A wide-baseline stereo object detection system

Abstract

Light Detection and Ranging (LiDAR) and stereo vision are two commonly used technologies that provide 3D sensing of surrounding traffic actors. LiDAR is expensive, and its powerful laser illuminating pulse poses safety concerns for human eyes and sensitive electronics [1]. Additionally, it is a shortcut that sidesteps the fundamental problems of visual recognition, unable to distinguish pedestrians and inanimate objects, needed by complete autonomous driving. As an attractive substitute for LiDAR, stereo vision is cost effective, uses passive illumination which is harmless to retina and sensitive electronics, and is capable of visual recognition. Yet, its depth accuracy is inferior to that of LiDAR because of stereo vision's limited baseline and erroneous photometric alignment between the pair of images.

In this research, we develop a deep learning based correspondence algorithm to tackle the alignment issue and show that the wide-baseline stereo system significantly outperforms a LiDAR using a synthetic dataset generated by CARLA [4].

Keywords: Stereo depth perception, Autonomous Driving, Object Detection (Faster-RCNN), CARLA Simulator

Research Objective

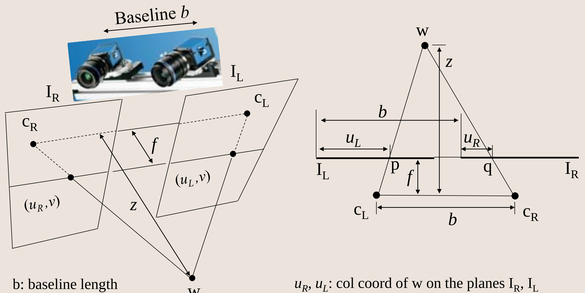
Develop a stereo vision system on a par with LiDAR

- Illustrate narrow baseline limits the accuracy of stereo vision system
- Use recent AI breakthrough (i.e., deep learning) to overcome the limiting barrier (i.e., stereo correspondence ambiguity)
- Show the wide baseline stereo system outperforms a LiDAR system
- Prove the concept using a synthetic dataset

Stereo Depth Perception

$$\delta z = -\frac{z^2}{fb}\delta d$$

Depth error δz is inversely proportional to baseline b. 6x baseline means 6x reduction in depth error!



b: baseline length
w: world point
I_R, I_L: left and right image planes
f: focal length
c_R, c_L: optical centers
v: row coord of w on the planes I_R, I_L

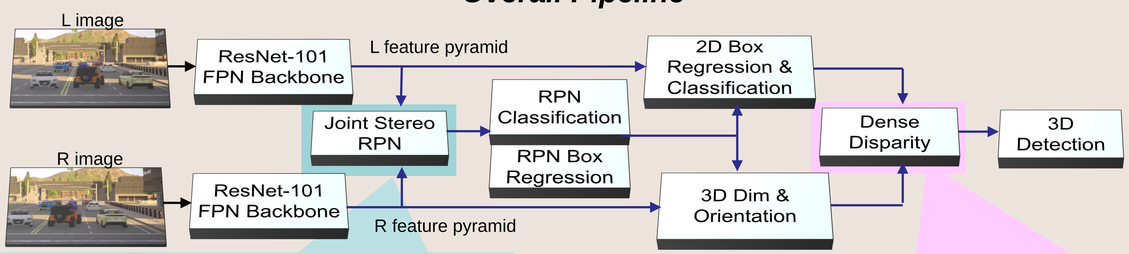
u_R, u_L: col coord of w on the planes I_R, I_L
z: depth of world point
δz : measurement error of depth
d: disparity d = $u_L - u_R$
δd: measurement error of disparity
p,q: pixels of w on planes I_R, I_L

$$\Delta wpq \approx \Delta w c_L c_R \Rightarrow \frac{z-f}{z} = \frac{b-u_L+uR}{b} \Rightarrow z = \frac{fb}{u_L-u_R} = \frac{fb}{d} \Rightarrow \delta z = -\frac{z^2}{fb}\delta d$$

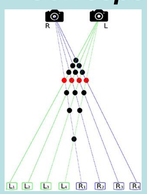
the depth of the scene point (w) is inversely proportional to the disparity. Taking the derivative, we arrive the main equation with a small rearrangement of variables.

Method Innovation

Overall Pipeline

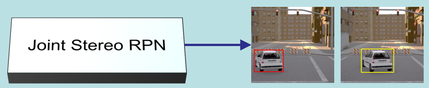


Correspondence ambiguity issue



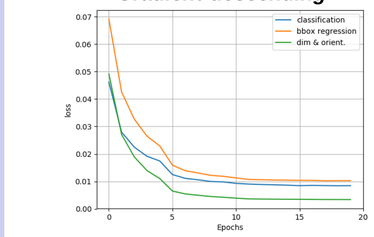
- The wider the baseline, the larger the disparity, which needs a larger search span in image, causing it to be computationally expensive
- Featureless zone

