Light Field Depth Estimation With Multi-Layer Perceptron

Sarah Xu
Stanford University
Electrical Engineering
sarahxu@stanford.edu

Abstract

Depth estimation is an important topic in computer vision, and light field, providing both spatial and angular information of a scene, is valuable in estimating the depth map of the scene. In this project, we explored a multi-layer perceptron (MLP) based approach of depth estimation for light field. First, we extract features from light field data such as the defocus and correspondence cues. Then, we combine the various features and exploit the relations between them using a MLP model to produce an accurate depth map. We trained the MLP model with data from the 4D Light Field Dataset [6] and successfully produced better depth estimation comparing to Markov Random Field (MRF) based method. The code for this project can be found at github.[1]

1. Introduction

Depth estimation has been and continues to be an important topic in computer vision, it provides critical geometric information about the scene that is being captured, and it has many applications in robotics, autonomous driving, gaming, etc. Light field models the light ray as a 5D plenoptic function of position, angle, wavelength, and time dimensions [2]. It was simplified to a 4D function as radiance is constant along a ray in empty space [7]. Compared to conventional 2D images, the light field contains both spatial and angular information and multiple cues about the scene that can be used for refocusing and estimating the depth map. With the recent advances in commercializing the light field micro-lens array devices, depth map estimation algorithms have been one of the focused research areas. Since current light field devices capture each pair of sub-aperture images with a very narrow baseline, the disparity of the images and the spatial resolution are restricted, making the accuracy of depth recovery limited. Depth map estimation is also very challenging with non-Lambertian surfaces, i.e. specular materials and with occlusion. In this project, we want to explore various cues, algorithms, and methods that leverage the structure and correlations in spatial, angular, (and temporal) dimensions to produce a depth map estimation from the light field data.

[1] https://github.com/ysx001/ee367-lightfield-depth

2. Related Work

2.1. Sub-aparture Image Matching Methods

The sub-aperture image matching method, Adelson and Wang in 1992 described the "single lens stereo" that can be used to estimate the depth map of light field by calculating the disparity of each pixel using spatial and viewpoint derivatives. [1] Ng et al. proposed a lightfield refocusing algorithm using the Fourier Slice Photography Theorem, where the derivative between different angular views reflects depth probabilty. [9] Tao et al., following Ng et al.'s work, used, defocus cue, correspondence cue and later shading cue to estimate the depth of the scene. [11, 12]. Furthermore, many research works have been done for estimating the depth in scattering medium, occlusion, noise and for glossy materials. [14, 17, 8, 13] Heber et al. also proposed a method that shears the light field view, and considered each warped image as a row in a matrix that is low rank. [3]

2.2. Combining the Visual Cues

To produce an accurate depth map incorporating difference visual cues and smooth out the outliers produced by each cue, there are two prominent methods based previous research work. Markov Random Field(MRF) is used by Tao et al. and others to combine the different visual cues and use the properties of the depth map to provide an estimated depth map. [11] Furthermore, Wanner and Goldluecke presented a total variation (TV) smoothing method for combining the different cues. [16]

2.3. Learning-based Methods

With the recent advance of Machine Learning-based approaches, there has been a plethora of research exploring learning-based methods that provides an end-to-end pipeline for depth estimation. Heber et al. proposed a CNN-based method that learns an end-to-end mapping and predicts depth information from light field data, and later improved from 5-layer CNN to a U-Net based method. [4, 5] However, the CNN-based method is restricted by the amount of training light field data with accurate ground truth depth map labels, which are difficult to obtain. Recent research tries to address this limitation by data augmentation, training an attention module, and unsupervised or zero-shot feature CNN that exploit the correlations between spatial, angular, and temporal dimensions. [10, 15]

3. Method

In this project, we used a two-step pipeline to estimate depth from light field data as shown in Figure 1 The first step is to extract various visual cues from the light field data based on Tao et al's work, and the second step is to effectively combine the cues using a MLP network. We choose to use an MLP network to explore the relationships between the features and cues because of it's ability to exploit the internal connections between data and the target and efficiently fit a model to the relationship. Further, from previous experiment results, MRF achieves limited success in combining the visual cues.

