A COST MINIMIZATION WITH LIGHT FIELD IN SCENE DEPTH MAP ESTIMATION

Yuchen Zhang *Hongkai Xiong*

Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

ABSTRACT

This paper proposes a scene depth map generation method based on lens-based light field cameras. In particular, it achieves the functions of the incident rays and the corresponding directional features, which can favor determining the coordinates of candidate space points. The light rays behind the aperture in the 4D light field would be converted into the rays before the aperture with their directions known. The light rays through the aperture center are denoted as reference light rays and keep their directions. With the probability (cost) of each reference light ray in each depth value, we obtain an initial depth map by selecting the depth value with minimum cost. It would be refined via multi-label optimization and weighted median filtering. Experimental results demonstrate the accuracy of the depth map estimated by the proposed method.

Index Terms— light field, depth map, directional features, multi-label optimization, weight median filtering

1. INTRODUCTION

In recent years, light field cameras [1, 2] have attracted more attention from both scientific and industrial researchers since the plenopic camera appears. Traditional cameras are designed to capture the amount of light radiation converging in each point in the image plane. However, the output is a 2D image with no more than the color information. The light field camera can provide more information, e.g. the full plenoptic function generated by each observed point, which includes the intensity of the light ray in all directions. Currently, there are two kinds of hand-held light field cameras [1, 2]. The first type was Lytro developed by Ren Ng using a micro-lens array in front of a camera sensor [3], which could generate a tiny sharp image of the lens aperture and estimate the directions of incoming rays through it. The second type is based on coded aperture [2], which means that a mask is set up in the aperture to encode the light rays. The additional information inherent in the light field allows a wide range of applications. For example, it could be made use of in light field rendering of computer graphics so as to generate a virtual viewpoint in the scene. Refocusing of the camera is another application based on light field data, which is a huge advantage in photography. Owing to the fact that the light field data contains multiple view images, the depth map can be desired for accurate estimation. For 3D scene reconstruction, 3D point cloud can be obtained by combining the depth map and the camera matrix.

In comparison with traditional approaches in stereo matching [4, 5], lenslet light field images exist very narrow baselines. Since blurriness emerges in the sub-pixel shift, the stereo matching based approaches do not work well. There are huge matching costs in stereo correspondence as well. Instead of correspondence matching, quite a few constraints have been considered in depth map estimation from a lenslet light field data. In 2014, Wanner and Goldluecke [6] calculated the vertical and horizontal slopes in the epipolar plane based on a structure tensor. Georgiev and Lumsdaine [7] have ever proposed a normalized cross correlation method between micro-lens images to produce the depth map. In 2012, Bishop and Favaro [8] indicated an interactive method based on light field data. In 2013, Tao et al. [9] introduced a framework combining the correspondences and defocus cues together in the disparity map estimation. In 2015, Hae-Gon Jeon et al. [10] obtained the disparity map by building the cost volume, and made refinement by multi-label optimization.

In this paper, we propose a novel depth map generation method based on a lens-based light field camera. Since the incident rays can be calculated by the corresponding emergent rays, we convert the light rays behind the aperture, which are recorded as 4D light field, into the rays before the aperture with their directions known. The light rays through the aperture center, which are denoted as reference light rays, do not change their directions. The candidate space points lie on the intersections of these rays and the plane with depth value d. Considering the adjacent light rays of the reference light rays, the probability of each reference light ray in each depth value is derived, which is represented as a cost. An initial depth map is generated by selecting the depth value with minimum cost. In order to enhance the smoothness, we refine the results via multi-label optimization and weighted median filtering.

2. SPACE POINTS VIA INCIDENT RAYS

First, we analyze the characteristics of the incident rays from the light field data, and then describe the features extracted from the incident rays. With these features along with their

The work was supported in part by the NSFC, under grants 61425011, U1201255, and 61271218.

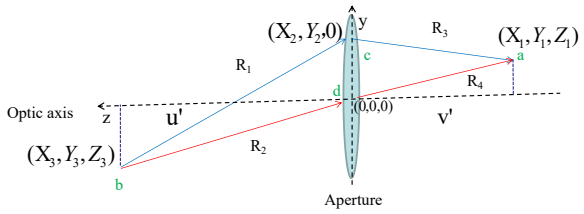

Fig. 1. b is the space point and a is the corresponding image point. Here, two light rays are represented as (R_1, R_3) and (R_2, R_4). The red light ray which propagates through the center of the aperture, is defined as a reference light ray. The blue light ray is in common.

