Light-Field Database Creation and Depth Estimation

Abhilash Sunder Raj
Stanford University
abhisr@stanford.edu

Michael Lowney
Stanford University
mlowney@stanford.edu

Raj Shah
Stanford University
shahraj@stanford.edu

Abstract

Light-field imaging research has been gaining a lot of traction in recent times, especially after the availability of consumer-level light-field cameras like the Lytro Illum. However, there is a dearth of light-field databases publicly available for research. The project is mainly divided into two parts - creation of an extensive light-field database and evaluating a state-of-the-art depth estimation algorithm. The database has over 250 light-fields made publicly available for research. These light-fields have been captured using the Lytro Illum camera. The depth-maps, output images, and metadata obtained using the Lytro command line tool have also been included. As the second part of our project, we have implemented the depth estimation algorithm proposed by Tao et al. [7] This algorithm estimates depth by combining correspondence and defocus cues. The algorithm uses the complete 4D epipolar image to create a depth-map which proved to be qualitatively much better than the depth-map generated by the Lytro command line tool.

1. Light-field Database

Light-field [4] cameras have the ability to capture the intensity as well as direction of the light hitting the sensor. This makes it possible to refocus the image [5] as well as shift one's viewpoint of the scene after it has been captured. This also facilitates estimation of depth using light-fields. Light-fields have a wide range of potential applications in fields such as computer vision, and virtual reality. The main motivation for creating a light-field database is the fact that there are few publicly available light-field databases to conduct research in this emerging field. Our light-field database consist of over 250 light-fields of natural scenes captured using the Lytro Illum camera. The scenes and objects are captured taking into consideration different kinds of applications researchers in this field may potentially work on and consequently, these light-fields are divided into different categories. The 9 different categories included in the database are - (1) Flowers and plants, (2) Bikes, (3) Fruits and vegetables, (4) Cars, (5) Occlusions, (6) Buildings, (7) People, (8) Miscellaneous, (9) Refractive and Reflective surfaces. These external standardized light fields extracted using the Lytro command line tool help in recreating the 14×14 shifted views, which can then be used for various applications. Along with the light-fields, the depth-maps, processed images and metadata obtained using the command line tool have also been included. The light-field database can be accessed at http://lightfields.stanford.edu/.

2. Depth Estimation

Depth estimation is one of the major applications of light-fields. The multiple perspectives generated from a single light-field image can be used to estimate depth using correspondence. Moreover, the refocusing ability of light-fields allow the computation of defocus cues, which can be used to generate a depth-map as well. Tao et al. [7] propose an algorithm which combines both the cues to obtain a better estimation of depth. As the second part of our project, we implement and evaluate this depth estimation algorithm and compare it with the depth-maps generated by the Lytro command line tool. The algorithm exploits the complete 4D epipolar image (EPI) [2]. It computes and combines both defocus and correspondence cues to obtain a single high-quality depth-map. These kinds of accurate depth-maps may find applications in computer vision, 3D reconstruction, etc.

2.1. Related Work

Over the years, researchers have proposed different methods to represent light-fields. Adelson et al. [1] represented a light-field using a seven-dimensional plenoptic function which stores 3D intensity values, direction, wavelength and time of the light emitted from a scene. This was reduced to a 4D plenoptic function by Levoy and Hanrahan [4], which uses a 2-plane parametrization. This is the representation of light-fields used in our algorithm. Another important aspect of the algorithm is computing 4D EPIs, which was initially explored by Bolles et al. [2] EPIs are

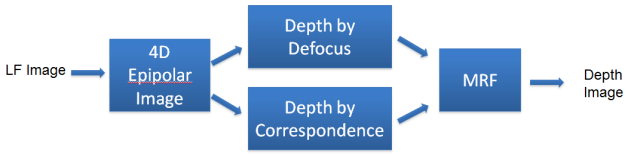


Figure 1. Pipeline of the depth estimation algorithm

explained in detail in 2.2.1.

Depth from correspondence cues has been studied extensively with stereo images. Initially, depth from defocus cues was made possible by taking images with focused at different distances and thus, creating a focal stack, or using multi-camera structures to obtain the defocus information in one exposure. The emergence of light-field cameras has made it possible to capture correspondence and defocus information in a single image. To the best of our knowledge, the algorithm which we are implementing is the first attempt to combine correspondence and defocus cues for light-field images. In one of the most recent works in depth estimation, Tao et al. [8] build upon this algorithm to use shading information, whereas Wang et al. [9] modify this approach to create occlusion-aware depth-maps.

