A Monocular Camera Depth Estimate Approximation using Deep learning

Rajanna
REVA Academy of Corporate Excellence-RACE
REVA University
Bangalore, India
rajannak.ai01@race.reva.edu.in

Rashmi Agarwal
Reva Academy of Corporate Excellence-RACE
REVA University
Bangalore, India
0000-0003-1778-7519

Jay Bharateesh Simha
Reva Academy of Corporate Excellence-RACE
Reva University
Bangalore, India
jb.simha@reva.edu.in

***Abstract*— In numerous applications, such as collision detection, Robotic handling, Robotic-based manufacturing facility, and Advanced Driver Assistance Systems (ADAS), depth estimation is crucial and the most significant task. Radar, Ultrasonic, Lidar technologies both operate by reflecting radio or sound wave or laser beams respectively. Stereo cameras used for depth estimation are costly and increase the cost of the system. There are a few Monocular camera depth estimate approaches that have evolved using mathematical calculation. One approach uses pixel to depth estimate mapping and the other uses the geometry of the road, the contact point of the vehicle on road, and camera properties. The proposed solution would implement depth detection using a monocular camera with deep learning. The objective of the proposed solution is to detect the depth from a monocular image and calibrate it with the actual depth. The pixel to distance data which derived from camera properties is used to run through varied hidden layers and nodes to conclude the implementation using Deep Learning (DL). Evidently, train and test data learn and converge more quickly on the deeper models which has more hidden layers and nodes.**

Keywords— ADAS, Lidar, Radar, Ultrasonic, Deep Learning, Monocular camera, Depth estimate

I. Introduction

For depth estimate, the most prevalent methods nowadays are based on Radar, Lidar, Ultrasonic, or a combination of these technologies. The radar technology measures metal object reflections and accounts for the Doppler effect to determine relative speed. The Lidar systems estimate distance using laser beams. Radar systems are generally relegated to high-end vehicles due to their prohibitively expensive price tags. Using stereo cameras, the conventional method of depth estimate [1] is limited in range and expensive. Depth estimation using a single camera is challenging because of distortions in the image caused by the camera's perspective. A monocular camera method exists to estimate an item's size, however, it is not very precise and has a tendency to estimate an unknown-sized object incorrectly. A significantly more precise estimate can be derived by considering the road geometry and the vehicle's contact point with the ground. The flatness of the road and the camera's optical axis being perpendicular to the road are prerequisites for this proposed method. In this research, the pixel to distance data which derived from camera properties is used to run through varied hidden layers and nodes to conclude the implementation using deep learning.

II. Literature review

Positive findings have emerged from recent studies of how artificial intelligence can be used in fields like image processing, driverless vehicles, and robotic manufacturing. One study [2][3] discusses how to take depth measurements using a single camera. Train a system to recognize a mapping between an object's pixel height and physical distance by observing and taking advantage of the correlation between these two quantities. The height of each test object in the image is then calculated by using this mapping. Few studies [4]-[6] discuss a further monocular vision system technique for determining range and measuring range. Since only one camera is used to gather data, all distances must be approximated in terms of perspective. Both the size of the truck in the photograph and its location near the bottom of the frame could be indicators. A width-based range estimate has a 30% margin of error because the width of an unidentified vehicle (car, van, truck, etc.) could be anywhere from 1.5 meters to 3 meters. Using the geometry of the road and the vehicle's contact point with the road, a far superior solution was provided. The assumption was a road surface that is flat and a camera with a parallel optical axis to the road surface. In [5], [7], [8] for example, it is suggested that a single forward-facing camera installed on the dashboard be used to perform both in-path and oblique distance calculations.

Deep learning is a subclass of machine learning that comprises neural networks with three or more layers. These neural networks aim to imitate the behavior of the human brain, albeit their ability to "learn" from massive amounts of data falls far short of that of the human brain. The first advantage of deep learning over machine learning is that feature extraction is redundant. Artificial neural networks utilizing deep learning do not require the feature extraction stage. Various studies showed promising results in applying mathematics estimation behind the depth perspective but not many employ deep learning [8] [9]. Therefore, this study addresses this research gap.

III. Problem statement

The depth estimation technique helps drivers avoid collisions using only one of the cameras often found in modern vehicles. Many different techniques have been devised to help drivers keep a safe distance between their host car and other vehicles on the road. Some approaches make use of the correlation between the actual height of an object and its pixel size. Making use of this connection by training a system to find a mapping between the height of a pixel and the actual distance to an object. The height of each test object in the image is then calculated by using this mapping. Also, a flat road, a perpendicular visual axis to the road, and camera properties are considered. All the methods use a mathematical calculation to estimate the depth. None of them are using contemporary adaptations of deep learning and computer vision models.

