Using Light-field Angular Coherency for Depth Estimation

Rutika Moharir
Carnegie Mellon University

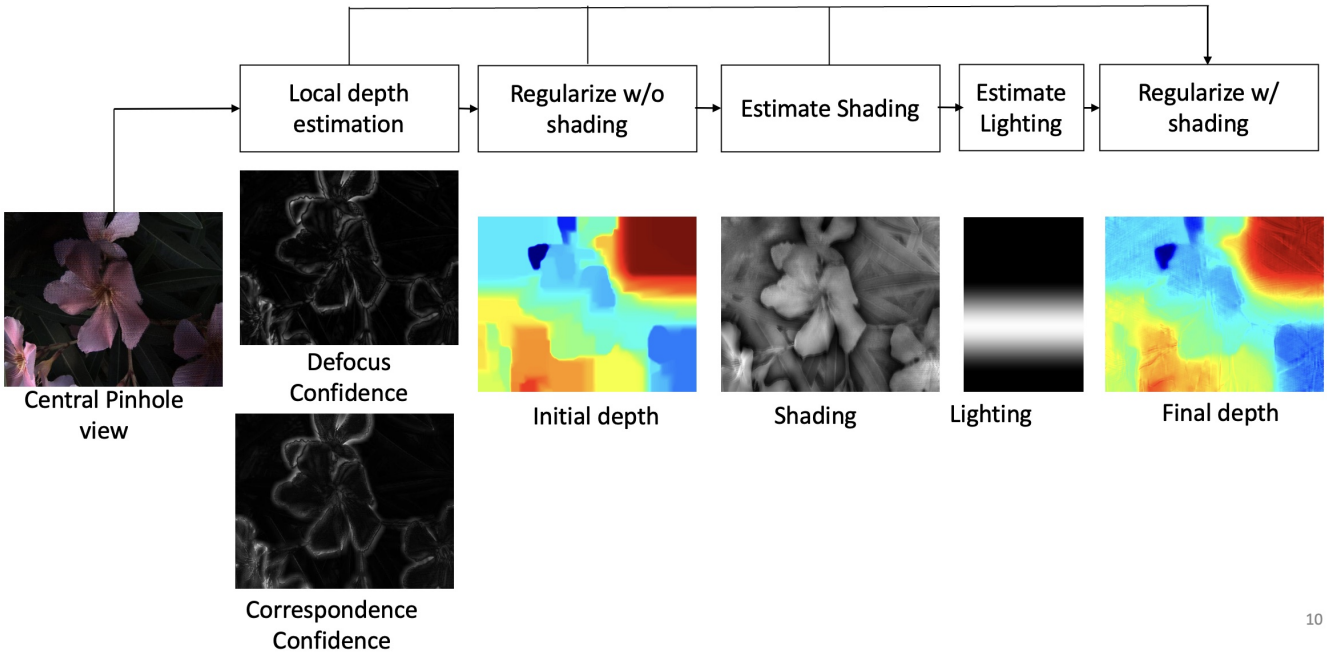

Figure 1: The proposed depth recovery pipeline using cues from defous, correspondence and shading.

ABSTRACT

Plenoptic cameras are rich with information about the scene and hence are increasingly employed in various consumer and industrial applications. One of the most important application of plenoptic cameras is practical depth recovery which can be done from just a single-shot capture. However, this recovered depth is restricted by the narrow baselines and spatial resolution of the camera. One possible solution is to add regulariztion to an initial depth estimate, but this inital estimate can itself be incorrect on smooth surfaces where the depth is mostly non-existent.

This report explores an approach which enables highly accurate depth recovery using not only the regularly used cues of defocus and correspondences, but improves the depth estimation by using cues from shading which helps refine the details in the objects's shape. The optimization framework first generates the all-in-focus image where the pixels exhibit angular coherency which is then used to optimize for the depth, lighting and shading to improve the depth estimates.

Additionally, since the approach is optimized for Lambertian objects, it fails for glossy or specular surfaces. Hence a method to estimate and remove the specular component is explored, thus enabling accurate depth estimation for any real world scenes.