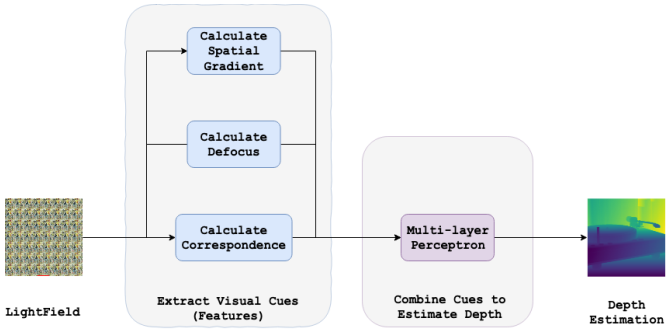

Figure 1. Depth Estimation Pipeline with Multi-Layer Perceptron

3.1. Feature Extraction

We are mainly interested in extracting two visual cues from the light field data for the scope of this project. Both the defocus cue and the correspondence cue are calculated from the shift focal stack using the shift and and algorithm as discussed in class. For a given shift metric α and a input light field $l_0(x, u)$, we calculate the shifted light field focal stack using

$$l_\alpha(x, u) = l_0(x + u(1 - \frac{1}{\alpha}), u)$$

The equation for shifting on full 4D data is, [ng et al citation]

$$L_\alpha(x) = L_0(x + u(1 - \frac{1}{\alpha}), y + v(1 - \frac{1}{\alpha}), u, v)$$

We then integrate the images on shifted sensors by adding them together using the following equation:

$$\bar{l}_\alpha(x) = \frac{1}{N_u} \sum_{u'} l_0(x + u(1 - \frac{1}{\alpha}), u')$$

Where N_u is the number of angular pixels in u axis.[2]

3.1.1 Defocus Cue

The defocus cue is computed relative to a patch around the pixel of interest

$$D_\alpha(x) = \frac{1}{|W_D|} \sum_{x' \in W_D} |\Delta_x \bar{l}_\alpha(x')|$$

where W_D is the size of patch and Δ_x is a spatial Laplacian operator along the x-axis. For each α, we calculate a corresponding defocus cue.

Additionally, we calculate the α values that maximize spatial contrast for defocus cue for each pixel

$$\alpha_D^*(x) = \underset{\alpha}{\operatorname{argmax}}\, D_\alpha(x)$$

3.1.2 Correspondence Cue

We first calculate the angular variance of a given spatial pixel given an α for correspondence.

$$\sigma_\alpha(x)^2 = \frac{1}{N_u} \sum_{u'} (l_\alpha(x, u') - \bar{l_\alpha(x)})^2$$

And for correspondence cue, we calculate it relative to a patch around the pixel of interest as well

$$C_\alpha(x) = \frac{1}{|W_C|} \sum_{x' \in W_C} |\sigma_\alpha(x')|$$

where W_D is the size of patch. For each α, we calculate a corresponding correspondence cue.

Additionally, we calculate the α values that minimize angular contrast for correspondence cue for each pixel

$$\alpha_C^*(x) = \underset{\alpha}{\operatorname{argmin}}\, C_\alpha(x)$$

[2]For more graphical illustration of the shift + add algorithm, please refer to Stanford EE367 problem session 5. `http://stanford.edu/class/ee367/`

3.2. Multi-Layer Perceptron

Multi-Layer Perceptron is a type of artifical neural network that contains multiple layer of perceptrons and non-linear activation functions that connects each percetron nodes. As shown in Figure 2[3], each input x_i is weighted by the weight $(w_j^{[1]})_i$ and the weighted input are summed together to compute the activation a_j, i.e.

$$a_j = A(\sum_{i=1}^{d}(w_j^{[1]})_i x_i)$$

where $A(x)$ is some non-linear activation function such as Rectified Linear Unit(ReLU), Sigmoid Function, Leaky ReLU or others. The weights on each edge of connecting the node are updated using back-propagation during each training iteration, allowing the network to fit a model to the representation of the input data and the relationships between each other and the ground truth.

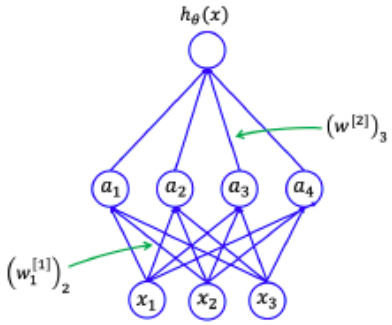

Figure 2. An example two-layer MLP

In this project, we used a 6-layer MLP with Leaky ReLU activation function between layers, and a ReLU activation function at the output layer. The input data is all the features for a single pixel in the input light field, and the ground truth is the normalized depth map of the single pixel.

4. Experiments

4.1. Dataset

The Dataset we used is the 4D Light Field Dataset[4], where 24 rendered light field data with their respective ground truth depth map are provided. [6] The 24 sets of light field data is further divided in 3 groups: 18 sets for training, 3 sets for evaluation and 3 sets for test. For each set of light field data, we have $9 * 9$ images of size $512 * 512 * 3$.