IV. Objectives of the Study

Deep learning is the latest trend in the industry and gives breaking results in various fields may it be Text analytics, Natural language processing, computer vision, and so on. DL approach in depth estimation will lead way for fast adaption of the latest models and further improves the potential of industry acceptance. The primary objective of this study is to implement the Depth estimation approximation of a monocular camera system using DL model. This enables the migration of the legacy system to DL model which helps better integrate to AI ecosystem overall. The secondary objective of the study helps in the selection of an optimal DL model to balance the complexity of the model and the cost associated with a dataset of applications that you are intended to use.

V. Project Methodology

Height, focal length, camera tilt, and pixel resolution are all characteristics of cameras. It is important to note that each camera's settings are specific to that camera. The acquired image is in the two-dimensional plane, and the distance can be estimated by determining the relationship between the two-dimensional image and the real three-dimensional view of the image. If an accident is avoided, it will be because the camera was mounted to the automobile and determined the exact distance of the obstacle. Under the assumption of a flat road, a camera is installed so that its optical axis is perpendicular to the road. If we place a point on the road, *Z* meters in front of the camera, that spot will be at a certain vertical position in the image captured by the camera as shown in Fig. 1.

An image plane *(I)* and pinhole *(P)* separated by a focal distance *(f)* make up the schematic pinhole camera shown in Fig. 1. The camera is mounted on the car *(A)* at a certain height *(H)*. The rare end of the car *(B)* is off to the side, at a distance of *Z1*. When viewed from above, the image plane is intersected by the vehicle's point of contact with the road *(Y1)*. Neither the f-number nor the Y-coordinate of the image, which are typically expressed in pixels, are to scale in this illustration.

When a further away car *(C)* makes road contact *(Z2)*, its contact point will be projected on its image plane at a lower *(y2)* location than *(y1)*.

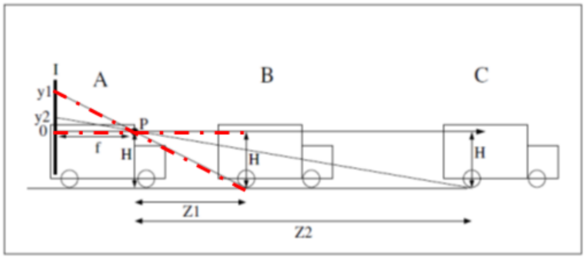

Fig. 1. Schematic diagram of the imaging geometry

The similarity of triangles allows for a straightforward derivation of Equation (1).

$$\frac{Y}{f} = \frac{H}{Z} \tag{1}$$

A. *Flowchart of the proposed methodology*

The workflow in the new approach is shown in Fig. 2. In the new approach, DL is used to compute the depth estimation approximation.

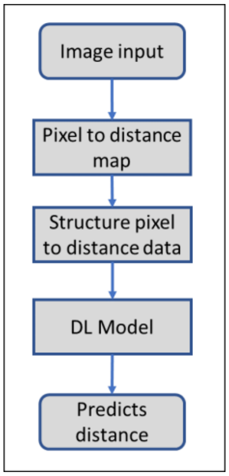

Fig. 2. Proposed methodology

B. *Dataset*

Data for the proposed implementation is simulated using camera parameters like focal length, vertical view angle, horizontal view angle, and pixel density in both vertical and horizontal. The Centre of the typical image as in Fig No. 3 used for data extraction and tabulated in TABLE I. Actual image size is 1280 x 720 pixels and since the 640 x 360-pixel point is the farthest vehicle can move and there is no depth information beyond this point. Also, Oblique distance calculation is represented in Fig. 4 and Equation (2).

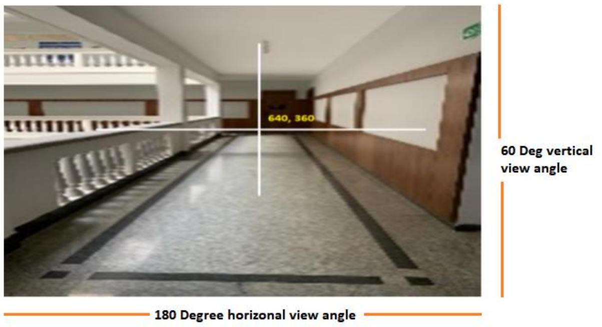

Fig. 3. Pixel to distance calculation

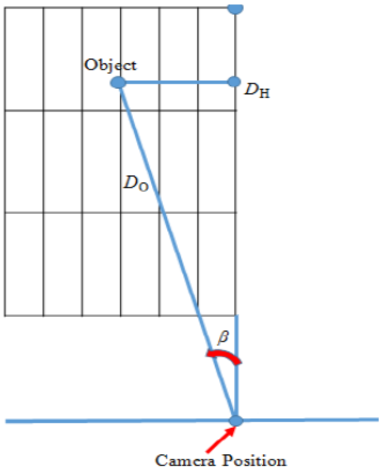

Fig. 4. Oblique distance calculation [7]

