CS766 Project Midterm Report

Xucheng Wan
9077946623
xucheng.wan@wisc.edu

Zhengyang Lou
9073000235
zlou4@wisc.edu

Xiao Wang
9078040434
xwang2229@wisc.edu

In this report, we first give our current progress and difficulties encountered in implementing existing algorithms (which will serve as baseline algorithms). Then we provide a short literary review of some of the papers that are not mentioned in the project proposal we found relating to our project on top conferences during the past few years. Lastly, we state our plan for the remaining part of the project.

Progress and Difficulties

Progress:

(1) We set up our programming environment with tensorflow-gpu 1.4.0, opencv-python 3.3.0.10 and numpy 1.13.3 on a windows machine. The GPU is a NVIDIA GTX 1080.

(2) We implemented the algorithm in S. Zhou, Z. Lou, Y. Hu and H. Jiang's [1] paper in Python. This will serve as one of the baseline algorithm of multiview disparity estimation under noisy condition.

Difficulties:

The open source code of J. Zbontar and Y. LeCun [2], and W. Luo, A. Schwing, and R. Urtasun [3] are written in torch, which makes it hard to run on a windows machine with CUDA. we tried to set up a ubuntu environment by using a dual boost. But it turns out that installing CUDA drive on ubuntu is also quite time consuming. Since our purpose is to use these open source code as a reference of our own implementation in tensorflow rather than training the model by ourselves using their code, we decided to use ubuntu subsystem under windows 10. The official implementation of [4] is not available and open source code cannot reach the results reported in the paper. So we will not re-implement the paper.

Literary Review:

Monocular Depth Estimation

Some current works[5,6] conduct monocular depth estimating methods simply using end-to-end learning approach under unsupervised training environment.

Clement et al.[7] point out that traditional monocular depth estimation requires extremely large quantities of ground truth paired depth image. Dataset of such large scale is hard to find, so they propose to train a convolutional neural network that maps between two paired-images. In this CNN model, denoted by $\hat{d} = f(I)$, the left image I is used as the input image and the right image $\hat{d}$ as the ground truth image. Such a model represents the intuition between stereo images and thus contains inference towards the 3D shape of a scene. In this paper, a novel loss function is used combining appearance matching loss, disparity smoothness loss and left-right disparity consistency loss.

Binocular Stereo Estimation

Patrick et al. [8] propose a novel hybrid model combining CNN with Conditional Random Field so that complex local matching cost and parameterized geometric priors can be combined in a global optimization approach. They use Structured output Support Vector Machine to train the joint model end-to-end and they conduct learning without any pre-processing.

Multi-view Stereo Estimation

The [9, 10, 11] focus on multi-view Stereo Estimation. C. B. Choy et al. [9] unified single-view and multi-view reconstruction with 3D convolutional LSTM based on their own assumption. Other researchers like Fabian et al. and Schonberger et al. do the stereo estimation in different ways for different aims.

Active method for Stereo Estimation

Aiming at solving specialized problem of stereo matching under active illumination, Sean Ryan Fanello et al. [12] propose an ‘active’ stereo which uses unsupervised greedy optimization to learn features for estimation infrared images. During the optimization process, a series of sparse hyperplanes are trained and optimized to map image patches into compact representations. Such ‘active’ stereo avoids the strict requirement of camera calibration procedures and it largely reduces adverse effect caused by overlapping sensors.

3D reconstructure

Similar as mentioned above, some current work focus on the 3D reconstruction with kinds of different views, such as fisheye camera [13], spherical panoramic camera [14,15] and getting the depth at the same time.

Cameras (such as fisheye cameras) which are able to capture color images with a large FoV, but lacking the 3D information. Alejandro et al. use such system which is proposed to extend the depth information to the fisheye image via layout estimation. Then they get a new 180 degree depth image where the center has the initial depth information and the periphery a good estimation of the structure of the room.