[3] http://cs229.stanford.edu/notes2020fall/notes2020fall/deep_learning_notes.pdf

[4] https://lightfield-analysis.uni-konstanz.de/

4.2. Preprocessing

We preprocessed the light field data to extract defocus and correspondence cues. For calculating the cues, we used $\alpha \in [0.2, 2]$ with step resolution of 64 steps ($\alpha_{step} = 0.028$ instead of $\alpha_{step} = 0.007$ as in the original paper to save some preprocessing time.) W_D and W_C are $9 * 9$ windows as suggested in the paper. For each cue, the preprocessed data size is $65 * 512 * 512 * 3$.

4.3. Hyper-parameters

The Hyper-parameters we experimented for this project includes the patch size of input to the MLP network, the learning rate and optimizer for the network, the loss function, the activation function of between the layers, and the size of the layers. For the patch size of the network, we experimented with (1 * 1), (3 * 3) and (5 * 5), where (3 * 3) achieves the best performance. The learning rate we used is $1e-5$ with the Adam optimizer, which is further reduced to $1e-7$ when loss is on plateau. We used $l1$ loss as the penalty function since we want the residuals to be closely constraint to 0. And we used Leaky ReLU as activation in between the middle layers, and ReLU on the output layer. We also added a dropout layer with rate of 0.2 since we have a very small dataset and we want to prevent overfitting. The size of the middle layers are $[256, 128, 64, 32]$ respectively.

5. Results

5.1. Qualitative Results

From Figure 3 we can observe that the MLP-based depth estimation method achieve visibly better depth map comparing to refocus depth (where patch window in defocus cue is 1), defocus depth, correspondence depth, defocus and correspondence combined depth on the three sets of test light field data. The MLP depth estimation has less noise on the background, and has sharper edges that estimates the contour of the objects well.

5.2. Quantitative Results

To perform a quantitative comparison of the depth estimation from different algorithms, we normalize the ground truth depth map as well as all the algorithmic depth estimation outputs to between 0.0 and 1.0, then we calculate the mean-squared-error(MSE) between each output and the grouth truth. The results is shown in Figure 4. As we can observe from the table, MLP achieves lowest MSE on the test data comparing to other algorithms. One notable pattarn is that defocus or correspondence based depth performs better under different situations, and the combined result

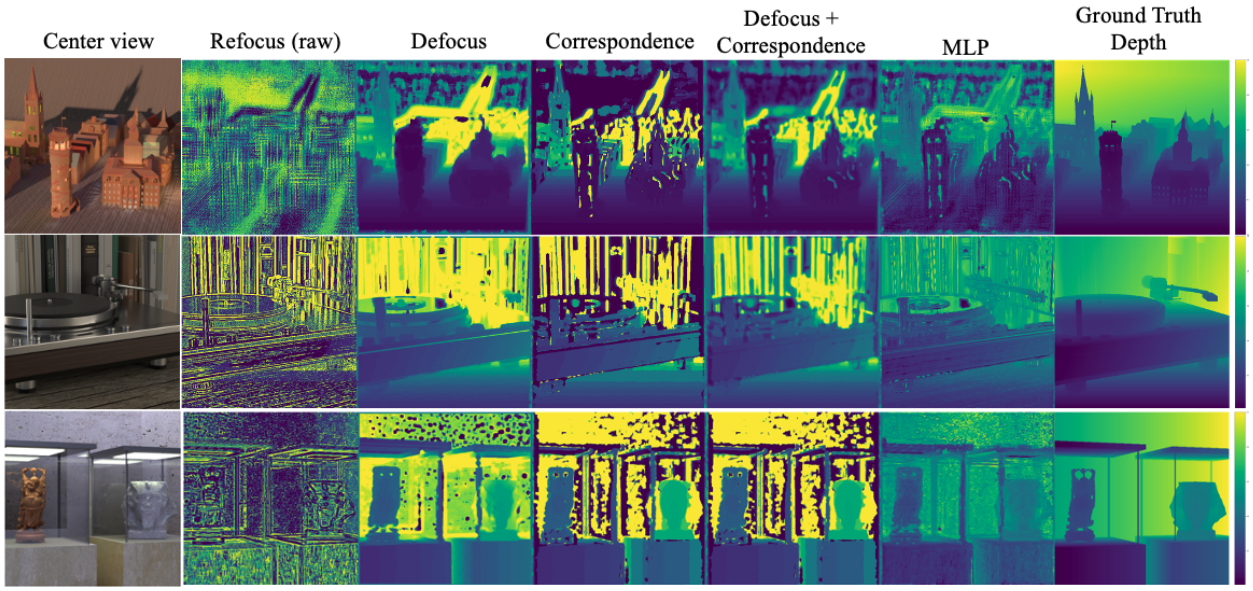

Figure 3. Depth Estimation Result Comparison between refocus (1 pixel gradient), defocus, correspondence, defocus and correspondence combined, MLP and ground truth depth map.

does not always produce a lower mse. In constast, the MLP model consistently find the right balance between different cues and reliably produce a better depth estimation.

