How Depth Estimation in Light Fields Can Benefit from Angular Super-Resolution?

Mandan Zhao[1], Gaochang Wu[2](✉), Yebin Liu[3], and Xiangyang Hao[1]

[1] Information Engineering University, Zhengzhou, China
mandanzhao@163.com, xiangyanghao2004@163.com
[2] Northeastern University, Shenyang, China
ahwgc2009@163.com
[3] Tsinghua University, Beijing, China
liuyebin@mail.tsinghua.edu.cn

Abstract. With the development of consumer light field cameras, the light field imaging has become an extensively used method for capturing the 3D appearance of a scene. The depth estimation often require a dense sampled light field in the angular domain. However, there is an inherent trade-off between the angular and spatial resolution of the light field. Recently, some studies for novel view synthesis or angular super-resolution from a sparse set of have been introduced. Rather than the conventional approaches that optimize the depth maps, these approaches focus on maximizing the quality of synthetic views. In this paper, we investigate how the depth estimation can benefit from these angular super-resolution methods. Specifically, we compare the qualities of the estimated depth using the original sparse sampled light fields and the reconstructed dense sampled light fields. Experiment results evaluate the enhanced depth maps using different view synthesis approaches.

Keywords: Light field · Angular super-resolution · View synthesis
Depth estimation

1 Introduction

Light field imaging [4,6] has emerged as a technology allowing to capture richer information from our world. Rather than a limited collection of 2D image, the light field camera is able to collect not only the accumulated intensity at each pixel but light rays from different directions. Recently, with the introduction of commercial and industrial light field cameras such as Lytro [1] and RayTrix [2], light field imaging has been one of a most extensively used method to capture 3D information of a scene.

However, due to restricted sensor resolution, light field cameras suffer from a trade-off between spatial and angular resolution. To mitigate this problem,

M. Zhao is student author.

J. Yang et al. (Eds.): CCCV 2017, Part I, CCIS 771, pp. 506–517, 2017.
https://doi.org/10.1007/978-981-10-7299-4_42

researchers have focused on novel view synthesis or angular super-resolution using a small set of views [7,8,10,15,16] with high spatial resolution. Typical view synthesis or angular super-resolution approaches first estimate the depth information, and then warp the existing images to the novel view based on the depth [12,16]. However, the depth-based view synthesis approaches rely heavily on the estimated depth, which is sensitive to noise, textureless and occluded regions. In recent years, some studies based on convolutional neural network (CNN) aiming at maximizing the quality of the synthetic views have been presented [5,14].

In this paper, we investigate how the depth estimation can benefit from these angular super-resolution methods. Specifically, we compare the qualities of the estimated depth using the original sparse sampled light fields and the reconstructed dense sampled light fields. Experiment results evaluate the enhanced depth maps using different view synthesis approaches.

2 Depth Estimation Using Angular Super-Resolved Light Fields

In this section, we describe the idea that uses super-resolved light fields in angular domain for depth estimation. We first investigate several angular super-resolution and view synthesis approaches. Then several depth estimation approaches are introduced using the super-resolved light field.

2.1 Angular Super-Resolution for Light Fields

Two angular super-resolution (view synthesis) approaches are investigated in the paper, which were proposed by Kalantari et al. [5] and Wu et al. [14]. Kalantari et al. [5] proposed a learning-based approach to synthesize novel views using a sparse set of input views. Specifically, they break down the process of view synthesis into disparity and color estimation and used two sequential CNNs to model them. In the disparity CNN (see Fig. 1), all the input views are first warped (backwarped) to the novel view with disparity range of $[-21, 21]$ and level of 100. Then the mean and standard deviation of all the warped input images are computed at each disparity level to form a feature vector of 200 channels.

In the color CNN, the feature vector is consisted of warped images, the estimated disparity and the position of the novel view, where the disparity is applied to occlusion boundaries detection and information collection from the adjacent regions, and the position of the novel view is used to assign the warped images with appropriate weights. The networks contain 4 convolutional layers with kernel sizes decreased from 7×7 to 1×1, where each layer is followed by a rectified linear unit (ReLU); the networks were trained simultaneously by minimizing the error between synthetic and ground truth views.

