-D depth maps for outdoor).

IV. THE PROPOSED SYSTEM AND METHODOLOGIES

our system proposes an Object distance estimator and notification warning system which uses as input images from monocular camera and uses deep learning models together with some computer vison techniques for processing and giving out the result in real-time.. We started this system by doing an analysis of the requirement necessary and a feasibility study . Then we we came forward with a schematic design on how the system should work as seen in the figure 1.

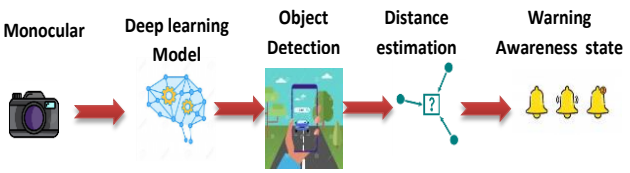

Figure 1: System data flow overview design.

From Our proposed system Shown in the figure1 , the sequences of images are collected from a monocular camera (or video file) and passed as input to the deep leaning model. The deep learning model here represent actually two models; one for object detection and the other for depth map generation. The object detector model will identify and locate the position of a target object in the image, while the depth map generator model will generate the depth map from the image, and this depth map will be used together with the coordinate of the of the object from the detector, to estimate the distance of this object from the camera. Once the object is detected, position known and distance from the camera calculated, we will now be able to tell if this distance is within the danger zone or not. If this distance is within the danger range, the warning state will be stated notifying the driver to take care and be more careful.

To Implement this project and make it realistic within the limited time frame, we divided it into 4 sub modules which are:

- Train model and build Object detector.
- Generate the depth maps from Color images (2D images)
- Integrate the depth map with Object detection
- Implement the Warning state.

A. Train model and build Object detector

In this part of the proposed system we are interested in bringing up an object detector to detect objects of specific classes that can intercept a moving vehicle that work in a real-time with high-precision. After studying the state-of-the-art object detection architectures (YOLOv3, Faster R-CNN, SDD), we note YOLOv3 has proven to be the best option for us since it has a very good balance between the accuracy and speed which is essential for this system. Then we chose dataset which will provide us with high quality labelled images of the classes of object (Car, Van, Bus, Truck, Pedestrian, Motorcycle, Dog, Traffic sign) which we will train our classifier. For this, we had two options; Google Open images dataset and Kitti dataset were more suitable for this. The training method we are interested in doing here is supervised learning using a pretrained model trained on the MS COCO dataset. Our training script was configured to train in two phases sequentially, each with 50 epochs but with batch size of 32 for the first and 8 for the second. We downloaded the labelled dataset from Kitti (7481 training images and 7518 test images, 12GB) and GOID (5000 images for each class). We started by training on the Kitti dataset which took 4 days to complete. We tested the resulting Keras model on a couple of images and the performance from first view was poor. The model could not accurately detect the target objects from the images unless these objects shaped was well shown in a good quality image. so We also train the original pretrained model on the GOID (Google open image dataset). It also took 4 days to complete the training. We tested the resulting Keras model and the result was encouraging. We tested with serval set of images and the model could identify the object in the images. We wrote a script to test this model in real time, with data

from a video file and the model was able to detect the object at an acceptable latency.

B. Generate the Distance from Color images (2D images)

depth estimation from images is a very important task in many applications and used in augmented reality usually for special visual effect and scene reconstructions. Depth map is a gray scale image that contains the distance of the objects from a viewpoint. In this phase, we are interested in generating depth images from the 2D RGB images fed to our system. This depth image would be necessary in to estimate the distance of the detected object.

we will use the deep learning supervised method which is one of the recommended ways for this with decent level of accuracy. We decide to go on form a project done and published by Ibraheem Alhashim and Peter Wonka in their paper title High Quality Monocular Depth Estimation via Transfer Learning[23]. They trained and tested their model on both the Kitti[25] and NYUv2[24] dataset. This was done using a simple and yet efficient architecture which produces quality depth map with good resolution. This architecture is an encoder-decoder architecture where this encoder make use of a pre-trained DenseNet-169 (trained on the ImageNet) and the decoder made up of two convolutional layers applied after a 2X bilinear sampling step.

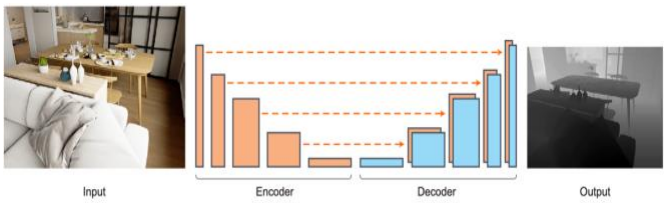

Figure 2: overview of Alhashim and Peter network architecture.

