Out-of-focus: Learning Depth from Image Bokeh for Robotic Perception

Eric Cristofalo and Zijian Wang, {ecristof, zjwang}@stanford.edu

Stanford University

MSL

Introduction

- Autonomous robotic systems require detailed environmental perception (e.g. 3D reconstructions) to complete challenging tasks such as object recognition, obstacle avoidance, planning, and manipulation
- Pixel depth estimation typically comes from triangulating image features, learning from single RGB images [1,2,3], or using laser rangefinders
- Instead, we employ the out-of-focus aesthetic – *bokeh* – of an image to determine a pixel-wise depth map via supervised learning using out-of-focus images from a camera with labeled RGB-D sensor information
- Current Depth from Focus (DfF) methods quantify the *focus measure* of out-of-focus images using specific focus descriptors [4]
- **Project Goal: Learn pixel-wise depth mapping from labeled *focal-stack* (sequence of images with varying *focal-depth*) using Convolutional Neural Networks (CNN)**

Data Acquisition

- Acquired 344 labeled data examples of office scenes include 52 image focal-stack and RGB-D image
- **Hardware** (Figure 1):

1) Logitech C920 USB webcam for varying focus images
2) Microsoft Kinect V2 for ground truth depth images

- Acquire data using ROS and OpenCV in C++
- Performed data filtering, alignment, and scaling in MATLAB using OpenCV (Figure 2)
- Replicated data using random 3D perspective transforms to yield **3674 data examples** for the network

Figure 1. Mobile experimental data collection setup.

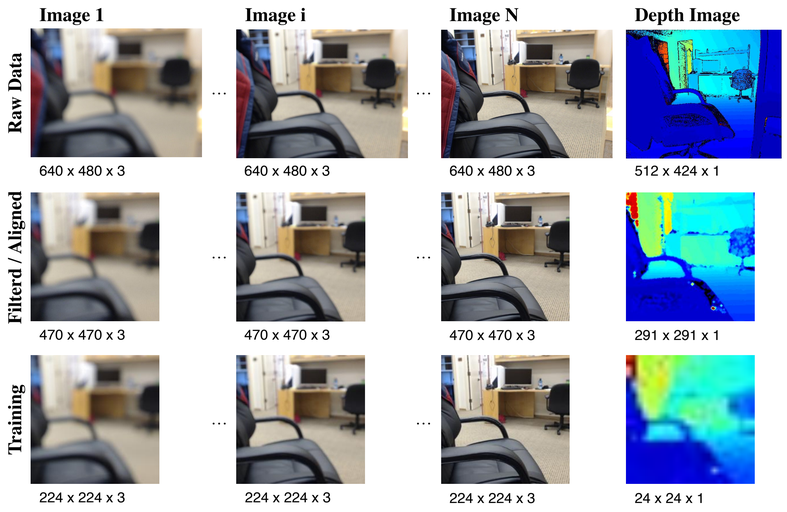


Figure 2. Example data from training set. Raw focal stack (N images with varying focal depth) and corresponding depth map (top row). Filtered and aligned data (middle row). Training images and labeled data (bottom row).

Convolutional Neural Network

- VGG-like Convolutional Neural Network (Figure 3)
- Input: 224x224 single or stacked RGB image(s). For N focal stack images input, we superpose them together and make the depth 3*N
- 6 convolution layers with stride 3x3 using ReLU activation
- 4 max pooling layers
- 2 fully connected layers
- Output: 24x24 (flattened) depth estimation image
- Loss function: mean square error for all pixel depth values

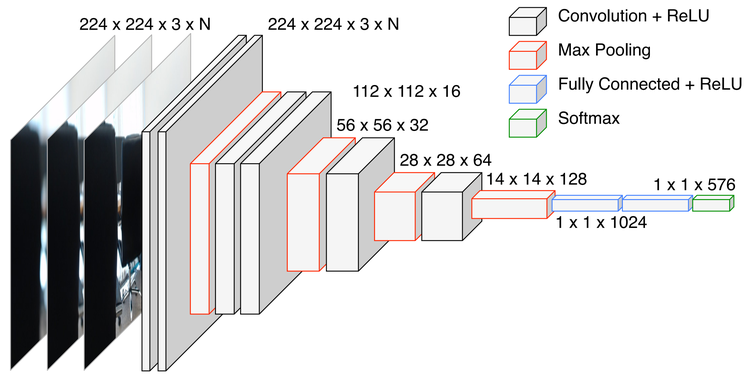


Figure 3. Convolutional Neural Network architecture using multiple input images from focal stack