Unlike Kalantari et al. [5] that super-resolves light fields directly using images, Wu et al. [14] super-resolve light fields using EPIs. They indicated that

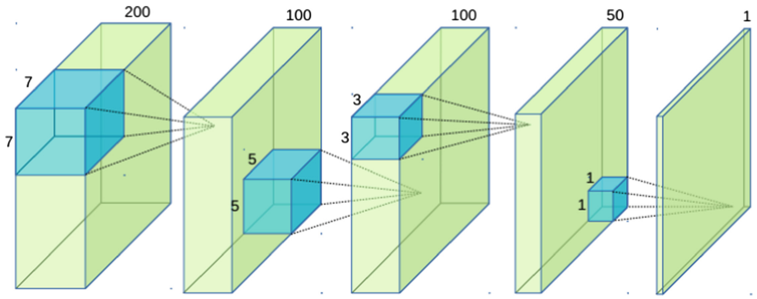


Fig. 1. The disparity CNN consists of four convolutional layers with decreasing kernel sizes. All the layers are followed by a rectified linear unit (ReLU). The color CNN has a similar architecture with different number of input and output channels.

the sparse sampled light field super-resolution involves information asymmetry between the spatial and angular dimensions, in which the high frequencies in angular dimensions are damaged by under-sampling. Therefore, they model the light field super-resolution as a learning-based angular high frequencies restoration on EPI.

Specifically, they first balance the information between the spatial and angular dimension by extracting the spatial low frequencies information. This is implemented by convolving the EPI with a Gaussian kernel. It should be noted that the kernel is defined in 1D space because only the low frequencies information in the spatial dimension are needed to be extracted. The EPI is then up-sampled to the desired resolution using bicubic interpolation in the angular dimension. Then a residual CNN is employed, which they called “detail restoration network” (see Fig. 2), to restore the high frequencies in the angular dimension. Different with the CNN proposed by Kalantari et al. [5], the detail restoration network is trained specifically to restore the high frequency portion in the angular dimension, rather than the entire information. Finally, a non-blind deblur is applied to recover the high frequencies depressed by EPI blur. Compared with the approach by Kalantari et al., Wu et al.’s approach has more flexible super-resolution factor; moreover, because of the depth-free framework, their approach achieves higher performance especially in occluded and transparent regions and non-Lambertian surfaces.

The architecture of the detail restoration network of Wu *et al.* is outlined in Fig. 2. Consider an EPI that is convolved with the blur kernel and up-sampled to the desired angular resolution, denoted as $\mathbf{E}'_L$ for short, the desired output EPI $f(\mathbf{E}'_L)$ is then the sum of the input $\mathbf{E}'_L$ and the predicted residual $\mathcal{R}(\mathbf{E}'_L)$:

$$f(\mathbf{E}'_L) = \mathbf{E}'_L + \mathcal{R}(\mathbf{E}'_L). \tag{1}$$

The network for the residual prediction comprises three convolution layers. The first layer contains 64 filters of size 1 × 9 × 9, where each filter operates on 9 × 9 spatial region across 64 channels (feature maps) and used for feature extraction. The second layer contains 32 filters of size 64 × 5 × 5 used for non-linear mapping.

The last layer contains 1 filter of size 32×5×5 used for detail reconstruction. Both the first and the second layers are followed by a rectified linear unit (ReLU). Due to the limited angular information of the light field used as the training dataset, we pad the data with zeros before every convolution operations to maintain the input and output at the same size.

The desired residuals are $\mathbf{r} = \mathbf{e}' - \mathbf{e}'_L$, where $\mathbf{e}'_L$ are the blurred and interpolated low angular resolution sub-EPIs. The goal is to minimize the mean squared error $\frac{1}{2}||\mathbf{e}' - f(\mathbf{e}'_L)||^2$. However, due to the residual network we use, the loss function is now formulated as follows:

$$L = \frac{1}{n}\sum_{i=1}^{n} ||\mathbf{r}^{(i)} - \mathcal{R}(\mathbf{e}'^{(i)}_L)||^2, \tag{2}$$

where n is the number of training sub-EPIs. The output of the network $\mathcal{R}(\mathbf{e}'_L)$ represents the restored detail, which must be added back to the input sub-EPI $\mathbf{e}'_L$ to obtain the final high angular resolution sub-EPI $f(\mathbf{e}'_L)$.

They apply this residual learning method for the following reasons. First, the undersampling in the angular domain damages the high frequency portion (detail) of the EPIs; thus, only that detail needs to be restored. Second, extracting this detail prevents the network from having to consider the low frequency part, which would be a waste of time and result in less accuracy.