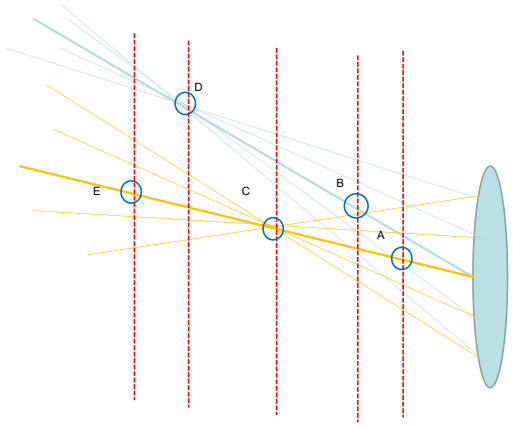

Fig. 2. The illustration for determining the reliable space points. C and D are the reliable points since all the rays with the same color are converge into these two points.

neighborhood, candidate space points are determined to generate a depth map.

2.1. Incident Light Rays Analysis

Let us denote the light field as $L(u, v, s, t)$, where (u, v) represents the main lens plane, and (s, t) means the image plane. Based on geometrical optics, light rays through the center of the main lens keep their directions. Considering light field, the rays can be represented by $L(u_0/2, v_0/2, i, j), i = 1, 2, ..., s_0, j = 1, 2, ..., t_0$, where (u_0, v_0) are the angular resolution, (s_0, t_0) are the spatial resolution. We seperate the light field into two groups, which are denoted by A_1 and A_2, respectively. A_1 is composed of light rays through the center of main lens, which can be formulated by $\{L(u_0/2, v_0/2, i, j), i = 1, 2, ..., s_0, j = 1, 2, ..., t_0\}$. The light rays in A_1 are denoted as the reference light rays. A_2 consists of the remaining light rays.

Fig. 1 reflects the relationship of the incident light rays and the emergent light rays. b is the object point and a is the corresponding image point. R_1 and R_2 are the incident light rays, while R_3 and R_4 are the corresponding emergent light rays. D_i denotes the direction vector of ray R_1. In order to obtain (X_1, Y_1, Z_1) and (X_2, Y_2, Z_2), in practice, we assume that the center of the image is on the principle optic axis of the aperture. We know the locate of point a in the image coordinate system and the length of one image pixel, so as to obtain its coordinate in the camera system. It is easy to know $D_4 = (X_1, Y_1, Z_1)$. Since the light rays R_2 and R_4 are in the same direction, $D_2 = (X_1, Y_1, Z_1)$. For ray R_3, the direction vector D_3 is $(X_1 - X_2, Y_1 - Y_2, Z_1)$. In order to obtain D_1, (X_3, Y_3, Z_3) needs to be calculated. Since the incident rays and the emergent rays are one to one mapping, we assume that the image point a is in focus. We denote the focus of the aperture as f, the object distance as u' and the image distance as v', then we have:

$$\frac{1}{u'} + \frac{1}{v'} = \frac{1}{f} \tag{1}$$

In Fig. 1, $v' = Z_1$, then we can obtain $u' = \frac{v'f}{v'-f}$. Since Z_3 is equal to u', based on the pinhole imaging model, we have the formula below:

$$\frac{X_3}{X_1} = \frac{Y_3}{Y_1} = \frac{Z_3}{Z_1} = \frac{f}{Z_1 - f} \tag{2}$$

Considering X_1, X_2, Y_1, Y_2, Z_1 and f are already known, we can simplify Eq. (2) as:

$$\begin{cases} X_3 = \frac{X_1 f}{Z_1 - f} \\ Y_3 = \frac{Y_1 f}{Z_1 - f} \\ Z_3 = \frac{Z_1 f}{Z_1 - f} \end{cases} \tag{3}$$

We denote (X, Y, Z) as an object point with depth Z in R_1, which can be obtained using:

$$\frac{X - X_2}{X_3 - X_2} = \frac{Y - Y_2}{Y_3 - Y_2} = \frac{Z}{Z_3} \tag{4}$$

Using Eq. (4), we can get the coordinate of the intersection of the ray R_1 and the plane with the function $z = Z$. We denote the directional feature of the light ray R_1 as $(p_1, p_2, p_3, p_4) = (\frac{1}{Z_3}(X_3 - X_2), \frac{1}{Z_3}(Y_3 - Y_2), X_2, Y_2)$. When the depth is d_i, the corresponding space point of ray R_1 is $(p_1 d_i + p_3, p_2 d_i + p_4, d_i)$. Using this feature, the candidate space point of ray R_1 can be obtained when the depth is determined.