1 INTRODUCTION

Light-field images provide a host of geometric cues, the most exploited ones being defocus and correspondence cues which help in initial depth recovery from a single input image. However, by exploiting the full angular data captured in the light-field image, we can extract additional cues of lighting and shading which can help us in regularizing the initial depth estimates particularly on smoother surfaces where otherwise the estimated depth is mostly planar. [9] show an optimization framework that exploits such angular coherency exhibited by light-field images for accurate depth estimation. This report aims to implement their algorithm and show its results on real world images captured using the Lytro camera.

The optimization framework makes the common assumption of Lambertian surfaces under general (distant) direct lighting. The 4D light-field data is represented as an epipolar image (EPI) with parameters (x, u) where u is the angular dimension (along on the lens aperture) and x is the spatial dimension (along the scanline). By integrating over the angular u dimension and computing the variance in x dimension we can get the defocus cue where high value means the point is in focus. While the correspondence cue can be calculated by measuring the the angular u variance, where low variance indicates correct correspondence. Thus using the EPI we can get defocus and correspondence cue. However, these measurements rely on high texture in the scene. Hence we need a stronger measurement for these cues.

We can get an all-in-focus image obtained by refocusing the light-field image. But this is not the only advantage of light-field refocusing. Each spatial pixel in the all-in-focus image represents viewpoints that converge on one point on the scene, thus exhibiting angular coherency. As stated in [6] angular coherency can give us constraints for photo-consistency, depth consistency and shading consistency. Thus by deriving a relationship between this angular coherency and refocusing, we can get stronger geometric cues to regularize the depth. Not only accurate depth and shading information, we can also extract lighting and normal estimates from this relationship between angular coherency and refocusing which would have been otherwise difficult for non-textured surfaces.

To extract and remove the specular component in the image, we try to estimate the confidence of the lighting estimate at each pixel and weight the lighting with this obtained confidence measure.

2 PREVIOUS WORK

2.1 3D using Photometric Stereo

Extracting depth information from a single-shot capture is a heavily underconstrained problem. Many prior work either assume known light source environemtns or make assumptions about the geometry. Some other techniques, like photometric stereo is based on extracting depth and 3D surfaces by varying lighting across multiple images. The classical methods [1], [11] focus on lambertian surfaces, more recent works have [7] addressed challenges posed by non-Lambertian reflectance, enhancing the method's versatility. One of the key advantages of photometric stereo lies in its ability to recover fine surface details and intricate geometries. Nevertheless, challenges exist, including sensitivity to specular reflections, the need for precise lighting control, and limitations in handling real-world scenes with varying reflectance properties and the obvious disadvantage of requiring multiple image captures.

2.2 3D from Disparity and Lightfields

Light-field images provide depth cues that can be used for accurate depth recovery as suggested in [4], [10]. However, these algorithms struggle with the accuracy of depth in low textured regions since they rely on local contrast, requiring texture and edges. One solution is using depth regularizers as in [2] but it results in planar depths. This report is based on the approach in [9] which is build on [8] to improve depth estimation from defocus and correspondence cues, and additionally incorporate shading information.

2.3 3D using Structured Light

Structured light 3D reconstruction is another widely employed technique for capturing and reconstructing the three-dimensional geometry of objects or scenes. This method involves projecting a known pattern, often a grid or series of stripes, onto the target surface and analyzing the deformation of the pattern to infer depth information. One of the pioneering works in this field is by [12], which introduced the concept of structured light triangulation for 3D reconstruction. More recent works [5] have explored the use of advanced coding techniques to enhance the robustness and precision of these structured light systems. The advantages of structured light include its capability to capture detailed surface information with high accuracy and are versatile and can be implemented in various environments. However, these methods are highly sensitive to ambient light, limitations in capturing mirror-like or transparent surfaces, and the need for careful calibration.

3 BACKGROUND

3.1 Refocusing using Angular Coherence

The goal is to solve for $\alpha^*(x, y)$ and the shading S in $P(x, y) = A(x, y)(x, y)$ where P is the central pinhole image of the light-field input L_0, A is the albedo and S is the shading

In order to refocus the light-field image to depth α we remap the light-field image as follows

$$L_\alpha(x, y, u, v) = l(x^f(\alpha), y^f(\alpha), u, v) \tag{1}$$

$$x^f(\alpha) = x + u(1 - \frac{1}{\alpha}) \tag{2}$$

$$y^f(\alpha) = y + v(1 - \frac{1}{\alpha}) \tag{3}$$

where L_α is the refocused image. The central viewpoint is located at $(u, v) = (0, 0)$