Compared with the approach by Kalantari *et al.*, Wu *et al.*'s approach has more flexible super-resolution factor; moreover, because of the depth-free framework, their approach achieves higher performance especially in occluded and transparent regions and non-Lambertian surfaces.

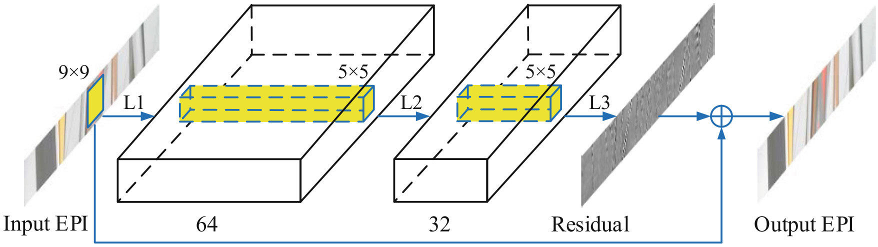


Fig. 2. Detail restoration network proposed by Wu et al. [14] is composed of three layers. The first and the second layers are followed by a ReLU. The final output of the network is the sum of the predicted residual (detail) and the input.

2.2 Depth Estimation for Light Fields

The Approach for Depth Measurement. In this subsection, we investigate several depth estimation approaches for light field data.

Tao et al. [9] proposed a depth estimation approach that combines depth cues from both defocus and correspondence using EPIs extracted from a light field. Since the slope of a line in an EPI is equivalent to a depth of a point in the scene [12], the EPIs are sheared to several possible depth values for computing defocus and correspondence cues responses. For a shear value α, a contrast-based measurement $\bar{L}_\alpha$ is performed at each pixel by averaging the intensity values in the angular dimension of the EPI. Then the defocus response D_α is measured by weighting the contrast-based measurements in a window in spatial dimension of the EPI. For the correspondence cue of a shear value α, the variance of each pixel in spatial dimension σ_α is computed, then the correspondence response C_α is the average of the variance values in a patch. After the computations of defocus and correspondence cue, a MRF global optimization is performed to obtain the final depth map.

Wang et al. [11] developed a depth estimation approach that treats occlusion explicitly. Their key insight is that the edge separating occluder and correct depth in the angular patch correspond to the same edge of occluder in the spatial domain. With this indication, the edges in the spatial image can be used to predict the edge orientations in the angular domain. First, the edges on the central pinhole image are detected using Canny operation. Based on the work by Tao et al. [9] and the occlusion theory described above, the initial local depth estimation is performed on the two regions in the angular patch of the sheared light field. In addition, a color consistency constraint is applied to prevent obtaining a reversed patch which will lead to incorrect depth estimation. Finally, the initial depth is refined with global regularization.

The Approach for 3D Measurement. In this subsection, we're going to focus on the 3D measurement of the light field imaging refocus: (a) We propose a novel digital refocusing algorithm to realize the refocus of the light field imaging. (b) We choose and refine the clarity-evaluation-function and interpolation method to establish the system of the evaluation and the refinement about light field imaging. (c) We take the corrected arithmetic mean filtering to refine the results of the 3D measurement.

(a) Refocus of the light field: Light fields transfer one plane to the other, the position coordinate can be defined as

$$[u', v', s', t'] = \begin{bmatrix} 1 & 0 & d & 0 \\ 0 & 1 & 0 & d \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot [u, v, s, t]^T, \tag{3}$$

where, $T_c = \begin{bmatrix} 1 & 0 & d & 0 \\ 0 & 1 & 0 & d \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$, $[u, v]$ represents the plane position of the camera array, $[s, t]$ represents the direction information.

First, we transform the light field information to the standard description of the light field. pix is the size of the pixel, and f_m is the focal length of the camera. We can get:

$$\begin{bmatrix} s \\ t \end{bmatrix} = \frac{pix}{f_m} \begin{bmatrix} p_x \\ p_y \end{bmatrix}, \tag{4}$$

$$\begin{bmatrix} u \\ v \end{bmatrix} = pix \begin{bmatrix} grid_x \\ grid_y \end{bmatrix}, \tag{5}$$

where, $[grid_x, grid_y]^T$ is the center of the camera pixel coordinate.

