Towards Robust Perception Depth Information For Collision Avoidance

1st Shadi Saleh
Department of Software Engineering
Technische Universität Chemnitz
Chemnitz, Germany
shadi.saleh@informatik.tu-chemnitz.de

1st Shanmugapriyan Manoharan
Embedded Systems
Technische Universität Chemnitz
Chemnitz, Germany
shanmugapriyan.manoharan@s2017.tu-chemnitz.de

2nd Julkar Nine
Department of Computer Engineering
Technische Universität Chemnitz
Chemnitz, Germany
julkar.nine@informatik.tu-chemnitz.de

3rd Wolfram Hardt
Department of Computer Engineering
Technische Universität Chemnitz
Chemnitz, Germany
wolfram.hardt@informatik.tu-chemnitz.de

2020 IEEE Congreso Bienal de Argentina (ARGENCON) | 978-1-7281-5957-7/20/$31.00 ©2020 IEEE | DOI: 10.1109/ARGENCON49523.2020.9505483

***Abstract* — Early detection of the obstacles and accurate estimation of the object's distance helps avoid fatal accidents. However, the existing object detection ignores debris and other object classes not included in the training process. On the other hand, the driving area is monitored and recognized using active sensors like LiDAR, RADAR, but expensive. This study introduces a modified architecture to estimate the depth map based on an unsupervised learning framework. Furthermore, understanding the color-encoded depth map helps identify the risk of collision. The decoding of the color-encoded depth map provides information about the distance of the object. Thus, we presented an efficient and robust algorithm to predict a potential collision in real-time based on the estimated depth map using predefined threshold values. This approach emphasizes integrating the estimated depth map with the level of comprehension of the situation awareness to enhance the ability to recognize and process predicted uncertainties in the environment. The better results are achieved for modified architecture in terms of ARD, RMSE, RMSE (log), accuracy, and other evaluation metrics achieved lower but comparable results to the state-of-the-art techniques with a maximum depth of 80 meters. The integrated collision avoidance algorithm with the depth estimation architecture achieved a performance of 25 FPS on RTX 2080Ti GPU.**

Keywords—depth map, unsupervised learning, monocular camera, collision avoidance.

I. Introduction

Nowadays, the automobile is the primary transportation means used by people in the modern world. The automotive domain's notable advancements are the Advanced Driver Assistance System (ADAS) and autonomous driving [1]. As the road traffic increases rapidly, these advancements mainly ensure the safety of the driver and passengers. Every year 20 to 50 million people are getting wounded, and around 1.2 million die in car accidents. Several applications developed in ADAS like Adaptive Cruise Control (ACC), collision & lane departure warning, blind-spot detection, reduce accidents [2]. However, all these systems are limited to one or more aspects. Besides, some of the systems' cost is high, and they are not widely used in standard vehicles, which increases the cost of self-propelled cars. The active sensors include Sound Navigation Ranging (SONAR), RAdio Detection And Ranging (RADAR), Light Detection and Ranging (LiDAR), can detect the obstacles but also offers many shortcomings (such as sparse depth, affected by weather conditions, and expensive). The available active sensors cannot achieve better Range, Resolution, Accuracy, and low Size, Weight, Power. These limitations make the available sensors inadequate for autonomous vehicles.

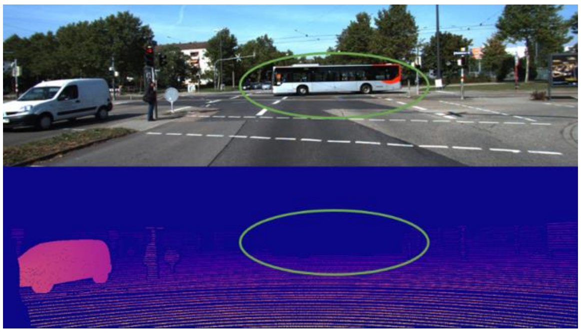


Figure 1: Shows the pretty sparse depth information obtained from KITTI dataset [4]

Object recognition is a crucial factor in advanced driver assistance systems and self-driving cars. However, familiar objects in traffic scenes, such as pedestrians, motorcycles, and different vehicle types, are often dependent on supervised learning techniques. These supervised models are achieved by using a large number of annotated datasets. Thus, the classifiers will be trained to recognize certain objects within the scene. Unfortunately, the above does not apply to different objects such as debris and roadside obstacles that may appear in the captured scene. Furthermore, distance perception is another significant perceptual technique used to avoid collisions with different objects that are too close to the obstacles; usually, it is known as depth estimation.