2.2. Reliable Space Points Computation

For each light ray recorded by the camera, we can acquire the corresponding directional feature. Based on these features,

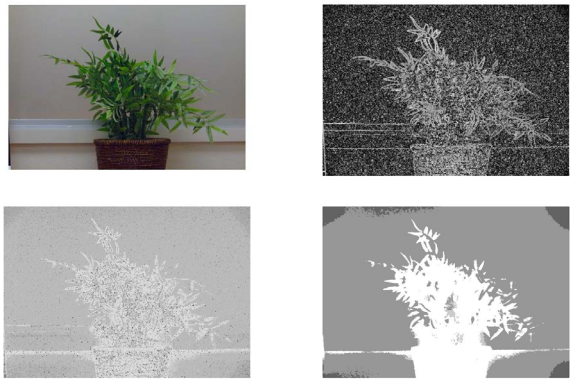

Fig. 3. The generated depth map samples. The upper left image is the reference view of the scene, and the upper right image is the depth map through the cost. The botom left image is the depth map taking local smoothness, and the botom right image is the final depth map after weighted median filtering.

the candidate space points can be attained. We denote d_i as a candidate depth value, where $i = 1, 2, ..., t$. For the plane $z = d_i$, the candidate space points are the intersections of the plane and the rays in A_i. Fig. 2 represents the effect for choosing a reliable space point with the ray directions known. Obviously, the points C and D are the correct object points for the light rays in the space, since the colors of the rays through the center of main lens and their neighborhood are same.

Considering the light field is four dimensional, we denote $R_{i,j,m,n}$ as the incident light ray for the corresponding light field $L(i, j, m, n)$. $R_{u_0/2,v_0/2,m,n}$ is in group A_1 and the other rays are in group A_2. We define $P_{i,j,m,n,k}$ as the candidate space point that is generated by the light ray $R_{i,j,m,n}$ in the plane $z = d_k$, which can be calculated using the corresponding feature and the depth. Let us denote cost $C_{m,n,k}$ for the space point generated by the light ray $R_{u_0/2,v_0/2,m,n}$ in the depth d_k. Thus, $C_{m,n,k}$ can be represented by:

$$C_{m,n,k} = \sum_{R_{i,j,a,b} \in N_{m,n,k}} \left(L(i, j, a, b) - L(u_0/2, v_0/2, m, n)\right)^2 \tag{5}$$

where $N_{m,n,k}$ consists of $u_0 v_0 - 1$ rays, and the corresponding space points have the minimum distances to $P_{u_0/2,v_0/2,m,n,k}$. For each depth label i and the corresponding image point (m, n), we calculate the cost $C_{m,n,k}$ for the space point. Obviously, the calculation of the space points is equivalent to the calculation of the depth map. The second image in Fig. 3 shows the resulted depth map via only the cost $C_{m,n,k}$. For the strong texture area, the depth map can be generated just based on the cost. However, for the weak texture area, the corresponding minimum costs are randomly distributed. Under this circumstance, the estimated depth map only from the cost is not reliable, and it is desirable to consider the neighboring area.

3. REFINEMENT

Once the candidate space points are calculated, for each ray $R_{u_0/2,v_0/2,m,n}$, we need to obtain a reliable space point in $\{P_{u_0/2,v_0/2,m,n,k}, k = 1, 2, ..., t\}$ in terms of the cost $\{C_{m,n,k}, k = 1, 2, ..., t\}$. Considering weak texture areas, multi-label optimization is leveraged to generate the depth map based on their neighborhood. After optimization, weighted median filtering is enabled to remove the noise points and smooth the depth map.

3.1. Multi-label Optimization

In order to correct the disparity map using neighboring estimation, we perform multi-label optimization using graph cuts. We denote $l(m, n)$ as the depth label for image point (m, n). The optimal depth map can be attained through minimizing

$$\begin{aligned} l_{\min} = \arg\min_{l} & \sum_{m,n} C_{m,n,l(m,n)} \\ & + \lambda \sum_{(m',n') \in N_{m,n}} ||l(m, n) - l(m', n')|| \end{aligned} \tag{6}$$

where $N_{m,n}$ contains the neighboring image points of (m, n). Eq. (6) consists of two terms, matching cost reliability ($C_{m,n,l(m,n)}$) and local smoothness ($||l(m, n) - l(m', n')||$). The third image in Fig. 3 shows an optimized depth map after multi-label optimization. It can be observed that many noise points still exist in the depth map, which will be refined by weighted median filtering.