Hand-held 360 VR cameras are released, but it cannot provide stereoscopic image (needs depth) which is essential for realistic VR contents. Sung Hoon et al. generate an all-around depth for 360 VR camera from 1-second small motion video and give out a sphere sweeping method on the basis of the unit sphere.

Plan

For the next step, we will finish the implementation of [1], and [2] as well if time permitting. For our own methods, we will first develop view selection networks for [1] to replace its view selection part. Namely, we will replace the view selection part of [1] with neural network. To do this in a supervised fashion, we need a binary image B where every pixel denotes the best choice of number of views used. For each pixel location, we compute all possible outputs with all possible combinations of views used of the algorithms. Then we find the best number of views that minimizes the 2-norm to the disparity ground truth. After that, we set the pixel location of the binary image B to that number. Then a network is trained to learn the best number of views. Besides, we will also investigate the possibility of integrating neural network with PGM by adding pre-trained neural networks from Patrick et al[8] to perform feature extraction, and use [16] for the view selection and depth prediction.

References

[1] S. Zhou, Z. Lou, Y. Hu and H. Jiang. Improving disparity map estimation for multi-view noisy images. In *International Conference on Acoustics, Speech and Signal Processing* (ICASSP) 2018.

[2] J. Zbontar and Y. LeCun. Computing the stereo matching cost with a convolutional neural network. CoRR, abs/1409.4326, 2014.

[3] W. Luo, A. Schwing, and R. Urtasun. Efficient deep learning for stereo matching. In *Computer Vision and Pattern Recognition* (CVPR), 2016.

[4] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry. End-to-end learning of geometry and context for deep stereo regression. arXiv preprint arXiv:1703.04309, 2017.

[5] J. Flynn, I. Neulander, J. Philbin, and N. Snavely. Deepstereo: Learning to predict new views from the word's imagery. *Computer Vision and Pattern Recognition*(CVPR), 2016.

[6] R. Garg, V. K. BG, G. Carneiro, and I. Reid. Unsupervised CNN for single view depth estimation: Geometry to the rescue. In *European Conf. Computer Vision* (ECCV), 2016.

[7] T. Zhou, M. Brown, N. Snavely, and D. Lowe. Unsupervised Learning of Depth and Ego-Motion from Video. In *Computer Vision and Pattern Recognition* (CVPR), pages 1851-1860, 2017.

[8] P. Knbelreiter, C. Reinbacher, A. Shekhovtsov et al. End-to-End Training of Hybrid CNN-CRF Models for Stere. In *Computer Vision and Pattern Recognition* (CVPR), pages 2339-2348, 2017.

[9] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European Conference on Computer Vision* (ECCV), 2016.

[10] Fabian Langguth, Kalyan Sunkavalli, Sunil Hadap, and Michael Goesele. 2016. Shading aware multi-view stereo. In *European Conference on Computer Vision* (ECCV), 2016

[11] J. L. Schonberger, E. Zheng, M. Pollefeys, and J.-M. Frahm. Pixelwise view selection for unstructured multi-view stereo. *European Conference on Computer Vision*(ECCV), 2016.

[12] S. R. Fanello, J. Valentin, C. Rhemann, et al. Ultra Stereo: Efficient Learning-based Matching for Active Stereo Systems. In *Computer Vision and Pattern Recognition* (CVPR), pages 2691-2700, 2017.

[13] A. Perez-Yus, G. Lopez-Nicol´as, and J. J. Guerrero, "Peripheral expansion of depth information via layout estimation with fisheye camera," in *European Conference on Computer Vision* (ECCV), 2016.

[14] S. Im, H. Ha, F. Rameau, and H. Jeon. All-around depth from small motion with a spherical panoramic camera. In *European Conference on Computer Vision*(ECCV), 2016.

[15] J. Ventura. Structure from motion on a sphere. In *European Conference on Computer Vision* (ECCV), 2016.

[16] E. Zheng, E. Dunn, V. Jojic, and J.-M. Frahm. PatchMatch based joint view selection and depth map estimation. In *Computer Vision and Pattern Recognition* (CVPR), 2014.