Nevertheless, depth information plays a crucial role in applications like autonomous driving, 3D scene reconstruction, and robotics [3]. Moreover, capturing ground truth is hard, for example, in the KITTI [4] dataset. This dataset was captured using a car-mounted stereo camera at the top and an expensive Lidar at the roof. Figure 1 shows that it is pretty sparse and does not have any depth information above the horizon line. Moreover, some missing objects, such as the

978-1-7281-5957-7/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: Universidad Nacional de Colombia (UNAL). Downloaded on February 10,2023 at 03:49:57 UTC from IEEE Xplore. Restrictions apply.

bus missing from the dataset because they are moving horizontally.

This study presents an accurate, cost-effective approach with lower power consumption to avoid the collision in real-time capabilities. An efficient approach was developed based on the CNN network to identify the depth information from a frontal monocular camera that is capable of perceiving which aspects of the scene are close and which are further away from the camera pose. The estimated depth allows us to observe objects and surfaces assumed to be in flat images as a 3D scene. The depth information is coded in the color map, where the close region (yellow color) means zero meters from the viewpoint and the far one (blue color) means the 50-meter depth information from the viewpoint. The color encoding of the estimated depth map helps find the objects near, far, and farthest in the environment.

II. Background Research

Every vision algorithms started to use the depth information, and almost all the state-of-the-art algorithms apply a combination of data to extract the depth information. The depth estimation techniques play a vital role in autonomous driving technology in terms of significantly improving accuracy.

A. *Depth Estimation based on Unsupervised Learning*

Numerous unsupervised methods proposed lately to address the problem of understanding 3D. The stereopsis-based auto-encoder was proposed by Garg et al. in order to estimate depth from a single view [5]. The training is sub-optimal as it is based on Taylor expansion. Garg's approach is advanced by Godard et al. [6], which includes left-right consistency and smoothness loss. Learning the monocular video's depth is made possible by incorporating the camera's pose estimation to the training pipeline, proposed by Zhou et al. [7]. Modeling the motion of rigid objects is included in the network by Vijayanarasimhan et al. [8].

B. *Obstacle Avoidance based on Stereo Depth Estimation*

Andert F. et al. [9] presented a stereo depth estimation approach to avoid a collision by the unmanned helicopter and its simulation environment. The method employs stereo cameras to capture the image pair, explores the free areas in the image, and returns the best free region. This approach provides an obstacle-free path and instructions to fly in this path, as shown in Figure 2. This method uses a stereo camera setup for the distance estimation. The stereo camera works poorly in the region without texture. This drawback degrades the performance of this approach. Therefore, in order to overcome this drawback, the distance estimation was realized with a monocular camera.

C. *Collision Avoidance based on Vision and Sensors*

Franke U. et al. [10] have implemented a stereo vision and motion analysis to recognize critical traffic conditions and examine a collision possibility. Wei Y. et al. [11] proposed a system that includes RADAR and LiDAR sensors to determine the object's position near the vehicle and predicts a collision based on the vehicle′ s speed and angular velocity. Prabha M. et al. [12] predicted the collision using the distance information between the two vehicles. The distance between vehicles is detected on both the front and back sides using an ultrasonic sensor node for a limited range.