Given the depth $\alpha^*(x, y)$ for each spatial pixel (x, y) we calculate L^*_α by refocusing each spatial pixel to its respective depth. All angular rays converge to the same scene point when refocused at α^*

$$L_{\alpha^*}(x,y,u,v) = L_0(x^f(\alpha^*(x,y)), y^f(\alpha^*(x,y)), u, v) \quad (4)$$

Thus L^*_α is the remapped lightfield image for the all-in-focus image. However, for the central pinhole image, the shear $x^f(\alpha), y^f(\alpha)$ are independent of (u,v). Hence at every α

$$L_\alpha(x,y,0,0) = P(x,y) \quad (5)$$

The central angular coordinate always images the same point in the scene, regardless of the focus. This property of refocusing allows us to exploit photo consistency, depth consistency, and shading consistency.

Photo consistency All angular rays converge to the same point in the scene at each spatial pixel, the angular pixel colors converge to $P(x,y)$

$$L^*_\alpha(x,y,u,v) = P(x,y) \quad (6)$$

Depth Consistency The angular pixel values should also have the same depth values, i.e. angular pixels (u,v) share the same depth for each (x,y)

Shading consistency Following from the photo consistency of angular pixels for each spatial pixel in L^*_α , shading consistency also applies, since shading is viewpoint independent for Lambertian surfaces.

$$S(x^f(\alpha^*(x,y)), y^f(\alpha^*(x,y)), u, v) = S(x,y,0,0) \quad (7)$$

Algorithm 1: Depth from Shading, Defocus, and Correspondence

$Z, Z_{conf} = LocalEstimation(L)$
$Z^* = Optimizedepth(Z, Z_{conf})$
$S = EstimateShading(L)$
$l = EstimateLighting(Z^*, S)$
$Z^* = OptimizeDepth(Z^*, Z_{conf}, l, S)$
return Z^*

4 METHOD

4.1 Defocus Response from Angular Coherence

The goal is to the find the depth α^* for each spatial pixel, i.e. from 4 we want to find α^* such that

$$\alpha^*(x,y) = argmin|L_0(x^f(\alpha), y^f(\alpha), u, v) - P(x,y)| \quad (8)$$

The equation enforces all angular pixels of a spatial pixel to equal the center view pixel color since the center pixel color P does not change regardless of the view α.

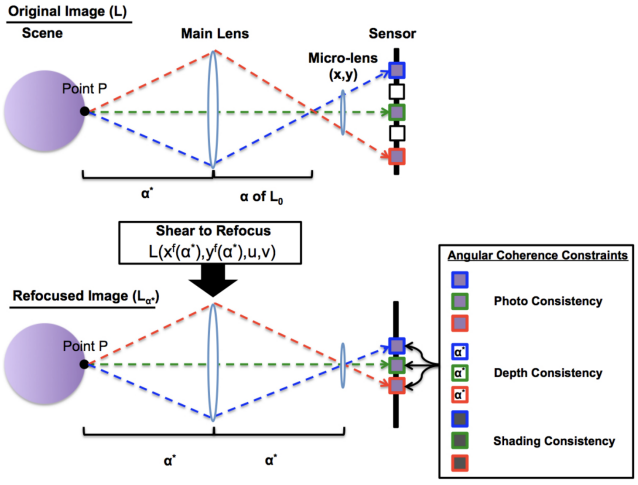

Figure 2: Angular Coherence and Refocusing

$$D_\alpha(x,y) = \sum_{(x',y')} |L_\alpha(x,y,u',v') - p(x',y')| \quad (9)$$

From 8 the defocus response is computed such that for each pixel in the image, we compare a small neighbourhood patch of the refocused image and its respective patch at the same spatial location of the center pinhole image.

4.2 Correspondence Response from Angular Coherence

Again by using 8 we can formulate a correspondence measure. Photo consistency is measured using the difference between the refocused angular pixels at α and their respective center pixel. This makes the measurement robust against small angular pixel variations due to noise.

$$C_\alpha(x,y) = \sum_{(u',v')} ||L_\alpha(x,y,u',v') - P(x,y)|| \quad (10)$$

4.3 Optimal Depth from Defocus and Correspondence

To get the optimal depth Z, a simple average of the defocus and correspondence cues, weighted by their confidences is used. The minimum of the combined response curve gives the optimal depth value for each pixel.