Stereo correspondence learning



- RPN (region proposal network) directly outputs left and right region proposal pair (i.e., red-yellow boxes) from the concatenated left and right feature pyramid.
- RPN uses the concatenated L and R feature pyramids (semantic context) to infer correspondence at object level

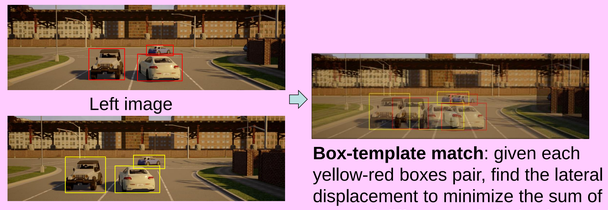
Gradient descending



Wide baseline to amplify disparity

Baseline = 0.3 m Baseline = 1.6 m

Dense disparity with subpixel accuracy



Box-template match: given each yellow-red boxes pair, find the lateral displacement to minimize the sum of squared difference of pixel values

Two-step training with synthetic datasets

Step 1: Training the network using KITTI dataset of 0.54 m baseline

Step 2: Fine-tuning the network using synthetic dataset of with different baselines (i.e., 0.3m, 0.8 m, and 1.6 m)

Use 500 synthetic training scenes to tune the weights of the neural network for the different baselines

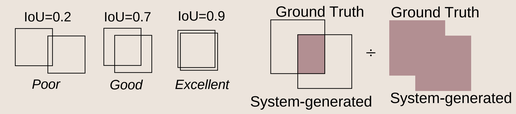
Illustrative Results

Data: The synthetic dataset contains 1000 scenes generated from CARLA [4] with average of 8 vehicles per scene. 500 scenes were used for fine-tuning while the other 500 scenes are for testing. Four samples of the results are illustrated as below:

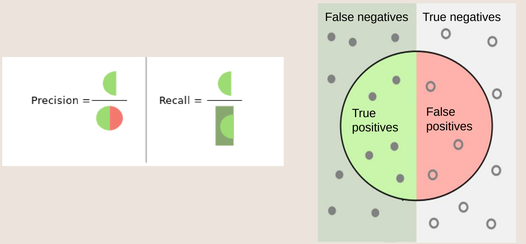
Evaluation & Comparison

1) Intersection over Union(IoU) used for bounding box accuracy

$$\text{IoU} = \frac{\text{Area of Intersection}}{\text{Area of Union}}$$



2) Average recall (AR) and average precision (AP) used for classification accuracy [https://en.wikipedia.org/wiki/Precision_and_recall]

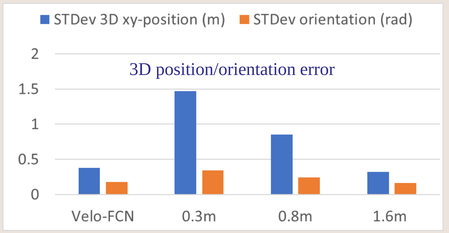


3) Standard deviation (STDev) used for the accuracy of vehicle 3D position and orientation angle

Systems	Sensor	AR (%)	AP (%) IoU_{2D}=0.5	AP (%) IoU_{3D}=0.5	STDev 3D xy-position (m)	STDev orientation (radian)
Velo-FCN[5]	LiDAR	*45.61*	n/a	55.26	0.38	0.180
Faster R-CNN [2]	Left mono	***72.23***	*71.54*	n/a	n/a	n/a
Ours @ 0.3m baseline	Stereo	*72.10*	*75.38*	50.97	1.47	0.340
Ours @ 0.8m baseline	Stereo	*71.32*	*74.61*	57.74	0.85	0.241
Ours @ 1.6m baseline	Stereo	*70.95*	***75.94***	**65.18**	**0.32**	**0.161**

* Above results are reported using the 500 synthetic testing scenes

- **The wide baseline of 1.6m outperforms the LIDAR system [5] in all categories**



Conclusion

We successfully developed a stereo system with wide baseline in CARLA [4] simulator and

- Proved our wide-baseline stereo system outperforms a LiDAR system [5] using the synthetic dataset of 1000 images
- Verified analytically and experimentally that the depth accuracy is inversely proportional to the stereo baseline
- Developed a neural network and overcame the correspondence barrier
- Showed the robustness in low lighting scenarios

References

[1] Jeff Hecht, https://spectrum.ieee.org/, Can Lidars Zap Camera Chips?
[2] S. Ren, et al., NIPS 2015, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks
[3] A Geiger, et al., The International Journal of Robotics Research 32(11) 1231–1237, 2013, Vision meets robotics: The KITTI dataset
[4] A. Dosovitskiy et al., arXiv:1711.03938, 2017, CARLA: An Open Urban Driving Simulator.
[5] B. Li, et al. In Robotics: Science and Systems, 2016. Vehicle detection from 3d lidar using fully convolutional network.
[6] L. Steffen et al, Front. Neurorobot., 28 May 2019, Neuromorphic Stereo Vision: A Survey of Bio-Inspired Sensors and Algorithms