Single view depth prediction
Amro Asali

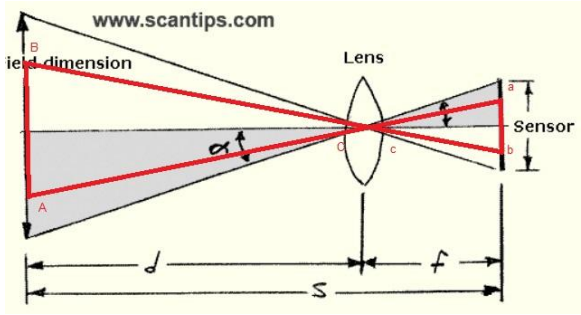
Abstract--in this project i use a single motionless image of a vehicle and estimate it's distance from the camera in meters assuming that i know the height of the vehicle which can be easily guessed for example the average height of an SUV is 1.5-1.7 meters but for the purpose of this project i measured the SUV's height , the camera's focal length and the sensor height are known .

Background work-- from a previous research [1] it has been obtained that $d = \frac{f*C*H}{S*O}$

While d is the depth that we want to obtain f is the focal length of the camera in Millimeters C is the height of the object in meters H is the height of the image in pixels S is the sensor's height in Millimeters and O is the height of the object in pixels

Methodology-- i used the method of bounding box regression using NN for finding the height of the vehicle in pixels , the NN is consisted of the body of VGG16 trained on imagenet dataset and used a new head which was trained on a 360 pictures dataset taken from caltech 101 dataset [2] where the labels are the top left and bottom right corners coordinates to localize the vehicle in the image and extract it's height in pixels and use that to figure out the depth using the formula

Background work--[3] the formula $d = \frac{f*C*H}{S*O}$ can be derived using triangular similarities



It's easy to see that ABC is similar to abc , let h=|ab| , H=|AB| so we get that

$$\frac{h}{H} = \frac{f}{d} \iff d = \frac{f\,H}{h}$$

And it holds that

$h_{mm} = \frac{O_p * S_{mm}}{S_p}$ while O_p is object height in pixels , S_{mm} is sensor height in mm , S_p is sensor height in pixels (image height)

So we substitute h and get the formula

Results-- here i show the estimated depth of using this method on 4 pictures from different distances from the car compared to the ground truth , i used a oneplus 5T for taking the pictures so the parameters are f=4.1 mm , s=5.22 mm , H is 480 in the first 3 and 240 in the 4th picture , C is 1.57.

	Estimated value	Ground truth
Image 1	4.45 m	4.5 m
Image 2	3.97 m	4 m
Image 3	4.93 m	5 m
Image 4	2.17 m	1.5 m

We can see in the first 3 pictures the results are accurate since the estimated height in pixels is close to the ground truth as shown in the pictures , in the 4th picture the bounding box's height is less than it should be making the denominator smaller than it should be resulting in a higher depth.

Failed results-- due to using a small dataset with more images where the car is close than far , the neural network fails to localize the vehicle appropriately thus resulting in a wrong estimate of the vehicles height in the image , using a larger dataset with more examples of cars further than 5 meters should solve this problem but since i labeled the images by hand and i already got good results for images in the range 2.5-5 meters it didn't seem necessary to do this, because this project is supposed to be a proof of concept and not a full research.

Conclusions and future work-- in this project i tackled the problem of depth estimation and object localization and as shown in results i was able to localize a vehicle in an image using a neural network and estimated it's distance from the camera successfully (with a reasonable error) as long as it was 1.5-5 meters far , using a larger more diverse dataset one should be able to estimate depth accurately at a larger range of distances and one could make a depth map for all objects in the image.

Resources --

[1] depth estimation in still images and videos using a motionless monocular camera
(PDF) Depth estimation in still images and videos using a motionless monocular camera (researchgate.net)

[2] caltech 101 dataset
Caltech101

[3] math of (FOV) scantips.com
The Math of camera Field of View Calculations (FOV) (scantips.com)