4.4 Regularizing Depth

The current obtained depth is a local depth estimation, to propagate this to regions with low confidence we optimize using the data constraint which retains the local depth values with high confidence and a smoothness constraint.

Data constraint: Using the confidence metric $Z_{conf}(x, y)$ from the defocus and correspondence cues, we define the data constraint as

$$E_d(x, y) = Z_{conf}(x, y) \cdot ||Z^*(x, y) - Z(x, y)||^2 \quad (11)$$

Smoothness Constraint: Smoothness constraint is implemented using 3 filters, laplacian, horizontal first derivative and vertical first derivative.

$$E_v(x, y) = \sum_{i=1,2,3} ||(Z^* \circledast F_i)(x, y)||^2 \quad (12)$$

However, this smoothness constraint propagates data with high local confidence resulting in planar depths. Hence we add shading constraints to give shape cues in low confidence regions.

Given the local depth estimation Z we want to find the optimized depth Z^* such that

$$E(Z^*) = \sum_{(x,y)} \lambda_d E_d(x, y) + \lambda_v E_v(x, y) \quad (13)$$

4.5 Shading from Angular Coherence

For shading estimation, instead of using the central pinhole image P, the entire light-field image L_0 is used to increase robustness. Hence we want to estimate S where $L_0 = AS$ where A is the albedo. The optimization solves for $S(x, y, u, v)$ in the log space, $logL_0 = log(A)$

The constraints to solve for A and S are obtained from depth, smoothness over local neighbours, and the angular coherency constraint which only constraints spatial pixels for the same angular viewpoint. For each pair of the set of (x, y, u, v) the shading constraint is as follows

$$E = \sum_{(p,q)} ||s(p) - s(q||^2 \quad (14)$$

4.6 Lighting from Angular Coherence

Given the shading S and central pinhole image P, the albedo A can be computed as $P = AS$

For lighting we have

$$P = A(x, y) \sum_{k=0}^{8} l_k H_k(Z^*(x, y)) \quad (15)$$

where l are the spherical harmonic coefficients and H_k are the spherical harmonic basis functions. Here the only unknown is l which is solved using linear least squares solver.

4.7 Final combining of all cues and Regularizing Depth

Finally, we add the shading constraint such that

$$E_s(x, y) = w_s(x, y) \cdot || \sum_{k=0}^{8} l_k H_k(Z^*(x, y)) - S||^2 \quad (16)$$

where $w_s(x, y) = 1 - Z_{conf}(x, y)$ to enforce shading constraint where the local depth estimation is not confident.

With the shading constraint the optimization objective becomes

$$E(Z^*) = \sum_{(x,y)} \lambda_d E_d(x, y) + \lambda_v E_v(x, y) + \lambda_s E_s(x, y) \quad (17)$$

which is solved using non-linear least squares.

5 EXTRACTING SPECULAR COMPONENT

For estimating the specular component, we want to classify if every pixel (x, y) is diffuse or specular. Such a per-pixel confidence measure can give us the final lighting estimation. The Attainable Maximum Likelihood (AML) [3] is used to measure the confidence $C(x)$ of the lighting estimation at each pixel. Then the final lighting estimation is

$$l^* = C(x)^{-1} \cdot l \quad (18)$$

To compute the confidence measure, we first remap the lighting estimate to the desired depths $(1, 256)$ using 2. To estimate the lighitng confidence at each pixel, we assume that the cost, here the lighting at each pixel, $C(x, d)$ is perturbed by a gaussian noise due to the specularity. We wish to estimate the true lighting l_0 at the true depth d_0 which does not have the lowest cost after the cost is perturbed. The likelihood is proportional to $e^{\frac{-(c(x,d)-c(x,d_0))^2}{\sigma^2}}$

The confidence C(x) is defined as the inverse of the sum of these probabilities for all possible depths

$$C(x) = (\sum e^{\frac{-(c(x,d)-c(x,d_0))^2}{\sigma^2}})^{-1} \quad (19)$$

This equation produces a high confidence when the cost has a single sharp minimum. The confidence is low when the cost has a shallow minimum or several low minima.