2.2. The Algorithm

The depth estimation algorithm can be visualized as a pipeline consisting of three stages as shown in Figure 1. In the first stage, a 4D epipolar image (EPI) is constructed from the light field image. This 4D EPI is then sheared by a metric α. The second stage of the pipeline uses this sheared EPI to compute depth-maps using defocus and correspondence cues. In the third stage, both these depth-maps are combined using a MRF (Markov Random Fields) global optimization process.

2.2.1 4D Epipolar image

In light-field imaging, we capture multiple slightly shifted perspectives of the same scene. These perspectives can be combined to form an EPI. To understand the concept of EPIs, assume that we capture multiple views of a scene by shifting the camera horizontally in small steps. If we consider a 1D scan-line in the scene and stack the multiple views of this scan-line on top of another, we obtain a 2D EPI as shown in Figure 2

In this algorithm, we make use of the entire 4D EPI. Each light-field image captured by the Lytro Illum camera contains 14×14 slightly shifted perspectives. When these shifted views are stacked on top of one another, we obtain the 4D-EPI, $L_0(x, y, u, v)$. Here x and y are the spatial dimensions of the image and u and v are the dimensions

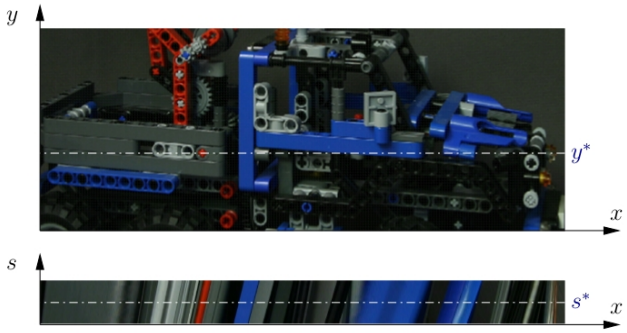


Figure 2. 2D Epipolar Image

corresponding to the location on the lens aperture.

Now, we shear the EPI by a metric α using the following formula:

$$L_\alpha(x, y, u, v) = L_0\left(x + u\left(1 - \tfrac{1}{\alpha}\right), y + v\left(1 - \tfrac{1}{\alpha}\right), u, v\right) \quad (1)$$

Here, L_0 is the input EPI and L_α is the sheared EPI.

2.2.2 Depth from defocus

To calculate depth from defocus, we first find the refocused image for the shear value α by summing over the u and v dimensions.

$$\overline{L}_\alpha(x, y) = \frac{1}{N} \sum_u \sum_v L_\alpha(x, y, u, v) \quad (2)$$

Here, N is the total number of images in the summation. Now, we can compute the defocus response by using:

$$D_\alpha(x, y) = \frac{1}{|W_D|} \sum_{(x,y) \in W_D} \left|\Delta \overline{L}_\alpha(x, y)\right| \quad (3)$$

Here Δ represents the 2D gradient operator and W_D is the window around the current pixel. $D_\alpha(x, y)$ is the defocus measure for the shear value α.

Once this is done for multiple values of the shear α, we find the α value that maximizes the defocus measure.

$$\alpha_D^*(x, y) = \arg\max_\alpha D_\alpha(x, y) \quad (4)$$

Now, $\alpha_D^*(x, y)$ is the depth-map obtained from defocus cues.

2.2.3 Depth from correspondence

To calculate depth from correspondence, we first calculate the angular variance for each pixel in the image for the shear

value α by using:

$$\sigma_\alpha(x,y)^2 = \frac{1}{N}\sum_u\sum_v \left(L_\alpha(x,y,u,v) - \overline{L}_\alpha(x,y)\right)^2 \tag{5}$$

Here, N is the total number of images in the summation and $\overline{L}_\alpha$ is the refocused image. To increase robustness, the variance is averaged across a small patch:

$$C_\alpha(x,y) = \frac{1}{|W_C|}\sum_{(x,y)\in W_C} \sigma_\alpha(x,y) \tag{6}$$

Here W_C is the window around the current pixel. $C_\alpha(x,y)$ is the correspondence measure for the shear value α.

Once this is done for multiple values of the shear α, we find the α value that minimizes the correspondence measure. The responses with low variances imply maximum correspondence.

$$\alpha^*_C(x,y) = \arg\min_\alpha C_\alpha(x,y) \tag{7}$$

Now, $\alpha^*_C(x,y)$ is the depth-map obtained from correspondence cues.