C. Integrate the depth map with Object detection for distance estimatio

In this phase of implementation, the input image passed into our model, the previous two steps will detecte the objects, and pridect boundary boxes to indicate their positions in the image and create a depth map from this image. From there we will use the object's coordinates and its boundary boxe and extract the pixel value from the depth map based on the coordinates of the center of the boundary boxe of the object detecte to give an estimate of that object's distance from the camera. when The boundary boxe of the object detected is big, the closer its distance to the camera, and the distance increases with the small boundary boxe of the object. We also need to understand the geometry of the pinhole camera projection and we need to know how we can map a point on the 2D to the 3D

D. Implement the Warning state in the sysytem

In this phase of the project, we want to able to notify the driver by anything which can call for his/her attention and let him know that one of the objects detected is at a distance equal or below the threshold distance. The notification is done by a sound together with an alert shown in the image displayed. In our prototype we displayed a text indicator in red and make a beep sound as means of notification.

V. RESULTS AND DISCUSSION

After training and being able to detect the target objects, now is time to test how accurate is our model. In Object detection, the accuracy of the model is tested not only with the real time performance on video or images, but with what is called Intersection Over Union (IoU) and mean Average precision(mAP) detector as shown in the figure 4. We need needed to check the mean average precision (mAP) at the minimum IoU which is conventional taken at 0.5 which gave mAP=61.

Precision and Recall are used to calculate the performance of the detector as shown in the figure 4.

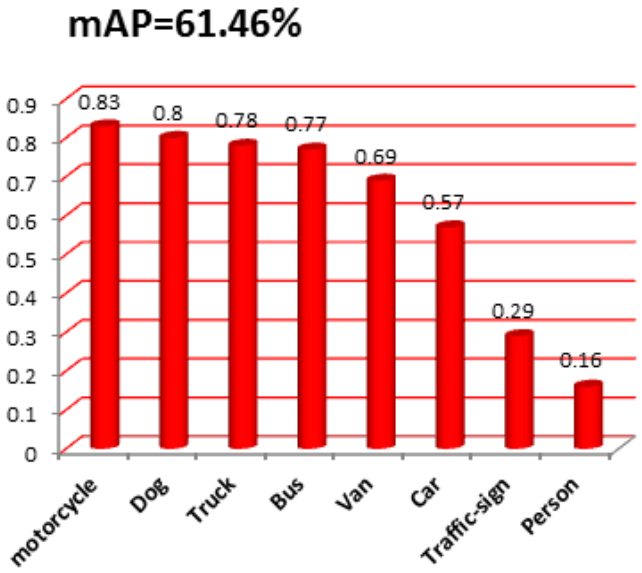

Figure 3: Testing the Accuracy of model

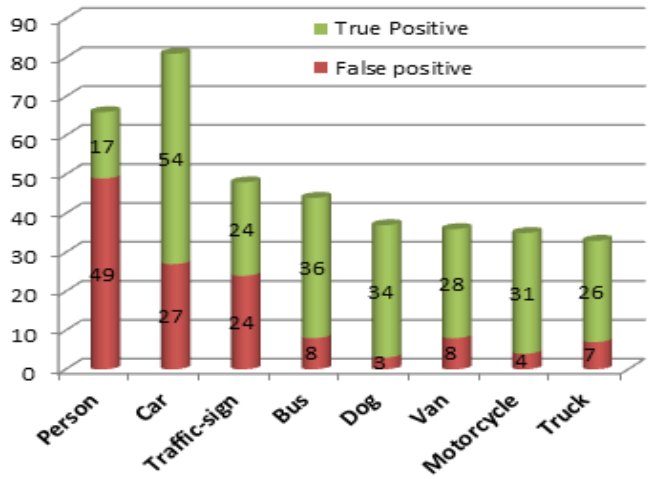

Figure 4: detection result (240 files and 8 detected class)

We tested our systems with various images taken at different angle and at different conditions of the day, also while testing our system we tweak from time to time the parameters to see how our results was affected by these changes. What we realized is that our model worked best for images taken horizontally and for objects not too close to the camera.

As we may have notice in figure 5, the first car detected has distance which is for sure not correct, but the subsequent cars detected have an estimated distance more or less close to the possible actual distances.