Implementation & Experiments

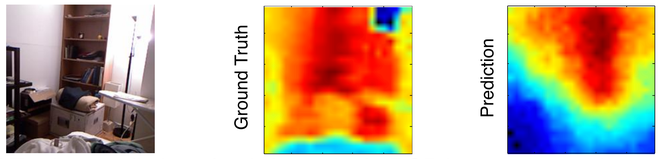
- Implemented in Keras with TensorFlow backend
- Data are split into ~ 85% training examples and 15% test examples
- Images are shuffled before training
- **[Exp1]** Training with NYU Depth V2 dataset [5]. 11670 examples (single RGB input to the neural net) in total. Took 6 hours for 30-epoch training on a 12-core, 64-gb memory workstation.
- **[Exp2]** Training with our own dataset. However, the input is just single RGB image, for the purpose of comparison with our focal stack method. Took 3.5 hours for 30-epoch training.
- **[Exp3]** Training with our own dataset (3674 examples). Each example contains 10 RGB images input taken under different focal lengths (36740 images used in total).
- Same network structure for fair comparison (see Table 1)

Table 1. Training and testing error results for all three data sets

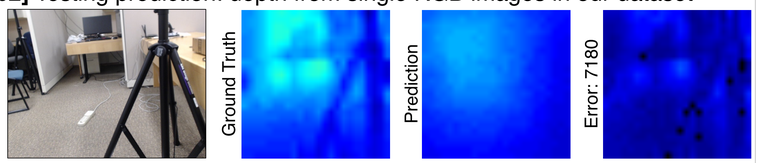
	NYU data, single RGB	Our data, single RGB	Our data, focal stack
Dataset size	11670	3674	3674
Training error (mse)	941.06	352.73	345.23
Mean abosolute error	22.93	12.92	12.79
Test error (mse)	698.99	355.27	335.88
Mean abosolute error	19.00	12.53	12.11

Results

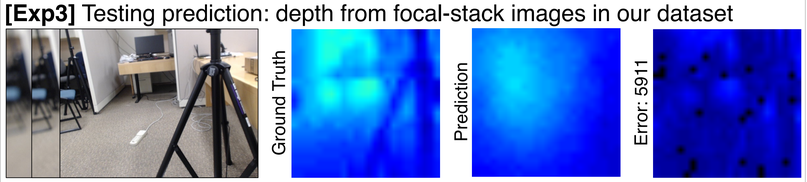
[Exp1] Testing prediction: depth from single RGB images in NYU dataset [5]



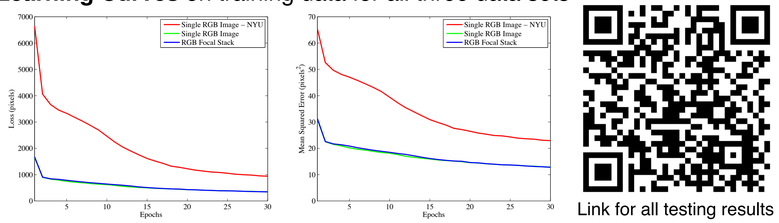
[Exp2] Testing prediction: depth from single RGB images in our dataset



[Exp3] Testing prediction: depth from focal-stack images in our dataset



Learning Curves on training data for all three data sets



Link for all testing results

Figure 4. Learning curves with training data showing the mse loss (left) and mean absolute error (middle) per pixel. Note that the final testing error for the focal stacks is slightly less than the single image data. Link to all testing results for our data set (right).

Conclusion and Future Work

- Showed CNN with focus data in images yields lower training and testing error than single RGB images
- Experimentally acquired novel data set of labeled, out-of-focus images
- Future work: Collect varied data in other environments (e.g. outdoors) with different RGB optics that contain more depth information at larger distances in the Kinect's workspace – [0.4,4] meters

Acknowledgements

We would like to acknowledge Professor Mac Schwager and the Multi-robot Systems Lab (MSL) in the department of Aeronautics and Astronautics at Stanford University.

References

[1] Liu, Fayao, et al. "Learning depth from single monocular images using deep convolutional neural fields," 2015.
[2] S. Choi, D. Min, B. Ham, Y. Kim, C. Oh, and K. Sohn, "Depth analogy: Data-driven approach for single image depth estimation using gradient samples," IEEE Transactions on Image Processing, vol. 24, no. 12, pp. 5953–5966, 2015.
[3] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view rgb-d object dataset," ICRA 2011
[4] S. Pertuz, D. Puig, and M. A. Garcia, "Analysis of focus measure operators for shape-from-focus," *Pattern Recognition,* 2013.
[5] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," ECCV, 2012.