Scene Representations from Focal Stack for Depth Estimation

Logan Bruns
Stanford University
lbruns@stanford.edu

1 Motivation

Humans are fairly good at estimating depth and relative depth of a scene using a number of different cues as well as prior knowledge about the physical world.

Understanding the depth of a visual scene has a variety of useful applications including augmented reality, visual editing, data capture, and robotics. With the ubiquity of mobile devices which are capable of creating focal stacks a system to create 3d depth estimations using mobile phone cameras would allow for all kinds of applications with existing hardware.

Focal stacks already provide some depth information by way of knowledge of the camera optics and examining which part of the scene is sharpest at which focal setting. However, this often requires a deep stack and significant computation time to create a high quality depth map. These approaches often also have problems with surfaces that do not have very high contrast detail.

The notion here is to see if it is possible to use deep learning to introduce some prior knowledge to the depth estimation process. Much as human recognizes objects like boxes or spheres and understands the relationships of their surfaces it may be possible to train a model to do the same and there by improve depth estimation. Improve it either in terms of quality of runtime for common cases.

2 Related Work

Suwajanakorn1 et al [Suwajanakorn et al.(2015)Suwajanakorn, Hernandez, and Seitz] did some impressive work in this area in their paper "Depth from Focus with Your Mobile Phone". Their work which did not use deep learning focused on using many images and optimization to push the limits of the optical information encoded in the focal stack. It looks like the code is available. Depending on level of bit rot and complexity to set up it may make for a good baseline.

Eslami et al [Eslami et al.(2018)Eslami, Rezende, Besse, Viola, Morcos, Garnelo, Ruderman, Rusu, Danihelka, Gregor, Reichert, B explored using neural scene representations and deep learning to learn a representation of the scene that encoded some prior knowledge of the spatial relationships. Their approach of training the model to optimize a mental rotation task to force the model to learn to encode spatial relationships then using these representations for other supervised tasks depending on spatial relations is partial inspiration for this project.

3 Project Overview

The proposed approach is to somewhat like Eslami et al [Eslami et al.(2018)Eslami, Rezende, Besse, Viola, Morcos, Garnelo, Ruder to pretrain a model to create neural scene representations that effectively encodes knowledge about depth of field and its impact on common objects in common scenes. Then use this pretraining to train a model to estimate a depth map for a scene.

Multiple loss functions and features will be considered and tried but the starting point for the pretraining will be to give the model several images from the focal stack and have it predict an image from the focal stack that it has not seen. Essentially asking the model to learn to predict depth of field effects on the image. To force a compact scene representation during this process there will be a bottleneck layer in the neural scene

model. This bottleneck layer's output will be the embedding that is used for the supervised model to predict the depth map.

To produce the focal stacks two datasets are being considered. Both datasets have raw images as well as high quality depth maps. The idea is to use an optical transform to create a synthetic focal stack from each of these images to use for training, validation, and evaluation. The two datasets are NYU's "Indoor Segmentation and Support Inference from RGBD Images" [Nathan Silberman and Fergus(2012)] dataset and Matterport's "Matterport3D: Learning from RGB-D Data in Indoor Environments" [Chang et al.(2017)Chang, Dai, Funkhouser, Halber, Niessner, Savva, Song, Zeng, and Zhang]. NYU's dataset is available for immediate download. Matterport's dataset requires asking permission.

Some thought may be needed to evaluate the appropriate metrics to use for the quality of the resulting depth maps. MSE may not be as useful as measure how well it does in relative fashion by object. Namely the precise depth is less important than knowing that this object is much farther away than that object. Also that the distances within an object are consistent with its shape.

4 Milestones, Timeline and Goals

Week of 2/10 Obtain datasets and setup data loading pipeline

Week of 2/17 Synthetic focal stack creation, initial model and baseline (maybe like HW)

Week of 2/24 Model iteration, figure out better quality metrics than training loss

Week of 3/2 Model iteration, draft of report and poster

Week of 3/9 Finish report and poster

References

[Chang et al.(2017)Chang, Dai, Funkhouser, Halber, Niessner, Savva, Song, Zeng, and Zhang] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. 2017. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*.

[Eslami et al.(2018)Eslami, Rezende, Besse, Viola, Morcos, Garnelo, Ruderman, Rusu, Danihelka, Gregor, Reichert, Buesing, Web
S. M. Ali Eslami, Danilo J. Rezende, Frederic Besse, Fabio Viola, Ari S. Morcos, Marta Garnelo, Avraham Ruderman, Andrei A. Rusu, Ivo Danihelka, Karol Gregor, David P. Reichert, Lars Buesing, Theophane Weber, Oriol Vinyals, Dan Rosenbaum, Neil Rabinowitz, Helen King, Chloe Hillier, Matt Botvinick, Daan Wierstra, Koray Kavukcuoglu, and Demis Hassabis. 2018. Neural scene representation and rendering. DOI 10.1126/science.aar6170.

[Nathan Silberman and Fergus(2012)] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. 2012. Indoor segmentation and support inference from rgbd images. In *ECCV*.

[Suwajanakorn et al.(2015)Suwajanakorn, Hernandez, and Seitz] Supasorn Suwajanakorn, Carlos Hernandez, and Steven M. Seitz. 2015. Depth from focus with your mobile phone. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.