(b) Principle of focus ranging: Light field camera still follow the principle of optical imaging. On the contrary, if the object deviating from the ideal image plane, beyonding the scope of the depth of field, which can cause the focal and imaging blur.

$$\frac{1}{f} = -\frac{1}{u} + \frac{1}{v}, \tag{6}$$

where, f represents the focal length of the light field camera. u represents the object distance. v represents the image distance. Here we use virtual image for distance measurement, so the value of the object distance is negative. Since light field camera focal length is known, and therefore is obtained by heavy focus on clear imaging like distance, which can calculate the distance. The above is the basic principle of focusing range.

(c) Corrected arithmetic mean filtering: Due to the light distribution or upheaval gaussian noise image gradient, so we choose the adjusted arithmetic average filtering processing.

$$f(x, y) = \frac{1}{\sqrt{m^2 + n^2}} \sum_{(s,t) \in s_{xy}} g(s, t), \tag{7}$$

where, $s_x y$ represents rectangular window in the coordinate (x, y), which size is $m \times n$. $g(x, y)$ represents the jamming image.

3 Experimental Result

In this section, the proposed idea is evaluated on synthetic scenes and real-world scenes. We evaluate the quality of super-resolved light fields by measuring the PSNR values of synthetic views against ground truth images. And the quality of estimated depth maps using super-resolved light fields is compared with those using low angular resolution light fields. For synthetic scenes, ground truth depth maps are further applied for numerical evaluations.

3.1 Synthetic Scenes

The synthetic light fields in HCI datasets [13] are used for the evaluations. The input light fields have 3×3 views, where each view has a resolution of 768×768 (same as the original dataset); and the output angular resolution is 9×9 for comparison with the ground truth images.

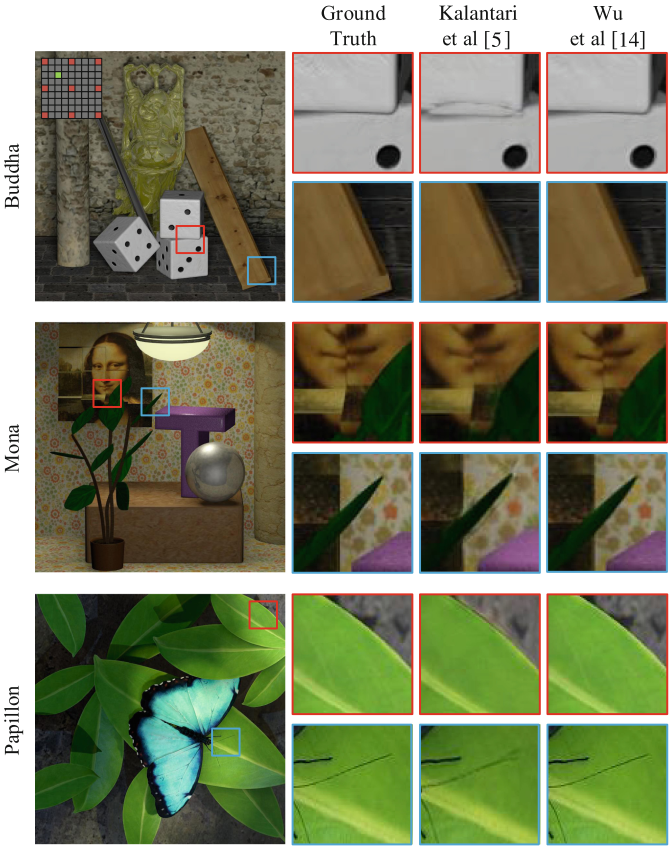


Fig. 3. Comparison of synthetic views produced by Kalantari et al.'s approach [5] and Wu et al.'s approach [14] on synthetic scenes. The ground truth views are shown on the left column, and the right columns show the close-up versions of the ground truth views, views produced by Kalantari et al.'s approach [5] and by Wu et al.'s approach [14], respectively.

Table 1. Quantitative results of reconstructed light fields on the synthetic scenes of the HCI datasets. The angular resolutions of input light fields are set to 3 × 3, and the output angular resolutions are 9 × 9.

	Buddha	*Mona*	*Papillon*
Kalantari [5]	34.05	32.53	28.26
Wu [14]	**43.20**	**44.37**	**48.55**

Table 1 shows a quantitative evaluation of the super-resolution approaches on synthetic scenes. The approach by Wu et al. [14] produces light fields of higher quality than those yielded by Kalantari et al. [5], because the CNNs in the latter approach are specifically trained on real-world scenes. Figure 3 shows the synthetic images in a certain viewpoint. The approach by Wu et al. [14] has better performance especially in the occluded regions, e.g., the board in the *Buddha* case, the leaves in the *Mona* case and the antenna in the *Papillon* case.