2.2.4 Confidence Measure

In order to combine the two cues, we need to find their confidence measures at each pixel. This is done by using the Peak Ratio metric introduced in [3]

$$D_{conf}(x,y) = \frac{D_{\alpha^*_D}(x,y)}{D_{\alpha^{*2}_D}(x,y)} \tag{8}$$

$$C_{conf}(x,y) = \frac{C_{\alpha^*_C}(x,y)}{C_{\alpha^{*2}_C}(x,y)} \tag{9}$$

where α^{*2} is the next local optimal value. This measure produces higher confidence when the value of the optima is farther away from the value of the next local optima.

2.2.5 Combining the Cues

Defocus and correspondences responses are finally combined using Markov Random Fields (MRFs) [6] as described in [7].

3. Results

The results of the depth-maps created using the methods described in [7] are compared to depth-maps extracted using the Lytro command line tool. In our implementation we used 256 different values of α between 0.2 and 2. We also used a window size of 9×9 for the window in both the defocus and correspondence depth estimates. The depth-maps are compared both qualitatively and quantitatively.

3.1. Qualitative Results

A qualitative comparison of the two algorithms can be made by examining the resulting depth-maps. The algorithm implemented in this paper is able to create a depth-map with much sharper contrast that allows for a large depth range. As seen in Figure 3, the shape of the plants in the background can easily be seen in the results of our algorithm, but in depth-map created with Lytros algorithm not much can be seen after the second plant in the scene. This shows that the algorithm implemented represents a better range of depth than the Lytro algorithm.

The trade offs between the defocus and correspondence cues can also be examined by comparing the output images. As mentioned in [7] the defocus cues work best in regions with high spatial gradients, and the correspondence cues provide accurate information in regions without strong gradients. However, the correspondence depth cues are more susceptible to noise in the image.

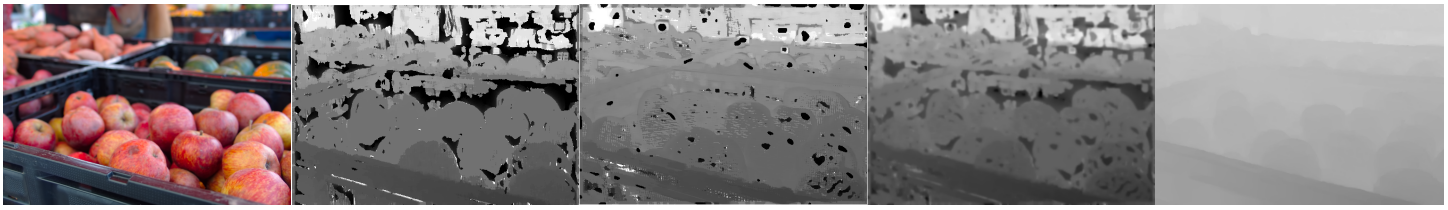
3.2. Quantitative Results

A quantitative comparison of the two algorithms was performed using a test image, where the distances were known. The image used was a bus schedule placed on a wall. The wall provides a fixed change in depth and the bus schedule adds gradients and other color information to the image. The angle of the camera with respect to the wall was calculated by measuring three distances between the camera lens and the wall. The first was a measurement between the camera and the wall in the direction the camera was pointing. The last two measurements were the shortest distance between the lens and the beginning of the poster and the end of the poster. Using basic geometry we were able to determine the angle between the image plane and the wall. Using this angle we can estimate what the change in depth should be for each pixel in the horizontal direction.

With the slope of the wall known, we now know the depth at each point along the wall. To compare the two algorithms, we calculated the mean square error (MSE) of the depth estimate, with the pixel values normalized between zero and one. This is done by first calculating the error for each pixel in the image. The error is the difference between the change in the intensity between two pixels in the horizontal direction and the expected change using the slope of the wall. The error for each pixel is then squared. The squared error is averaged over each measurement to give the resulting MSE.