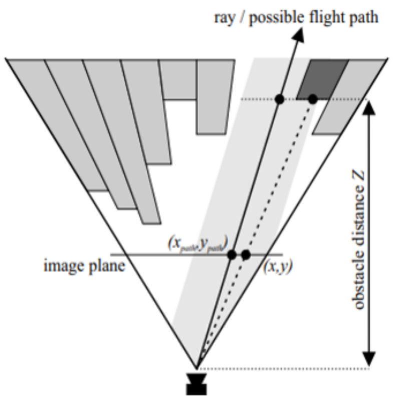


Figure 2: Determining the Flight path using depth values to avoid obstacles [9]

III. Implementation

A. *Comprehension of Color-Encoded Depth Map*

The color-encoded depth information was explained using the RGB color values and distance information (0 - 50 m). The red, green, blue colors indicate that the objects are near, far, and farthest from the viewpoint. Also, the location of the object represented using red, green, blue. At a distance of 0 - 16m, 17m - 34m, 35m - 50m is assigned red, green, and blue colors. The 3D space representation (see Figure 3) depicts that the red, green, blue regions are near, far, and farthest from the camera's viewpoint.

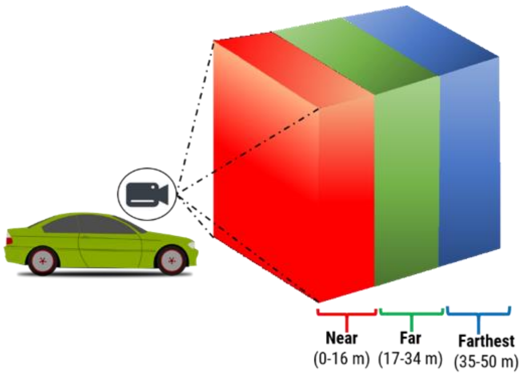


Figure 3: Depth Color-Encoding Comprehension

B. *Proposed System*

The entire system was developed in two steps, as presented in Figure 4. Firstly, the depth estimation method, inspired by the SFMLearner method [7]. A new architecture was developed based on SFMLearner, which is trained to estimate the single view's accurate depth. Then, a color decoding algorithm was implemented to identify the collision free-zone. The monocular camera continuously feeds the acquired frames into the depth model. Our trained depth model accurately predicts the corresponding single-view depth information for each captured frame with real-time capability. Simultaneously, the Region of Interest (ROI) within the camera's field of view is identified to provide the highest quality in areas or a scene of most interest while reducing uninterested regions' quality to accelerate processing time. The estimated depth map within the ROI is then transformed into a plasma color encoding, and based on the volume of closed colors, the probability of collision is estimated.

Authorized licensed use limited to: Universidad Nacional de Colombia (UNAL). Downloaded on February 10,2023 at 03:49:57 UTC from IEEE Xplore. Restrictions apply.

C. Depth Estimation Architecture

The modified architecture (see Figure 5). The 7th convolutional and deconvolutional layers have the 1024 output channels indicated with white font. The increase in the number of output channels increases the number of abstractions that network architecture can extract from the input RGB image. The Kernel size of the first four convolutional layers is 7, 7, 5, 5, and for the remaining layers, it is 3, 32, 64, 128, and so on are the output channels. The total number of parameters for this architecture is ~61:5 million compared to the SFMlearner model is ~33.2 million. This architecture also includes skip-connections and is trained using the KITTI dataset [4], Chainer deep learning framework, and RTX 2080Ti GPU. Hyperparameters used are Adam Optimizer, the learning rate of 0.0002, momentum 0.9, batch size 2, and epochs are 20.

D. Color-Decoding Algorithm

The algorithm for color decoding is implemented, as shown in Figure 6. The output of the depth model is passed as input to our color decoding algorithm. The plasma colors map encodes the estimated depth. In plasma color encoding, the yellow color indicates that the objects are close to the camera position. The blue color identifies that the objects are farthest from the camera position. There are ten colors present in the plasma color; thus, K means clustering is applied to cluster the colored depth map based on different color values (where k =10 is the number of clusters) and not used to cluster the various objects the scene. The histogram normalization is performed on clustered colors to achieve better quality without losing important information. An iteration process is then executed for all clusters to obtain the percentile color ratio that represents each cluster. The collision thresholds considered in the RGB color space are [240, 249, 33] (Th. 2) and [253, 202, 38] (Th. 1). The main reason for including two thresholds is to ensure high safety criteria. Thus, if the color cluster's threshold values have been detected, a collision warning is triggered as a hazard output.