The numerical results of depth maps using the approaches by Tao et al. [9] and Wang et al. [11] are tabulated in Table 2. And Fig. 4 demonstrates the depth maps estimated by Wang et al.'s approach [11] on the *Buddha* using input low angular resolution (3 × 3) light field, ground truth high resolution (9 × 9) light field and super-resolved light fields (9 × 9) by Kalantari et al. [5] and Wu et al. [14], respectively. The depth estimation using super-resolved light fields show prominent improvement when compared with the results using input low resolution light fields. In addition, due to the better quality of synthetic views

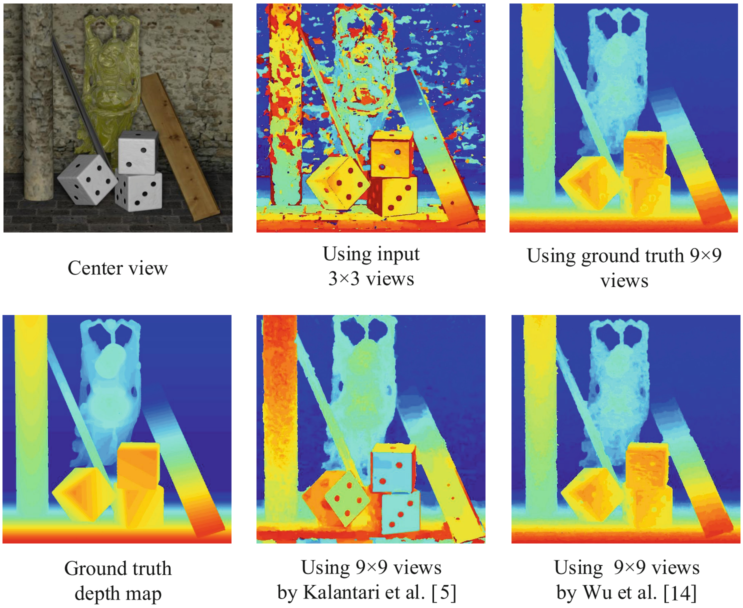

Center view
Using input 3×3 views
Using ground truth 9×9 views
Ground truth depth map
Using 9×9 views by Kalantari et al. [5]
Using 9×9 views by Wu et al. [14]

Fig. 4. Comparison of depth maps estimated by Wang et al.'s approach [11] using light fields of different angular resolutions on the *Buddha*.

Table 2. RMSE values of the estimated depth using the approaches by Tao et al. [9]/Wang et al. [11] on synthetic scenes of HCI datasets.

	Buddha	*Mona*	*Papillon*
Input views	0.2642/0.2926	0.2115/0.2541	0.1871/0.1533
Kalantari [5]	0.1721/0.1576	0.0876/0.0829	0.1665/0.1430
Wu [14]	0.0550/0.0401	0.0678/0.0517	0.0610/0.0532
GT Light Fields	0.0543/0.0393	0.0652/0.0529	0.0583/0.0455