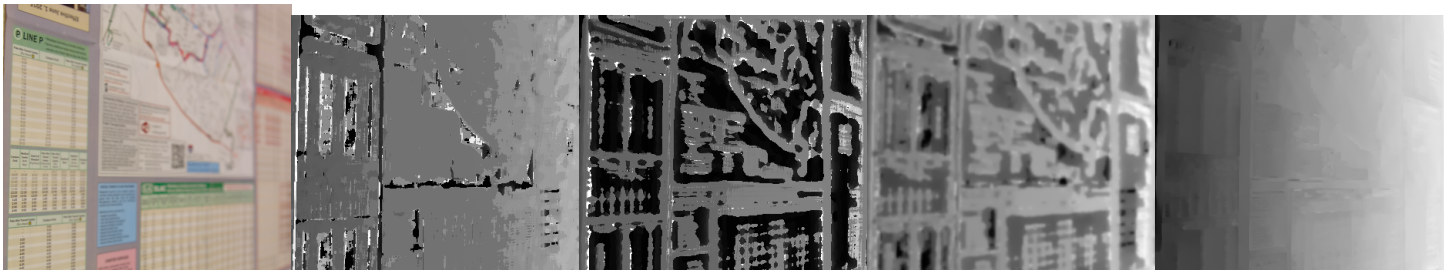
As shown in Table 1 the depth-map estimated using Lytro's method is more accurate than the depth-map we implemented. We believe this is due to the inaccurate labeling of the features on the bus schedule. In Figure 4, it can be seen that in our implementation the text on the bus schedule is highlighted as being farther away than the rest of the

(a) Plants RGB image (b) Plants Defocus depth (c) Plants Correspondence depth (d) Plants Final Depth (e) Plants Lytro Depth



(f) Apples Input RGB image (g) Apples Defocus depth (h) Apples Correspondence depth (i) Apples Final Depth (j) Apples Lytro Depth

Figure 3. Input images (a,f), depth-map based on defocus cues (b,g), depth-map from correspondence cues (c,h), final depth-map using Tao's method (d,i), Lytro's depth-map (e,j)



(a) Bus Schedule RGB image (b) Defocus depth (c) Correspondence depth (d) Final Depth (e) Lytro Depth

Figure 4. Input images (a), depth-map based on defocus cues (b), depth-map from correspondence cues (c), final depth-map using Tao's method (d), Lytro's depth-map (e). Note that the defocus image provides a much more accurate depth-map than the correspondence image. The main source of error occurs from merging the two sets of information.

Method	MSE
Lytro	4.2655e-06
Ours	3.8806e-05

Table 1. Results for the two algorithms

poster. These changes come from the correspondence cues in the image. The abrupt changes correspond to a high error calculation in our method of measuring the MSE. Based on the way we calculated the MSE, the correspondence cues have a negative impact on the depth estimation for this test image. While the algorithm we implemented may not always out perform Lytro's depth estimation, in general it appears to create depth-maps with a wider range of depths.

4. Conclusion

As the first part of this project, we have created a light-field database of natural scenes publicly available for research. Additionally, we evaluated a state-of-the-art depth estimation algorithm. The depth estimation algorithm, which was proposed by Tao et al. [7], leverages the multiple perspectives as well as refocusing abilities of a light-field by combining correspondence and defocus cues to generate a high-quality depth-map. Qualitatively, the depth-maps generated by this algorithm appeared to be better than the ones obtained using the Lytro command line tool. A quantitative evaluation of the algorithm was also done using a test image for which the depths in the scene were calculated. The MSE for Lytro's depth-map turned out to be better for the test image used. This is because the test image has a lot of

repeating features which were mislabeled by our algorithm, and correctly estimated by the algorithm implemented in the command line tool. However, our algorithm may give a lower MSE for other kinds of images.

5. Future Work

The time taken to generate depth-maps using the algorithm is much higher when compared to the time taken to generate depth-maps using the Lytro command line tool. Hence, an analysis of the trade-off between time and quality can be done in the future. One of the ways to decrease the computation time is to compromise on the number of bits used to represent depth. A lower-bit resolution will definitely affect the accuracy of the result, but may give running times beneficial for application which need depth-maps to be generated quickly.

The light-field database can be expanded by capturing scenes, which could be useful for specific applications in computer vision and image processing.

References

[1] E. H. Adelson and J. R. Bergen. *The plenoptic function and the elements of early vision*.

[2] R. C. Bolles, H. H. Baker, and D. H. Marimont. Epipolar-plane image analysis: An approach to determining structure from motion. *International Journal of Computer Vision*, 1(1):7–55, 1987.

[3] H. Hirschmüller, P. R. Innocent, and J. Garibaldi. Real-time correlation-based stereo vision with reduced border errors. *International Journal of Computer Vision*, 47(1-3):229–246, 2002.

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

[5] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan. Light field photography with a hand-held plenoptic camera.

[6] T. Pock, D. Cremers, H. Bischof, and A. Chambolle. Global solutions of variational models with convex regularization. *SIAM Journal on Imaging Sciences*, 3(4):1122–1145, 2010.

[7] M. 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.

[8] 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 (CVPR)*, 2015.

[9] T.-C. Wang, A. Efros, and R. Ramamoorthi. Occlusion-aware depth estimation using light-field cameras. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.