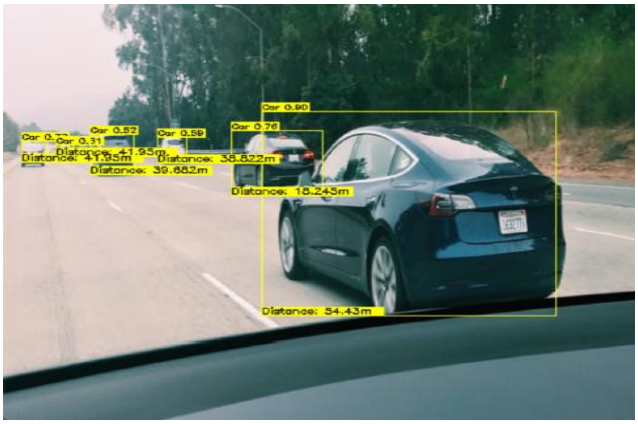

Figure 5: Sample result 1.

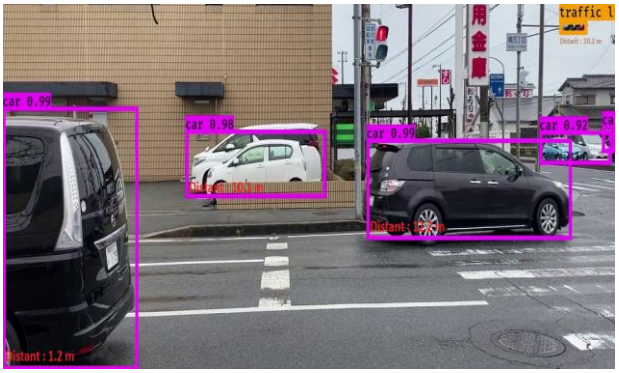

Figure 6: Sample result 2.

From the figure 7, we can see a difference in distance between the Traffic sign and the second car which may lead to some questions. As I said earlier, even though the accuracy of this system needs to be enhanced, the images at a certain angle like the traffic sign may even be more affected.

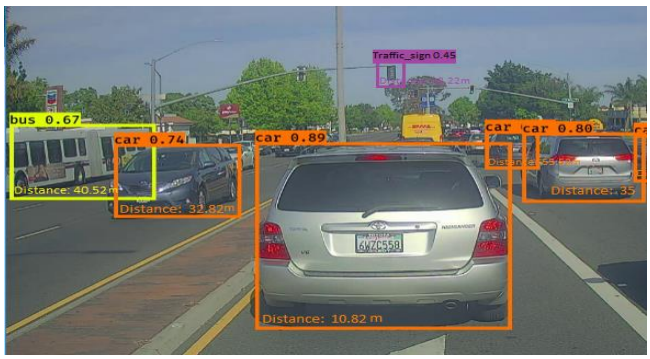

Figure 7: Sample result 2.

Figure 8: Sample result 1.

VI. Conclusion and Future Work

This is quite an interesting and challenging system which shows how much we can exploit the potential of machine learning in solving problems which causes much harm to the society. At this point, we generated our depth map using a pretrained model which was trained on NYUv2 and Kitti depth maps and has a high quality for the depth map images created from it. This allowed us to use information about the objects detected in the first step and the corresponding depth map images to be able to extract information from the depth map at specific point and analyses it to get an estimation of the distance from this point to the camera. We were able to detect and estimate the distance of the detected objects from the camera. Even though there is still much room for improvement, we firmly believe we are on the right track and that is a very good opportunity to contribute to something which can ameliorate the living conditions of people and save lives.

References

[1] Masaru Yoshioka, Naoki Suganuma, Keisuke Yoneda, and Mohammad Aldibaja," Real-time object classification for autonomous vehicle using lidar",*International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), pages 210–211. IEEE, 2017.*

[2] Alex Teichman and Sebastian Thrun," Practical object recognition in autonomous driving and beyond", I*n Advanced Robotics and its Social Impacts, pages 35–38. IEEE, 2011.*

[3] Shu Kong and Charless C Fowlkes," Recurrent pixel embedding for instance grouping",*In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 9018–9028, 2018.*

[4] Liang-Chieh Chen, Alexander Hermans, George Papandreou, Florian Schroff, Peng Wang, and Hartwig Adam. Masklab, "Instance segmentation by refining object detection with semantic and direction features", *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4013–4022, 2018.*

[5] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei," Relation networks for object detection", *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3588–3597, 2018.*

[6] Andy Zeng, Shuran Song, Matthias Nießner, Matthew Fisher, Jianxiong Xiao, and Thomas Funkhouser, "3dmatch: Learning local geometric descriptors from rgb-d reconstructions", *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1802–1811, 2017.*