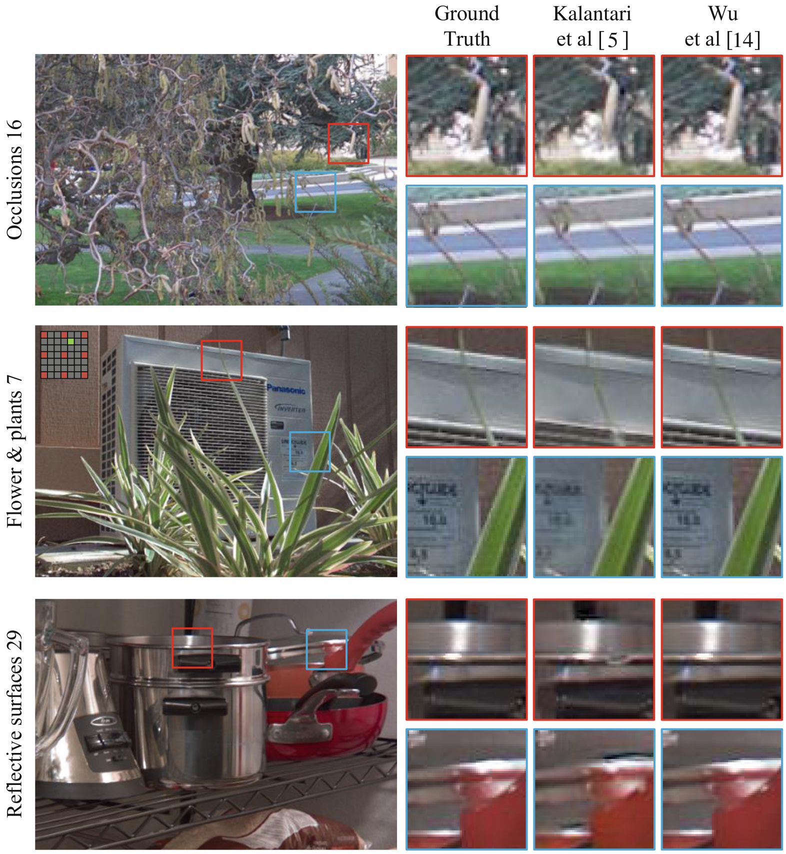


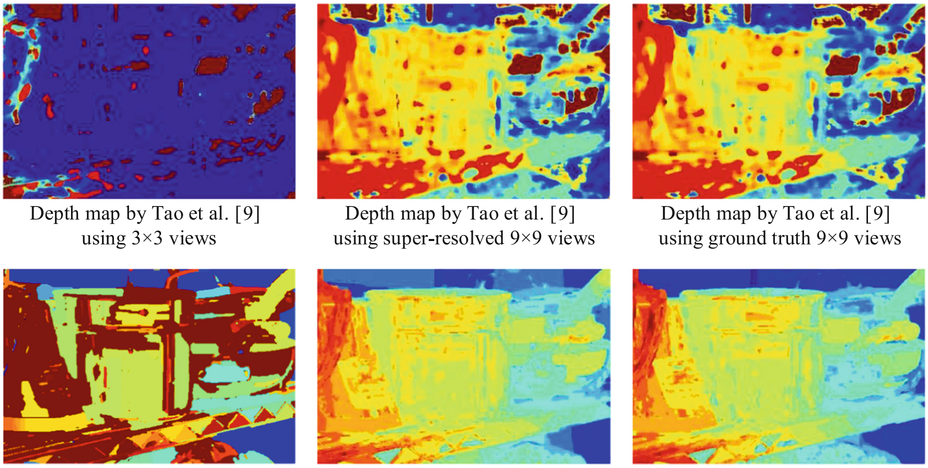
Fig. 5. Comparison of synthetic views produced by Kalantari et al.'s approach [5] and Wu et al.'s approach [14] on real-world scenes.

produced by Wu et al.'s approach [14], especially in the occluded regions, the estimated depth maps are more accurate than those using super-resolved light fields by Kalantari et al.'s approach [5].

3.2 Real-World Scenes

The Stanford Lytro Light Field Achieve [3] is used for the evaluation on real-world scenes. The dataset is divided into several categories including occlusions, and refractive and reflective surfaces, which are challenge cases to test the robustness of the approaches. We use 3 × 3 views to reconstruct 7 × 7 light fields.

Table 3 lists the numerical results of the super-resolution approaches on the real-world scenes. The approach by Wu et al. [14] shows better performance in terms of PSNR. Figure 5 shows some representative cases that contains complex occlusions or non-Lambertian surfaces. The networks proposed by Kalantari



Depth map by Tao et al. [9] using 3×3 views

Depth map by Tao et al. [9] using super-resolved 9×9 views

Depth map by Tao et al. [9] using ground truth 9×9 views

Depth map by Wang et al. [11] using 3×3 views

Depth map by Wang et al. [11] using super-resolved 9×9 views

Depth map by Wang et al. [11] using ground truth 9×9 views

Fig. 6. Comparison of depth maps estimated by Tao et al. [9] and Wang et al. [11] using light fields of different angular resolutions on the *Reflective surfaces* 29.

et al. [5] were specifically trained for Lambertian regions, thus tend to fail in the reflective surfaces, such as the pot and pan in the *Reflective* 29 case. In addition, due to the depth estimation-based framework, the synthetic views have ghosting and tearing artifacts in the occlusion boundaries, such as the branches in the *Occlusion* 16 case and the boundary between air-condition and grass in the *Flowers & plants* 7 case.

