Data-driven Depth Inference from a Single Still Image

Kyunghee Kim
Computer Science Department
Stanford University
kyunghee.kim@stanford.edu

Abstract

Given an indoor image, how to recover its depth information from one single image? This problem has been studied before for many years. But previous research mainly focused on using manually designed features, heuristics, or structure information. Lacking enough training data limits the methods that can be used to deal with this problem. However, with Kinect, it is now much cheaper to get ground truth depth information for indoor images. The purpose of this project is to use a lot of training data to obtain a more data-driven approach for recovering depth information given a single image.

1. Introduction

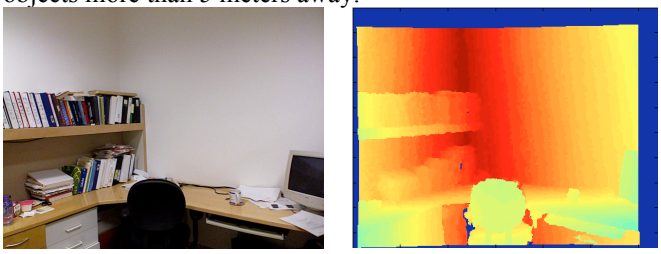
Depth estimation from images and reconstruction of 3D structure of the images has been of interest to computer vision researchers for many years. Saxena et al. [1][2] used Markov Random Field (MRF) to model the depths and relation between depths at different parts of the image. Scharstein and Szeliski [3] produced a dense disparity map using two-frame stereovision. Torralba and Oliva [4] proposed a way to obtain the properties of the structure in the image from Fourier spectrum and infer the depth from this information. Saxena, Chung, and Ng [5] inferred depth from monocular image features. This project will use a MAP-MRF approach similar to [1], [2] and [6] and use massive amount of indoor images collected with Kinect [7] to infer the depth from a single image.

2. Experiment

We will formulate the problem as an energy minimization problem as in [1], [2], [6] and [8] and before writing an energy function which consists of the unary term that models the relationship between the features in each pixel to the depth information and the pair-wise term that models the relationship between two neighboring pixels and depth information, we run some preliminary experiment to examine the properties of the unary term.

2.1 Data

The images were collected with Microsoft Kinect [7] RGB camera and depth camera that contains the indoor images with 4 scene categories i.e., office, kitchen, bedroom, and living room. We collected 200 RGB images of these indoor sceneries and each RGB image has a corresponding ground truth depth image created with the Kinect depth camera. Since the Kinect depth camera measures the depth information accurately within ~5 meters, indoor images seem to be more proper for this experiment than out door images that usually can have objects more than 5 meters away.



(a) (b)

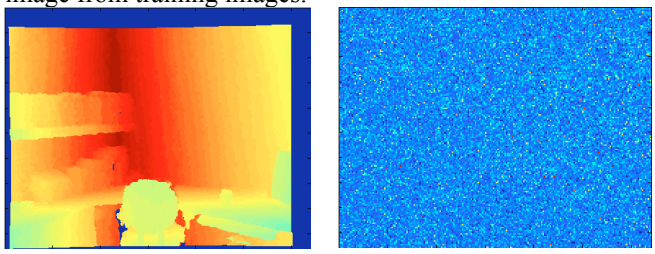
Figure 1: **One sample image.** The image on the left is an RGB image of 'office' scene category and the image of the right is the corresponding ground truth depth image.

2.2 Unary term experiment

Experiment Procedure In this experiment we infer the depth of a test image from training images and compare the inferred depth with the ground truth depth. We select one image among 200 images to use it as a test image and use the remaining 199 images as training images. And we scan the test image with 3 by 3 window size patch from the first pixel of the image until the end of the image without any overlapping patch. While this window slides over the entire pixels on the test image, for each patch we select a patch of the same size i.e., 3 by 3 patch from training images that shares the most similar features with

the patch from the test image. As features we use RGB color by calculating how much the RGB values in test patch are different from the path in selected among the training patch. From training images we randomly select 1000 patches for each 3 by 3 patch from the test image and finally choose only one patch among 1000 patches from training images as the best match to the test patch. We hypothesize that since these images are from similar indoor scenes if the patches have similar features in RGB images then their depth information would also share quite a lot of similarities.

Experiment result Figure 2. shows the result of this experiment. The image on the left is the ground truth depth of the test image, the same image previously shown in Figure 1 on the right side and the image on the right side of the Figure 2 is the result obtained from this experiment. This result image was generated by concatenating patches from the training images that matched the best with each patch on the test image while the 3 by 3 patch was sliding over the test image. Since these two images don't look similar it seems like we did not successfully infer the depth information from the training images for the given test image. In the next experiment we plan to use other features such as features obtained from SIFT [10] rather than using the RGB color comparison to find the best matched patch to the test image from training images.