6. Discussion

In this work, we demonstrated that the MLP-based depth estimation has the ability to combine different extracted features from light field data, find the relationship between the different features, and produce a depth estimation that is better than the MRF based approach. However, due to time and resource limitation, we did not explore automatic extraction of light field features using a convolutional neural network (CNN) or an attention network. The next step of the project could involve exploring other light field visual cues that can be used for depth estimation, and use an appropriate CNN architecture for feature extraction step.

Moreover, one of the main usage of depth map is providing information for 3D reconstruction. We also need to verify that if the MLP-based depth estimation produces reliable 3D reconstruction results.

Lastly, the training and test data of this project was limited to the rendered light field data provided by 4D Light Field Dataset. This project could be expanded and generalized if we train it on a broader dataset that contains both rendered data and real-world light field data, e.g. Lytro Camera data.

7. References

References

[1] E. H. Adelson and J. Y. Wang. Single lens stereo with a plenoptic camera. *IEEE transactions on pattern analysis and machine intelligence*, 14(2):99–106, 1992.

[2] A. Gershun. The light field. *Journal of Mathematics and Physics*, 18(1-4):51–151, 1939.

[3] S. Heber and T. Pock. Shape from light field meets robust pca. In *European Conference on Computer Vision*, pages 751–767. Springer, 2014.

[4] S. Heber and T. Pock. Convolutional networks for shape from light field. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3746–3754, 2016.

[5] S. Heber, W. Yu, and T. Pock. U-shaped networks for shape from light field. In *BMVC*, volume 3, page 5, 2016.

[6] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke. A dataset and evaluation methodology for depth estimation on 4d light fields. In *Asian Conference on Computer Vision*, pages 19–34. Springer, 2016.

[7] M. Levoy and P. Hanrahan. Light field rendering. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 31–42, 1996.

[8] H. Lin, C. Chen, S. B. Kang, and J. Yu. Depth recovery from light field using focal stack symmetry. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3451–3459, 2015.

MSE	Refocus (raw)	Defocus	Correspondence	Defocus + Correspondence	MLP
Tower	0.172	0.115	0.220	0.148	**0.082**
Vinyl	0.230	0.649	0.144	0.057	**0.030**
Museum	0.301	0.163	0.070	0.163	**0.038**

Figure 4. Depth Estimation MSE Comparsion between refocus (1 pixel gradient), defocus, correspondence, defocus and correspondence combined, MLP and ground truth depth map.

[9] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan. *Light field photography with a hand-held plenoptic camera*. PhD thesis, Stanford University, 2005.

[10] J. Peng, Z. Xiong, Y. Wang, Y. Zhang, and D. Liu. Zero-shot depth estimation from light field using a convolutional neural network. *IEEE Transactions on Computational Imaging*, 6:682–696, 2020.

[11] M. W. Tao, S. Hadap, J. Malik, and R. Ramamoorthi. Depth from combining defocus and correspondence using light-field cameras. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 673–680, 2013.

[12] M. W. Tao, P. P. Srinivasan, J. Malik, S. Rusinkiewicz, and R. Ramamoorthi. Depth from shading, defocus, and correspondence using light-field angular coherence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1940–1948, 2015.

[13] M. W. Tao, J.-C. Su, T.-C. Wang, J. Malik, and R. Ramamoorthi. Depth estimation and specular removal for glossy surfaces using point and line consistency with light-field cameras. *IEEE transactions on pattern analysis and machine intelligence*, 38(6):1155–1169, 2015.

[14] J. Tian, Z. Murez, T. Cui, Z. Zhang, D. Kriegman, and R. Ramamoorthi. Depth and image restoration from light field in a scattering medium. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2401–2410, 2017.

[15] Y.-J. Tsai, Y.-L. Liu, M. Ouhyoung, and Y.-Y. Chuang. Attention-based view selection networks for light-field disparity estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12095–12103, 2020.

[16] S. Wanner and B. Goldluecke. Variational light field analysis for disparity estimation and super-resolution. *IEEE transactions on pattern analysis and machine intelligence*, 36(3):606–619, 2013.

[17] W. Williem and I. K. Park. Robust light field depth estimation for noisy scene with occlusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4396–4404, 2016.