Figure 6 shows the depth maps estimated by Tao et al.'s approach [9] and Wang et al.'s approach [11] using input low angular resolution (3×3) light field, super-resolved light fields (7×7) by Wu et al. [14] and ground truth high resolution (7×7) light field, respectively. The quality of estimated depth maps are significantly improved using super-resolved light fields.

Table 3. Quantitative results of reconstructed light fields on the real-world scenes.

	Kalantari [5]	Wu [14]
Occlusions 2	28.90	**38.12**
Occlusions 16	32.24	**38.86**
Flowers & plants 7	26.70	**38.70**
Flowers & plants 12	34.97	**42.27**
Reflective surfaces 17	28.84	**42.28**
Reflective surfaces 29	37.70	**46.10**

4 Discussion and Future Work

There are several limitations of this idea. First, existing algorithms cannot handle the large parallax. They may occur the tearing artifacts when the parallax is too large. Second, light field super-resolution in the spatial domain is not considered. We believe that we can get the better results in depth estimation when jointing the spatial super-resolution. Finally, the computational efficiency of these approach is unsatisfactory. In the future work, we consider the super-resolution in the spatial domain and focus on the computational efficiency.

5 Conclusions and Discussion

We have presented an idea that uses an angular super-resolved light field to improve the quality of depth estimation. A straightforward way is to estimate a depth map using input low angular resolution light field, and render novel views using DIBR techniques. However, this approach always leads to error accumulation when recompute depth map using synthetic views. We therefore investigate approaches that directly minimize the quality of super-resolved light fields rather than depth maps. We evaluate this idea on synthetic scenes as well as real-world scenes, which contains non-Lambertian and reflective surfaces. The experimental results demonstrate that the quality of depth map is significantly improved using angular super-resolved light field.

References

1. Lytro. https://www.lytro.com/
2. RayTrix: 3D light field camera technology. http://www.raytrix.de/
3. Stanford Lytro Light Field Archive. http://lightfields.stanford.edu/
4. Ihrke, I., Restrepo, J.F., Mignard-Debise, L.: Principles of light field imaging: briefly revisiting 25 years of research. IEEE Signal Process. Mag. **33**(5), 59–69 (2016). https://doi.org/10.1109/MSP.2016.2582220
5. Kalantari, N.K., Wang, T.C., Ramamoorthi, R.: Learning-based view synthesis for light field cameras. ACM Trans. Graph. (TOG) **35**(6), 193–202 (2016)
6. Levoy, M., Hanrahan, P.: Light field rendering. In: Siggraph, pp. 31–42 (1996). https://doi.org/10.1145/237170.237199
7. Pujades, S., Devernay, F., Goldluecke, B.: Bayesian view synthesis and image-based rendering principles. In: CVPR, pp. 3906–3913 (2014)
8. Shi, L., Hassanieh, H., Davis, A., Katabi, D., Durand, F.: Light field reconstruction using sparsity in the continuous fourier domain. ACM TOG **34**(1), 12 (2014)
9. Tao, M.W., Hadap, S., Malik, J., Ramamoorthi, R.: Depth from combining defocus and correspondence using light-field cameras. In: ICCV, pp. 673–680 (2013)
10. Vagharshakyan, S., Bregovic, R., Gotchev, A.: Image based rendering technique via sparse representation in shearlet domain. In: ICIP, pp. 1379–1383. IEEE (2015)
11. Wang, T.C., Efros, A.A., Ramamoorthi, R.: Occlusion-aware depth estimation using light-field cameras. In: ICCV, pp. 3487–3495 (2015)
12. Wanner, S., Goldluecke, B.: Variational light field analysis for disparity estimation and super-resolution. IEEE TPAMI **36**(3), 606–619 (2014)

13. Wanner, S., Meister, S., Goldlücke, B.: Datasets and benchmarks for densely sampled 4D light fields. In: Vision, Modeling & Visualization, pp. 225–226 (2013)
14. Wu, G., Zhao, M., Wang, L., Chai, T., Dai, Q., Liu, Y.: Light field reconstruction using deep convolutional network on EPI. In: CVPR (2017)
15. Yoon, Y., Jeon, H.G., Yoo, D., Lee, J.Y., So Kweon, I.: Learning a deep convolutional network for light-field image super-resolution. In: CVPRW, pp. 24–32 (2015)
16. Zhang, Z., Liu, Y., Dai, Q.: Light field from micro-baseline image pair. In: CVPR, pp. 3800–3809 (2015)