IV. Result and Evaluation

The performance of our model is evaluated using metrics such as RMSE (Root Mean Square Error), RMSE (log), accuracy, ARD (Absolute Relative Difference), and SRD (Squared Relative Difference). Our model's depth information compared with the ground truth depth for the total number of pixels in that image using the RMSE metrics. The lower the value of the RMSE metrics indicates the model performance is good. The absolute and squared relative difference of the predicted depth and ground truth depth calculated using ARD and SRD. The model can predict the depth for a minimum value of zero meters and a maximum of 80 meters. We have evaluated our depth estimation model using the Eigen Test Set of KITTI datasets [4].

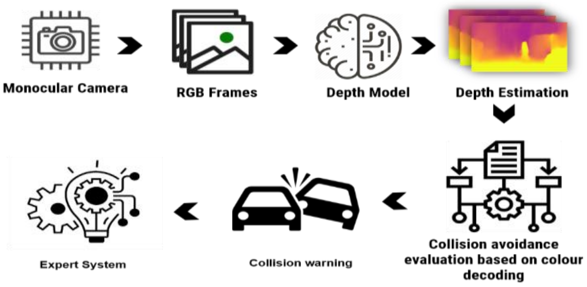


Figure. 4: Systematic Design of the Proposed System.

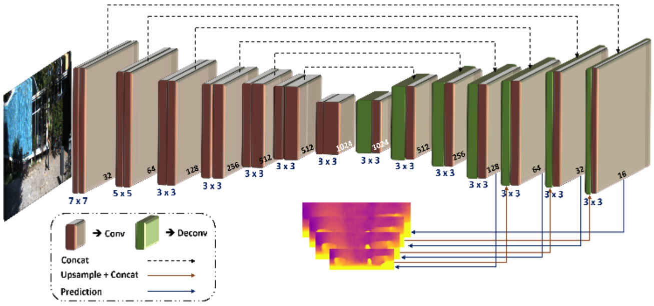


Figure 5: Modified Depth Network Architecture

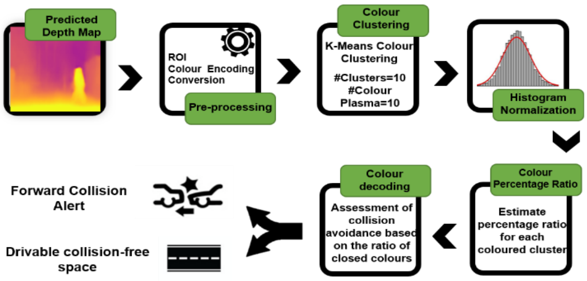

Predicted Depth Map
ROI Colour Encoding Conversion
Pre-processing
Colour Clustering
K-Means Colour Clustering
#Clusters=10
#Colour Plasma=10
Histogram Normalization
Colour Percentage Ratio
Estimate percentage ratio for each coloured cluster
Colour decoding
Assessment of collision avoidance based on the ratio of closed colours
Forward Collision Alert
Drivable collision-free space

Figure 6: Colour-Decoding Algorithm

Figure 7 represents the RGB input frame and the predicted depth map from our model. The nearby (yellow color) means zero meters from the viewpoint, and the far-away (blue color) indicates the 80-meter depth information from the viewpoint. Compared to the other researchers' studies, the evaluation of our model is presented in Table 1. The RMSE value defines how many pixels in the image can calculate the depth pixels. An ARD, RMSE, and RMSE (log) value of 0.194, 1.513, and 0.272 is achieved. Moreover, accuracy values of 0.706 ($\delta < 1.25$), 0.896 ($\delta < 1.252$) are better results achieved, and other metrics values are close to the state-of-the-art.

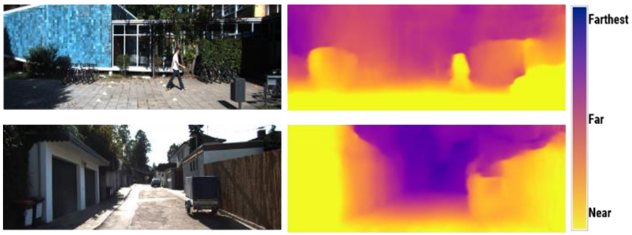


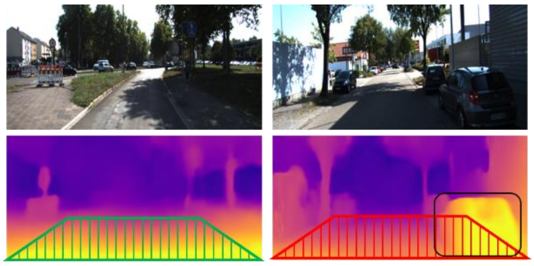
Figure 7: Input Image (left) and Predicted Depth (right)