6 RESULTS

Qualitative results on both uniform and non-uniform albedo examples on real images are shown. These images were captured using the Lytro camera. Additionally, results for specular reflection removal are also shown.

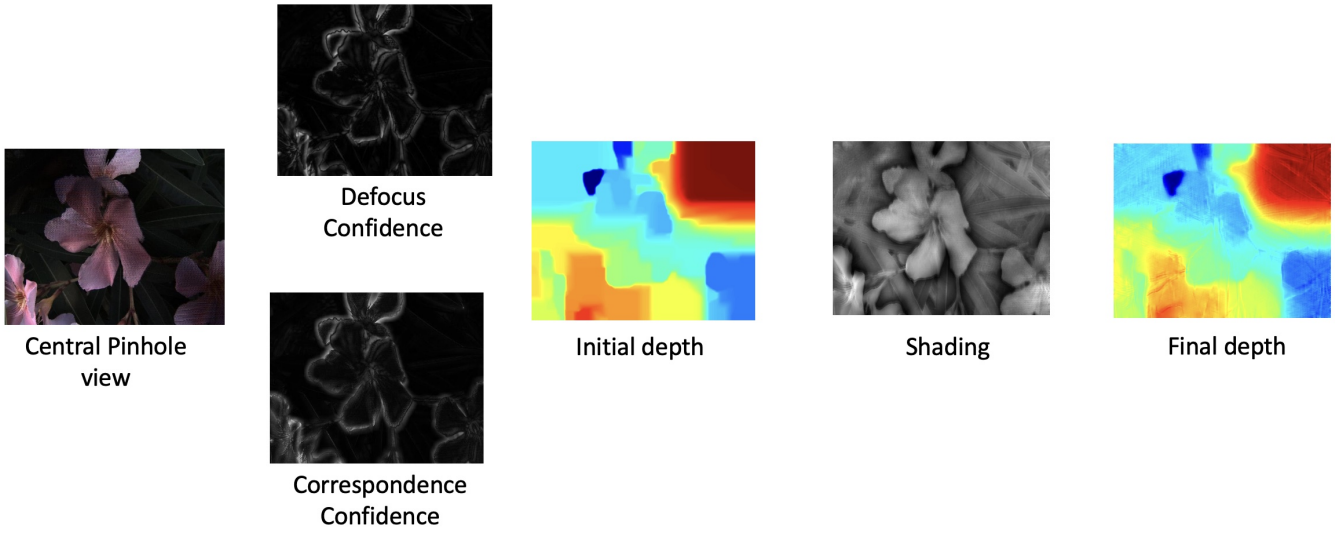

Figure 3: Result on the flower image. Defocus consistently shows better results at noisy regions and repeating patterns, while correspondence provides sharper results. But still even after combining both cues you can see the initial depth is mostly planar in low texture areas. However, we can see how the angular coherency help estimate highly accurate shading and hence detailed depth. Hence proving the need for shading constraints.

