Depth Estimation from a Single Image Using a Deep Neural Network

Rawan Alghofaili

January 2015

1 Introduction

By using the intrinsic and extrinsic camera parameters, Multi-view Stereo has been applied to accurately estimate depth maps. Although this can produce impressive results, the camera parameters(e.g. focal length, disparity and baseline) are necessary for the estimation. This means that the depth can not be estimated unless we have prior knowledge about the image's origin. In addition, two slightly different views of a particular scene are needed in order to reconstruct its depth map.

As interest grows in deep neural networks and with the introduction of readily available depth sensors such as Microsoft Kinect, depth estimation from single view images has become an open and interesting research problem in the computer vision community.

2 Method

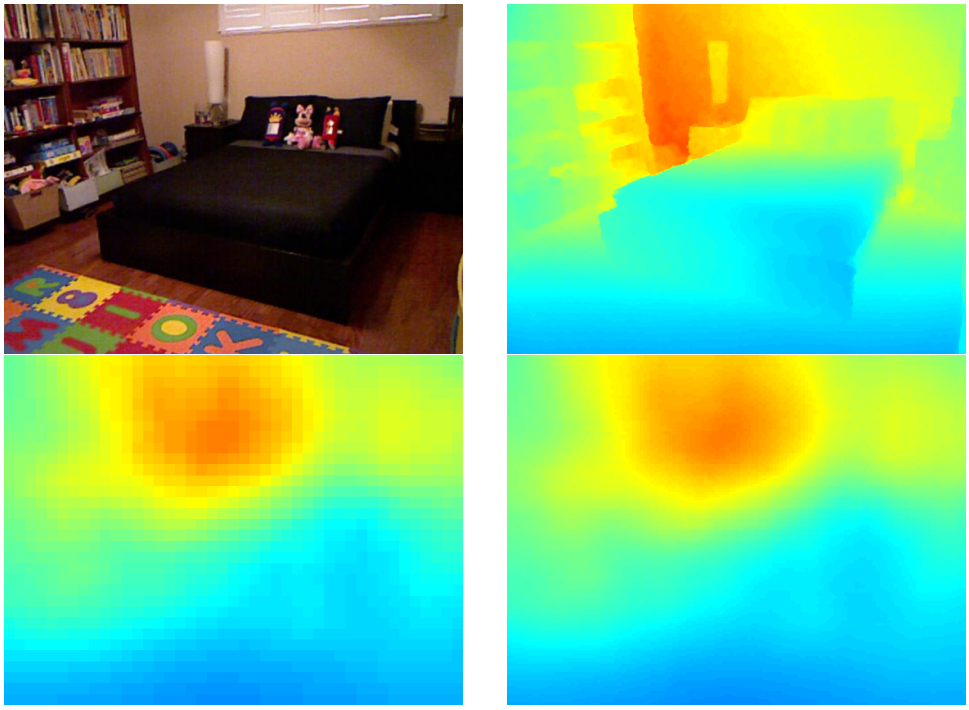

Figure 1: Results from [1]. Top left : input image, top right : ground truth, bottom left : coarse prediction, bottom right : refined prediction.

From [1] and [4] we concluded that scene information give us a good coarse depth estimate. Meaning, knowledge of the scene will provide us with global cues about the image's spatial structure such as the location of floor and ceiling edges. And by using image features extracted from the PLACES pool5 layer we can exploit scene information in estimating our depth map.

[1] [4] also deduced that further local refinements of the global estimate can give us a high accuracy depth map. As in [1] I will as well attempt to use PLACES[5] as a coarse estimate. However, I will use a deep neural network to learn the depth dictionary and transformation T instead of the coupled regression approach presented in the paper.

I will be using the CAFFE[2] framework to train, test, and update my convolutional neural network's topology.

3 Data

I intend to use the NYUv2[4] benchmark as a ground truth measure of depth while training. NYUv2 consists of depth maps of indoor scenes measured with a Microsoft Kinect.

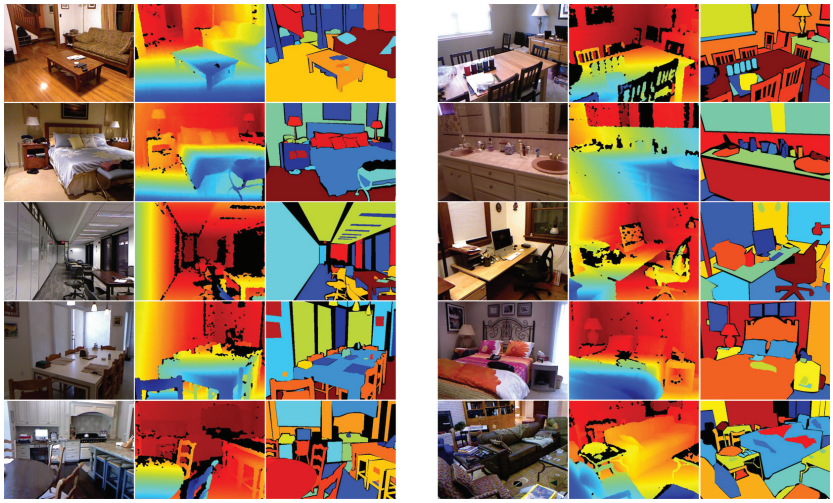

Figure 2: RGB images,their corresponding depth images and labeled images taken from the NYUv2 dataset

4 Milestone Goal

Have a functioning convolutional neural network structure that had already been finetuned to fit our problem and performs more accurately than the previous models. The next step is to experiment with multi-task learning and tweak the neural net structure accordingly.

References

[1] Baig, M., Torresani, L.: *Coarse-to-fine Depth Estimation from a Single Image via Coupled Regression and Dictionary Learning.* arXiv preprint arXiv:1310.1531, 2013.

[2] Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: *Decaf: A deep convolutional activation feature for generic visual recognition.*

[3] Seitz, S., Curless, B., Diebel, J., Scharstein, D., Szeliski, R.: *A comparison and evaluation of multi-view stereo reconstruction algorithms.* In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition. (2006).

[4] Silberman, N., Derek Hoiem, Fergus, R.: *Indoor segmentation and support inference from rgbd images.* In ECCV, 2012.

[5] Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: *Learning Deep Features for Scene Recognition using Places Database.* Advances in Neural Information Processing Systems 27 (NIPS), 2014.