3.2. Weighted Median Filtering

After multi-label optimization, the weighted median filtering [11] is used for the depth map refinement. Hence, the depth map points are weighted in the local histograms:

$$h(m, n, i) = \sum_{(m',n') \in N(m,n)} \omega(m, n, m', n') \delta(V(m', n') - i) \tag{7}$$

The weight $\omega(m, n, m', n')$ depends on the reference image V, which is denoted as $V(m, n) = L(u_0/2, v_0/2, m, n)$. The bilateral weight ω suppresses the pixels with different color from the center pixel. Finally, the median value is calculated by accumulating this histogram. The last image in Fig. 3 illustrates the depth map after weighted median filtering.

4. EXPERIMENTAL RESULTS

Considering the proposed approach employs the parameters of the light field camera, e.g. the sensor size, f-number and focal length, the performance was evaluated based on the datasets captured by the Lytro Illum. For a Lytro image, the total time cost is about 10 minutes, most of which was

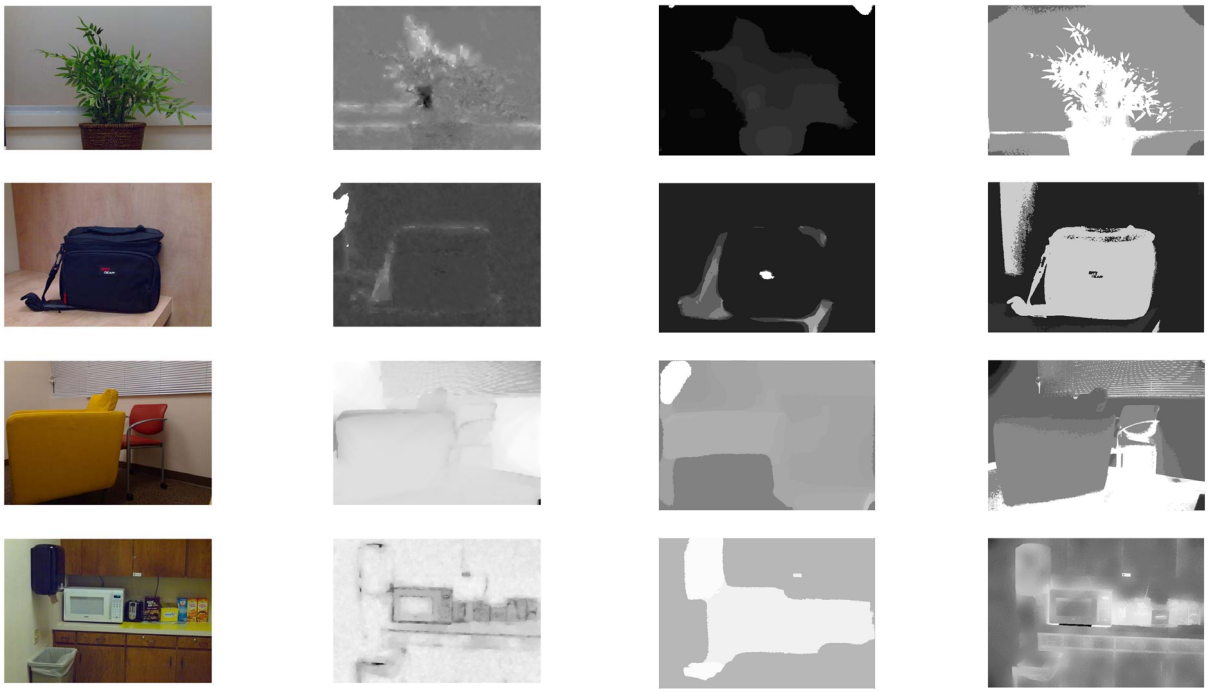

Fig. 4. The generated depth map results on four datasets. In each row, the images from left to right are the original image, the depth map generated by the Lytro desktop software, the depth map generated by Jeon's approach [10], and the proposed method.

spent by the computation of cost C in Eq. 5. The proposed approach was implemented in Matlab, and the operations were run on a Ubuntu 14.04 server with Intel Xeon CPU E5-2687W @ 3.10GHz and 256 GB memory.