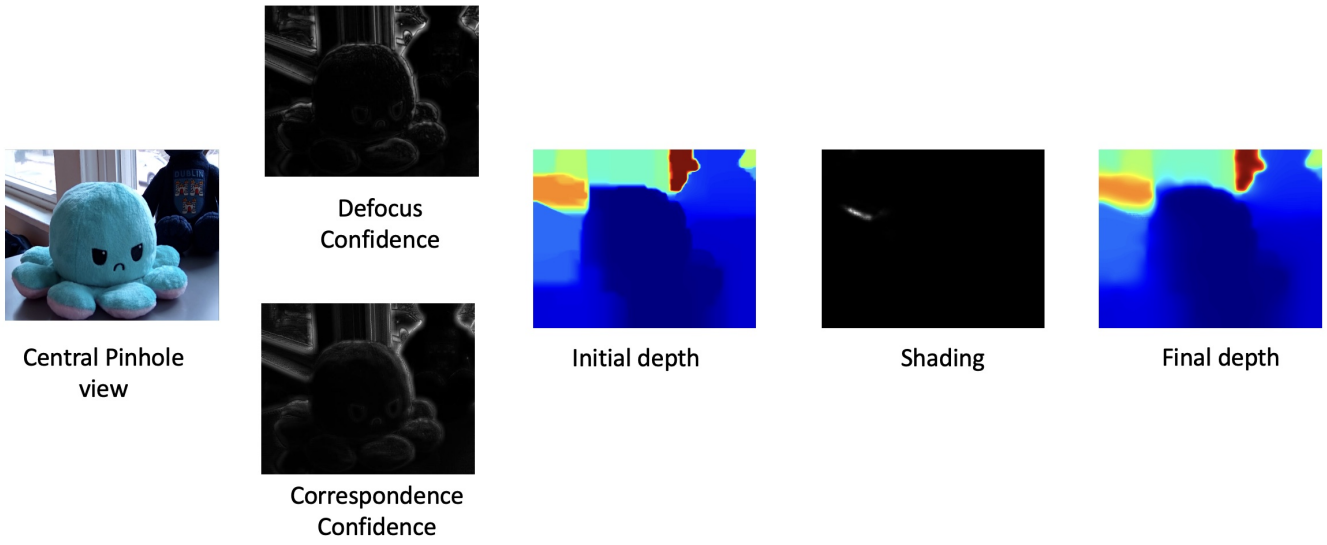

Figure 4: Result on the toy image. Here the shading estimate is inaccurate, hence the final depth is not that different from the initial depth.

7 IMPLEMENTATION CHALLENGES

- Since the original code on which this report is based [9] was in MATLAB, it was very computationally inefficient, I tried to re-implement it in python and was able to successfully get the defocus cues, correspomndecne cues and the initial depth estimates. However, post that including the shading constraint for depth regularization became increasingly tricky, hence due to lack of time left to debug, I resorted to using the mATLAB code itself. However, I plan of completing the implentation and open source the code for future research.

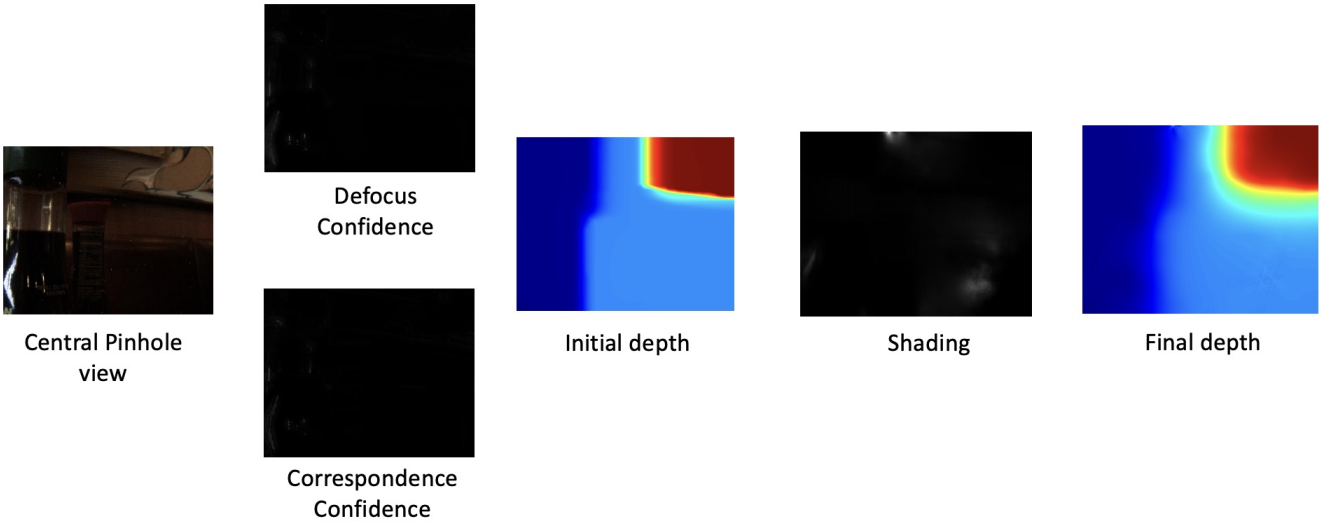

Figure 5: Another failure case is specular objects like this image of a bottle. Since the algorithm works on the assumption that we are picturing lambertian surfaces, the defocus and correspondence response is mostly negligible hence giving us planar depth.