(a) (b)

Figure 2: Ground truth depth image on the left and the result image inferred on the right.

We also use the norm-2 measurements of all the pixels in the ground truth depth image and the result image as in (1) to evaluate the result more quantitatively.

$$diff = sqrt(\sum_{(x,y)\in G, G(x,y)\neq 0} \frac{(G(x,y) - P(x,y))^2}{\#\, of\,(x,y)\in G, G(x,y)\neq 0}) \qquad (1)$$

where $G(x,y)$ means the ground truth depth at pixel (x,y) and $P(x,y)$ is the inferred depth at pixel (x,y) in the result image such as Figure 2 (b). We should notice that in (1) we are summing over all the pixels except for the ones with depth values are zeros. This is because in the ground truth depth image such as Figure 3 there exist some pixels with error marked with black circles and ellipses in the Figure 3 and also the boundary of the ground truth image is surrounded with depth zeros regardless of what the ground truth depth values are. We exclude these pixels in our formula (1) to calculate the difference between the ground truth depth and the predicted depth from the training images. For the result image in Figure 2 (b) we obtained 0.0010289 as an estimate using the formula (1).

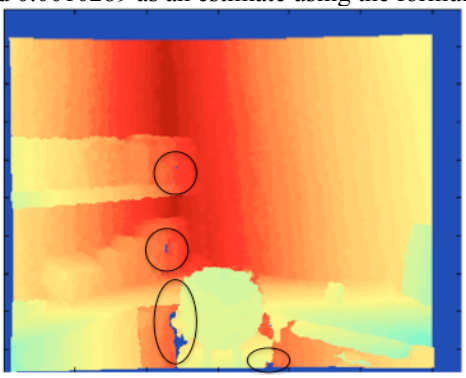


Figure 3. The pixels with error in depth information are marked with black circles and ellipses.

3. Plan

From this preliminary experiment we could observe what we need to improve to make our energy minimization approach to work. We explain our plan during the remaining time of the project as follows.

Feature detectors As we mentioned in the experiment result session above since simply comparing RGB values do not seems to find the good match for the test image in the training images we will use other feature detectors such as SIFT [10].

Speed up the code This experiment was performed with MATLAB m-files and it took 5 hours to get one result image in Figure 2 (b) for one test image. Since we want to run more experiments with more images and run several control experiments we need to make the code faster. Therefore we plan to make the code in mex files that make the computation more efficient when we use for-loops.

Experiments of more images We used only one image as a test image in this experiment. We plan to use 2 images for each scene as test images. Since there are 4 scenes, i.e., office kitchen bedroom and living room, we will use 8 images as test images.

Control experiments For each test image we will run three kinds of experiments. In the first control experiment we will use all the other 199 remaining images as training images. In the second control experiment we will use only the images in the same scene category as training images. Lastly we will use the images in the different scene category as training images. We expect to see the second control experiment performed the best and the third one performs the worst. We will also change the number of

training patches from 100 to 5000 whereas we used only 1000 training patches in this experiment. And we will also vary the size of the super pixels from 1by 1 to 10 by 10 whereas we used only the 3 by 3 size patch in this experiment.

Pair-wise term model After we get some intuition from the unary experiment we will model the pair-wise term in the energy function that models the depth value with respect to the relationship between neighboring super pixels and implement our energy minimization algorithm to infer the parameters to reconstruct depth information of the test image.

References

[1] Ashutosh Saxena, Sung Chung, and Andrew Ng. 3-D Depth Reconstruction from a Single Still Image. IJCV 2008.

[2] Ashutosh Saxena, Min Sun, Andrew Ng. Make3D: Learning 3D Scene Structure from a Single Still Image. PAMI 2008.

[3] Daniel Scharstein and Richard Szeliski. A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. IJCV 2002.

[4] Antonio Torralba and Aude Oliva. Depth Estimation from Image Structure. IEEE Trans Pattern Analysis and Machine Intelligence (PAMI), vol. 24, no. 9, pp.1-13, 2002.

[5] Ashutosh Saxena, Sung H. Chung, and Andrew Y. Ng. Learning Depth from Single Monocular Images. Neural Information Processing Systems (NIPS) 2005.

[6] Sara Vicente, Vladimir Kolmogorov, and Carsten Rother. Joint Optimization of Segmentation and Appearance Models. ICCV 2009.

[7] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. Real-Time Human Pose Recognition in Parts from a Single Depth Image. CVPR 2011.

[8] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. GrabCut: Interactive Foreground Extraction using Iterated Graph Cuts. SIGGRAPH 2004.

[9] James Hays and Alexei Efros. Scene Completion using Millions of Photographs. SIGGRAPH 2007.

[10] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints, IJCV 2004.