An example of the input for the color decoding algorithm, shown in Figure 8. The green and red regions in Figure 8a & 8b are the ROIs. The ROI of Figure 8a does not include the threshold values set in the algorithm that signifies the risk of collision is significantly less. In contrast, the ROI of Figure 8b does include the threshold values set in the algorithm that shows the possibility of collision. The algorithm evaluation also applies the consecutive frames of a single video. The region matches the threshold value of the collision avoidance algorithm indicated with the black box. Figure 9a and Figure 9b show that car and cyclist are near such that collision warning generated from the predicted depth map. However, the truck and tree are not visible in the depth map (see Figure 9c and Figure 9d) still the collision warning is generated. This algorithm solely depends on the depth map's color-decoding rather than the objects present in the scene. Additionally, this algorithm evaluation also applies to a single video's

Authorized licensed use limited to: Universidad Nacional de Colombia (UNAL). Downloaded on February 10,2023 at 03:49:57 UTC from IEEE Xplore. Restrictions apply.

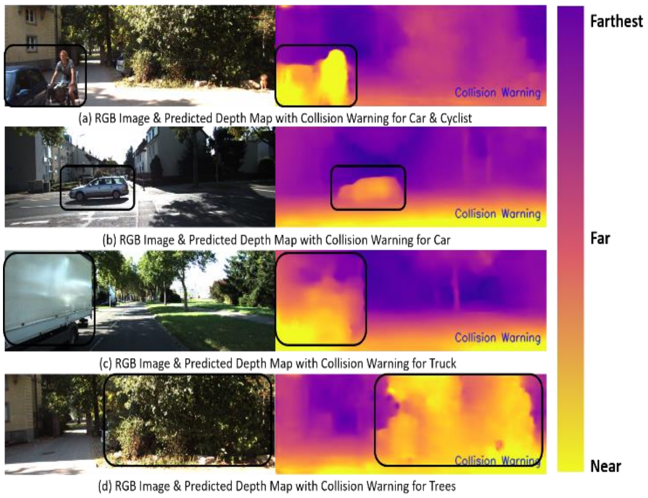
Table 1: Depth Estimation Comparison over the KITTI Eigen Test Set

Method	**Lower is better**				**Higher is better**		
	Abs Rel	**Sq Rel**	**RMSE**	**RMSE (log)**	**Acc. 1 δ<1.25**	**Acc. 1 $\delta<1.25^2$**	**Acc. 1 $\delta<1.25^3$**
Eigen et al. (Coarse) [14]	0.214	1.605	6.563	0.292	0.673	0.884	0.957
Eigen et al. (Fine) [14]	0.203	1.548	6.307	0.282	0.702	0.890	0.958
Liu et al. [15]	0.202	1.614	6.523	0.275	0.678	0.895	0.965
Zhou et. al. [7]	0.208	1.768	6.856	0.283	0.678	0.885	0.957
Ours	**0.194**	**1.513**	6.522	**0.272**	**0.706**	**0.896**	0.961



(a) No Collision (b) Collision

Figure 8: Collision Avoidance Algorithm Input



(a) RGB Image & Predicted Depth Map with Collision Warning for Car & Cyclist

(b) RGB Image & Predicted Depth Map with Collision Warning for Car

(c) RGB Image & Predicted Depth Map with Collision Warning for Truck

(d) RGB Image & Predicted Depth Map with Collision Warning for Trees

Figure 9: Collision Avoidance Algorithm Output

consecutive frames and achieved 25 FPS using RTX 2080Ti GPU.