$$Do = D_H / Cos(\beta) \tag{2}$$

Do is the oblique distance, D_H is the straight-line distance from the camera position and *Cos(β)* is the cosine of angle *β* to obtain the oblique distance.

TABLE I. INPUT AND TARGET VARIABLES FOR MODELING

Properties	Input variables
Input	Horizontal pixel position
	Vertical pixel position
Output	Depth estimate

C. *Deep learning model configurations*

Many simple to decent complex model has been implemented in order to arrive at the optimal configuration. Four different model configurations as listed in TABLE II are tried, and a comparison is done to arrive at the final solution. Note that the proposed configuration is subjected to train and test datasets split of 70% and 30% respectively. Mean absolute Error (MAE) and Root Mean Squared Error (RMSE) is chosen as loss functions with metric as accuracy. Adam's algorithm is chosen as an optimizer.

TABLE II. MODEL CONFIGURATIONS FOR TEST AND VALIDATION

	No. of nodes at layer 1	No. of hidden layers	No. of nodes in the hidden layer
Model 1	13	1	10
Model 2	13	4	10
Model 3	50	1	100
Model 4	50	4	100

VI. RESULT AND ANALYSIS

Application of different varied DL model configurations performed on dataset and validation metrics of the models are noted to understand the best fit model for a particular design in Table III. Model 1 is taken as the baseline model and all other models are compared with the baseline model. It can be observed that Model 2, MAE and RMSE are 75% less than the baseline model. Similarly, additional layers and neurons will reduce this to almost zero in Model 4, but the cost of building the complex model is very high. Hence, we recommend Model 2 for less critical applications like the one that has been tested.

TABLE III. TRAIN AND TEST METRICS FOR VARIES MODEL CONFIGURATIONS AT EPOCH 25

	MAE		RMSE	
	Train	***Test***	***Train***	***Test***
Model 1	2.0245	1.8609	2.6184	2.4958
Model 2	0.5060	0.5226	0.6550	0.6694
Model 3	0.1605	0.1675	0.2327	0.2494
Model 4	0.0623	0.0630	0.0837	0.0839

VII. CONCLUSION

The results from the previous section indicate that deeper and more nodes in the model, train and test converges will happen faster. Model 2 being the best choice for the dataset for depth estimate that we have proposed.

The end result demonstrates a single-camera Depth Estimation algorithm. One can equip car with this device to help you avoid collisions. To avoid crashes, a safe distance between the host vehicle and other vehicles must be maintained; The proposed system assesses this distance. With a higher-resolution camera, the same method can be applied to considerably greater distances. Additionally, the approach can be improved for steep grades and winding routes.

REFERENCES

[1] S. Kumar, D. Gupta, and S. Yadav, "Sensor Fusion of Laser & Stereo Vision Camera for Depth Estimation and Obstacle Avoidance," ©2010 Int. J. Comput. Appl., vol. 1, no. 26, pp. 975–8887.

[2] A. Rahman, A. Salam, M. Islam, and P. Sarker, "An image based approach to compute object distance," Int. J. Comput. Intell. Syst., vol. 1, no. 4, pp. 304–312, 2008, doi: 10.1080/18756891.2008.9727627.

[3] N. Yamaguti, S. Oe, and K. Terada, "A Method of Distance Measurement by Using Monocular Camera.

[4] I. Gat, M. Benady, and A. Shashua, "A monocular vision advance warning system for the automotive aftermarket," SAE Tech. Pap., 2005, doi: 10.4271/2005-01-1470.

[5] A. Joglekar, D. Joshi, R. Khemani, S. Nair, and S. Sahare, "Depth Estimation Using Monocular Camera," Int. J. Comput. Sci. Inf. Technolgies, vol. 2, no. 4, pp. 1758–1763, 2011.

[6] E. Dagan, O. Mano, G. P. Stein, and A. Shashua, "Forward collision warning with a single camera," IEEE Intell. Veh. Symp. Proc., no. July 2004, pp. 37–42, 2004, doi: 10.1109/ivs.2004.1336352.