The Lytro Illum camera is 8x optical zoom (30-250mm) at a constant f/2.0 aperture. The spatial resolution of the light field image is 433×625. The angular resolution is 15×15. The active area of the image sensor is $10.82mm \times 7.52mm$. The f-number of the camera is given by f/D, where f is the focal length, and D is the diameter of the effective aperture. The f-number of the camera is 2. In order to calculate the candidate space points which are represented in the camera system, the corresponding coordinates of the points (u, v) and (s, t), which are the image point in the aperture plane and the image plane respectively, are used to generate the four dimensional light field $L(u, v, s, t)$. Since the angular resolution is 15×15, the coordinate of (u, v) in the camera system is $((8 - u)f/30, (8 - v)f/30)$. Similarly, (s, t) is mapped to $(0.0174s, 0.0174t)$ in the image plane. After the raw data is obtained from the Lytro Illum, LFToolbox [12, 13] is used to decode and calibrate the raw data. After decoding and calibration, a light field function L is generated for later calculation.

Fig. 4 shows the results over four datasets, e.g. flower, bag, sofa and kitchen, compared with Jeon et al. [10] and the Lytro software. The focal lengths of these datasets are 75mm, 35mm, 32mm and 67mm respectively, while the magnifications are 35, 25, 33 and 36, respectively. It can be seen that both Lytro desktop software and Jeon's approach [10] would result in coarse and discontinuous effects. In specific, they cannot display the edges of the foreground objects and background objects correctly, while the proposed method could express the foreground and the background preciously enough.

5. CONCLUSION

To generate an accurate scene depth map, this paper proposes a cost minimization method based on lens-based light field. With the intrinsic parameters of the camera, it leverages the functions of the incident rays and the corresponding directional features to determine the coordinates of candidate space points. Through the defined cost for each candidate space point, an initial depth map would be obtained by selecting the depth value with minimum cost. Based on multi-label optimization, a reliable depth map was generated and refined using weighted median filtering.

6. REFERENCES

[1] Ren Ng, Marc Levoy, Mathieu Brédif, Gene Duval, Mark Horowitz, and Pat Hanrahan, "Light field photography with a hand-held plenoptic camera," *Computer Science Technical Report CSTR*, vol. 2, no. 11, 2005.

[2] Anat Levin, Rob Fergus, Frédo Durand, and William T Freeman, "Image and depth from a conventional camera with a coded aperture," in *ACM Transactions on Graphics (TOG)*. ACM, 2007, vol. 26, p. 70.

[3] Ren Ng, *Digital light field photography*, Ph.D. thesis, stanford university, 2006.

[4] Jian Sun, Nan-Ning Zheng, and Heung-Yeung Shum, "Stereo matching using belief propagation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 7, pp. 787–800, 2003.

[5] W Williem, Ramesh Raskar, and In Kyu Park, "Depth map estimation and colorization of anaglyph images using local color prior and reverse intensity distribution," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3460–3468.

[6] Sven Wanner and Bastian Goldluecke, "Variational light field analysis for disparity estimation and super-resolution," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 3, pp. 606–619, 2014.

[7] Todor Georgiev and Andrew Lumsdaine, "Reducing plenoptic camera artifacts," in *Computer Graphics Forum*. Wiley Online Library, 2010, vol. 29, pp. 1955–1968.

[8] Tom E Bishop and Paolo Favaro, "The light field camera: Extended depth of field, aliasing, and superresolution," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 5, pp. 972–986, 2012.

[9] Michael W Tao, Sunil Hadap, Jagannath Malik, and Ravi Ramamoorthi, "Depth from combining defocus and correspondence using light-field cameras," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 673–680.

[10] Hae-Gon Jeon, Jaesik Park, Gyeongmin Choe, Jinsun Park, Yunsu Bok, Yu-Wing Tai, and In So Kweon, "Accurate depth map estimation from a lenslet light field camera," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1547–1555.

[11] Ziyang Ma, Kaiming He, Yichen Wei, Jian Sun, and Enhua Wu, "Constant time weighted median filtering for stereo matching and beyond," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 49–56.

[12] Donald G. Dansereau, Oscar Pizarro, and Stefan B. Williams, "Decoding, calibration and rectification for lenselet-based plenoptic cameras," in *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*. IEEE, Jun 2013.

[13] Donald G. Dansereau, Oscar Pizarro, and Stefan B. Williams, "Linear volumetric focus for light field cameras," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 2, Feb. 2015.