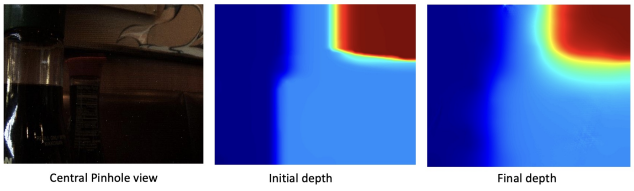

Figure 6: The suggested algorithm of extarcting specular component, did not help much in improving the depth estimation. Maybe the low light conditions in this image is also a factor hindering accurate depth estimation.

- The images from Lytro camera are RAW, hence they needed to be demosaicked and calibrated before usage.

8 EXTENSION OF PRIOR WORK

This report is an extension of prior work [9] through the following approaches -

- Show accurate depth estimations on complex real world images.
- Provides a simple algorithm to remove specular components in captured image, with the aim of improving depth recovery and not restricting the algorithm to only lambertian surfaces. Even though the intial results do not that look very good, the approach can be further poslished with the aim of correctly estimating the specular component in an image.

9 CONCLUSION

This report aimed to test and extend the works of [9] on real world images. Qualitative results were shown and failure cases were identified as those light-field images which have specular reflections present. A simple approach to work around such specularities was suggested and analysed.

A possible future direction could be to experiment with add confocal constancy in the pipeline for even better depth estimation. Another possible direction could be to explore some other sophisticated approach for dealing with non-lambertian surfaces.

REFERENCES

[1] P.N. Belhumeur, D.J. Kriegman, and A.L. Yuille. 1997. The bas-relief ambiguity. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 1060–1066. https://doi.org/10.1109/CVPR.1997.609461

[2] Bastian Goldlücke and Sven Wanner. 2013. The Variational Structure of Disparity and Regularization of 4DLight Fields. In *Proceedings : 2013 IEEE Conference on Computer Vision and Pattern Recognition ; CVPR 2013*, Patrick Kellenberger (Ed.). IEEE, Piscataway, 1003–1010. https://doi.org/10.1109/CVPR.2013.134

[3] Xiaoyan Hu and Philippos Mordohai. 2012. A Quantitative Evaluation of Confidence Measures for Stereo Vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34, 11 (2012), 2121–2133. https://doi.org/10.1109/TPAMI.2012.46

[4] Changil Kim, Henning Zimmer, Yael Pritch, Alexander Sorkine-Hornung, and Markus Gross. 2013. Scene Reconstruction from High Spatio-Angular Resolution Light Fields. *ACM Trans. Graph.* 32 (07 2013), 73. https://doi.org/10.1145/2461912.2461926

[5] Matthew Robinson, Hongwei Zhang, and Jian Zhang. 2017. Structured light systems: Design and applications. *Journal of Manufacturing*

Processes 25 (2017), 119–131.

[6] S.M. Seitz and C.R. Dyer. 1997. Photorealistic scene reconstruction by voxel coloring. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 1067–1073. https://doi.org/10.1109/CVPR.1997.609462

[7] Boxin Shi, Zhe Wu, Zhipeng Mo, Dinglong Duan, Sai-Kit Yeung, and Ping Tan. 2016. A Benchmark Dataset and Evaluation for Non-Lambertian and Uncalibrated Photometric Stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[8] Michael W. Tao, Sunil Hadap, Jitendra Malik, and Ravi Ramamoorthi. 2013. Depth from Combining Defocus and Correspondence Using Light-Field Cameras. In *2013 IEEE International Conference on Computer Vision*. 673–680. https://doi.org/10.1109/ICCV.2013.89

[9] Michael W Tao, Pratul P Srinivasan, Jitendra Malik, Szymon Rusinkiewicz, and Ravi Ramamoorthi. 2015. Depth from Shading, Defocus, and Correspondence Using Light-Field Angular Coherence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[10] Sven Wanner and Bastian Goldluecke. 2012. Globally consistent depth labeling of 4D light fields. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 41–48. https://doi.org/10.1109/CVPR.2012.6247656

[11] A. Yuille and D. Snow. 1997. Shape and albedo from multiple images using integrability. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 158–164. https://doi.org/10.1109/CVPR.1997.609314

[12] Zhengyou Zhang. 1999. Flexible Camera Calibration by Viewing a Plane from Unknown Orientations. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, Vol. 1. IEEE, 666–673.