V. CONCLUSION AND FUTURE ASPECTS

An unsupervised learning approach is employed in the presented study to estimate the depth map from a single RGB frame. Our proposed solution was able to estimate the depth map from a monocular camera with better results in terms of RMSE, Sq Rel, and its accuracy is quite close to state-of-the-art. A robust collision zone detection approach was developed based on the color-encoding of the estimated depth map. The collision avoidance approach's outcome provides the possibility of the collision. However, the transparent surfaces and occlusion regions will output invalid depth information. These limitations and results should be improved in future work.

Moreover, integrating object detection with depth estimation can be improved 3D perception of the scene to facilitate and avoid hazards in the surrounding. However, this should be addressed in future work. Furthermore, to enable cooperative avoidance systems and ensure more accurate in-vehicle route guidance systems, the thresholds should be dynamically adjusted according to vehicle speed. The ROI should also be defined based on the freely drivable space, which should be taken into account in the future.

Lightweight network architecture is extremely convenient and highly recommended on low-power devices, such as an Internet of Things (IoT) device or embedded system. Therefore, we intend to employ this approach in future work to optimize our application for outdoor navigation for visually impaired people [13]. Furthermore, it is also vital to refine the lightweight depth map model's accuracy for real-time applications.

REFERENCES

[1] Filip Mróz, and Toby P Breckon, An empirical comparison of real-time dense stereo approaches for use in the automotive environment, EURASIP Journal on Image and Video Processing, 2012(1):13, 2012.

[2] Jun-Su Kang, Jihun Kim, and Minho Lee, Advanced driver assistant system based on monocular camera, IEEE International Conference on Consumer Electronics (ICCE), pp. 55-56, 2014.

[3] Nikolai Smolyanskiy, Alexey Kamenev and Stan Birchfield, On the importance of stereo for accurate depth estimation: An efficient semi-supervised deep neural network approach, arXiv preprint arXiv: 1803.09719, 2018.

[4] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision Meets Robotics: The Kitti Dataset", The International Journal of Robotics Research, 32(11):1231–1237, 2013.

[5] R. Garg, G. Carneiro, and I. Reid, Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue, In ECCV, 2016.

[6] C. Godard, O. Mac Aodha, and G. J. Brostow, Unsupervised Monocular Depth Estimation with Left-Right Consistency, In CVPR, 2017.

[7] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, Unsupervised Learning of Depth and Ego-Motion from Video, In CVPR, 2017.

[8] S. Vijayanarasimhan, S. Ricco, C. Schmid, R. Sukthankar, and K. Fragkiadaki, SFM-Net: Learning of Structure and Motion from Video, CoRR, abs/1704.07804, 2017.

[9] Franz Andert, Gordon Strickert, and Frank Thielecke, Depth Image Processing for Obstacle Avoidance of an Autonomous VTOL UAV, ISSN 0700-4083, 2006.

[10] Uwe Franke and S. Heinrich, Fast obstacle detection for urban traffic situations, in *IEEE Transactions on Intelligent Transportation Systems*, 3(3), pp. 173-181, Sept. 2002.

[11] Yimin Wei, Huadong Meng, Hao Zhang, and Xiqin Wang, Vehicle Frontal Collision Warning System based on Improved Target Tracking and Threat Assessment, *IEEE Intelligent Transportation Systems Conference*, Seattle, WA, pp. 167-172, 2007.

[12] M. Prabha, M. Seema, P. Saraswathi, Distance based Accident Avoidance System using Arduino, International Research Journal of Engineering and Technology (IRJET), 2(7), pp. 777-780, 2015.

[13] S. Saleh, H. Saleh, M. Amin Nazari and W. Hardt, "Outdoor Navigation for Visually Impaired based on Deep Learning", in Proceedings of the 6th International Conference Actual Problems of System and Software Engineering, Moscow, 2019, Vol-2514, pp. 397-406.

[14] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In Advances in Neural Information Processing Systems, 2014.

[15] F. Liu, C. Shen, G. Lin, and I. Reid. Learning depth from single monocular images using deep convolutional neural fields. IEEE transactions on pattern analysis and machine intelligence, 38(10), pp. 2024–2039, 2016.

Authorized licensed use limited to: Universidad Nacional de Colombia (UNAL). Downloaded on February 10,2023 at 03:49:57 UTC from IEEE Xplore. Restrictions apply.