[7] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia ,"Icnet for real-time semantic segmentation on high-resolution images", *In Proceedings of the European Conference on Computer Vision (ECCV), pages 405– 420, 2018.*

[8] "WHO | Global status report on road safety 2018," *WHO, 2019.*

[9] Girshick, R. Donahue, J. Darrell, T. Malik, J. Rich, " feature hierarchies for accurate object detection and semantic segmentation" , *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.*

[10] Girshick, R," Fast R-CNN", I*n Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; pp. 1440–1448.*

[11] Ren, S.Q. He, K.M. Girshick, R.Sun, J. "Faster R-CNN: Towards real-time object detection with region proposal networks", *IEEE Trans. Pattern Anal. Mach. Intell. 2017, 39, 1137–1149.*

[12] He, K.M.Zhang, X.Y. Ren, S.Q. Sun, J. "Spatial pyramid pooling in deep convolutional networks for visual recognition", *IEEE Trans. Pattern Anal. Mach. Intell. 2015, 37, 1904–1916.*

[13] Redmon, J.Farhadi, "YOLOv3: An Incremental Improvement", *Apr. 2018.*

[14] Liu, W. Anguelov, D.Erhan, D. Szegedy," SSD: Single shot multibox detector", *In Proceedings of the 14th European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.*

[15] J. Redmon and A. Farhadi,"YOLOv3: An Incremental Improvement."

[16] Chang, X.J. Yang," Semi-supervised feature analysis by mining correlations among multiple tasks", *IEEE Trans. Neural Netw. Learn. Syst. 2017, 28, 2294–2305.*

[17] A. S. Suneel, "Person or Object Tracking and Velocity Estimation in RealTime Videos", *Publ. Probl. Appl. Eng. Res. Pap., vol. 04, no. 01, pp. 292–299, 2013.*

[18] A. Yadav and T. B. Mohite-Patil, "Distance Measurement with Active & Passive Method Distance Measurement with Active & Passive Method", Int. J. Comput. Sci. Netw, vol. 1, no. 4, pp. 2277–5420, 2012.

[19] Birchfield, S. and Tomasi," Depth discontinuities by pixel-to-pixel stereo", *International Journal of Computer Vision, C. (1999), 35(3):269–293.*

[20] Committee for Development Policy,"Complete dataset", Available online https: // www.un.org/development/desa/dpad/wp-content /uploads /sites/45/page/LDC_data.xls *at Accessed October 29. 2017.*

[21] R. Garg, G. Carneiro, and I. Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. ECCV, 2016, S.S. Tung and W.L. Hwang, "Depth Extraction from a Single Image and Its Application", *in Pattern Recognition - Selected Methods and Applications, IntechOpen, 2019.*

[22] "Depth Estimation - BeyondMinds - Medium" [Online]. Available: https://medium.com/beyondminds/depth-estimation-cad24b0099f. *[Accessed: 25-Mar-2020]*.

[23] I. A. Kaust and P. Wonka, "High Quality Monocular Depth Estimation via Transfer Learning", *arXiv:1812.11941v2 [cs.CV] 10 Mar 2019.*

[24] "NYU Depth V2 « Nathan Silberman", [Online]. Available: https://ww.cs.nyu.edu/silberman/datasets/nyu_depth v2.html. *[Accessed: 14- May-2020]*.

[25] The KITTI Vision Benchmark Suite." [Online]. Available: http://www.cvlibs.net/datasets/kitti/eval_object.php?ob benchmark=2d. [Accessed: 16-Jan-2020].

[26] Saleh, S., Khwandah, S., Heller, A., Mumtaz, A. and Hardt, W.. Traffic Signs Recognition and Distance Estimation using a Monocular Camera. In: 6th International Conference Actual Problems of System and Software Engineering. [online] Moscow: IEEE, (2019) pp.407-418. Available at: http://ceur-ws.org/Vol-2514/

[27] Saleh, S., Hadi, S., M., Amin Nazari, and Hardt, W." Outdoor Navigation for Visually Impaired based on Deep Learning". In: 6th International Conference Actual Problems of System and Software Engineering. [online] Moscow: IEEE, (2019).pp.397-406. Available at: http://ceur-ws.org/Vol-2514/.

[28] S. M. Saleh, S. A. Khwandah, W. Hardt, M. Hilbrich and P. I. Lazaridis, "Estimating the 2D Static Map Based on Moving Stereo Camera," 2018 24th International Conference on Automation and Computing (ICAC), Newcastle upon Tyne, United Kingdom, 2018, pp. 1-5, doi: 10.23919/IConAC.2018.874900