Monocular 3D Object Detection for Traffic Analysis

Himanshu Pahadia, MS Computer Science
Mentor: Dr. Yezhou Yang, Assistant Professor
School of Computing and Augmented Intelligence, Ira A. Fulton Schools of Engineering, ASU

Introduction

Monocular 3D object detection predicts 3D bounding boxes using a single monocular, typically RGB image. Identifying 3D bounding boxes is a difficult task as RGB images lack critical depth channel information.

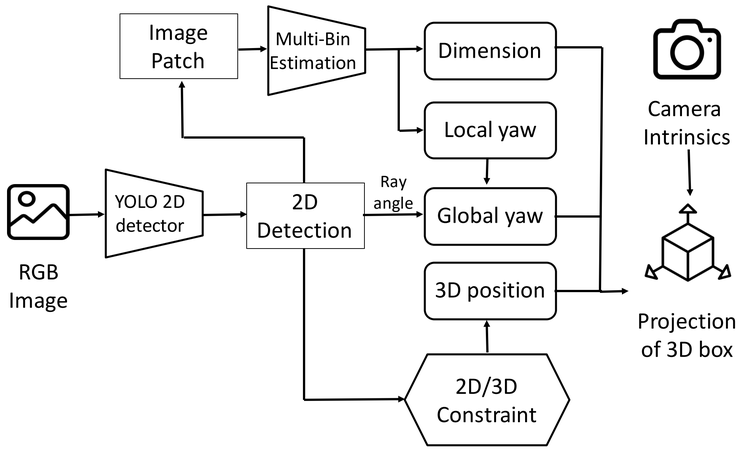
Motivation

- Alternatives systems include LIDAR sensors that are expensive and sensitive to adverse weather.
- 3D bounding box regression can help in more accurate traffic study - analysis, data archiving, speed estimation, reconstruction, etc.

Contributions

- System to regress 3D bounding boxes of vehicles and pedestrians on the road in real-time.

Network architecture



Methodology

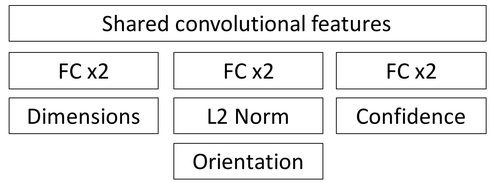
The system is divided into 2 main parts: 2D Object Detection & 3D Detector.

2D Object Detection

- We have used a single-stage feed-forward architecture of YOLO for the 2D object detection module.
- Features - skip connections, no pooling, and 3 prediction heads (processing at different spatial compressions).

3D Object Detection

- Architecture of Multi-Bin estimation takes in the cropped image patches (resized to 224x224) and outputs – orientation (local yaw) and Dimension of objects.



Multi-Bin Estimation Architecture

Orientation estimation

- Assumes pitch and roll to be zero as objects lie on a plane, only yaw needs to be estimated.
- The cropped patch may show a change in orientation while the vehicle is going straight. So, network is trained on local angle (α).

Dimension estimation

- Regression for dimensions prediction is performed wrt average, which is computed class-wise based on the training set.
- 2D box and VGG-19 extracted features are passed to the branch and outputs – Height, Width, and Length. Uses L2 Loss.

2D/3D Constraints

- Given the pose of the object in the camera coordinate frame (R, T) ∈ SE(3) and the camera Intrinsics K, the projection of a 3D point $X_o = [X, Y, Z, 1]^T$ in the object's coordinate frame into the image $x=[x,y,1]^T$ is: $x = K [R \ T] X_o$

Experiments and Results

Dataset - KITTI 3D object detection dataset consists of 7481 training images and 7518 test images with 3D annotations.

ARGOS Vision camera

Metric important for use-case – Inference time

Evaluation metric - Intersection over Union (IoU) and Average Orientation Similarity (AOS)

Per pose inference time PPIT (Titan-X) -

- PPIT KITTI – 0.112 s
- PPIT ARGOS – 0.098 s

Conclusion & Future Work

- We present a deep learning architecture that regresses 3D bounding boxes of objects from monocular RGB images.
- The key idea is that the perspective transformation of a 3D bounding box should fit tightly in the 2D bounding box

Work in progress and future work

- Transforming UA-DETRAC 2D object detection dataset into a 3D detection dataset. Testing the system with IoU and AOS.
- Training the network on the dataset and testing inference time on the ARGOS Vision camera.
- Swap the Yolo 2D detector with Tiny-Yolo, SqueezeDet.
- Swap the 2D/3D constraint flow with a deep neural network for location estimation of the 3D box' central point.

References

- Arsalan Mousavian et al. "3d bounding box estimation using deep learning and geometry"
- Geiger, Andreas, Philip Lenz, and Raquel Urtasun. "The kitti vision benchmark suite."

MORE
Masters Opportunity for Research in Engineering

Ira A. Fulton Schools of Engineering
Arizona State University