DEPTH MAP ESTIMATION USING CENSUS TRANSFORM FOR LIGHT FIELD CAMERAS

Takayuki Tomioka, Kazu Mishiba, Yuji Oyamada and Katsuya Kondo

Department of Information and Electronics, Tottori University, Tottori, Japan

ABSTRACT

Depth estimation for the lense-array type cameras is a challenging problem because of sensor noise and radiometric distortion which is a global brightness change between sub-aperture images caused by a vignetting effect of the micro-lenses. We propose a depth map estimation method which has robustness against the sensor noise and the radiometric distortion. Our method first binarizes sub-aperture images by applying the census transform. Next, the binarized images are matched by computing the majority operations between corresponding bits and summing up the Hamming distance. An initial map obtained by matching has ambiguity caused by extremely short baselines among sub-aperture images. We refine an initial map by the optimization which uses the assumption that the variations of the depth values in the depth map and of the pixel values in the texture-less objects are similar. Experiments show that our method outperforms the conventional methods.

Index Terms— light field camera, Lytro, depth map, census transform

1. INTRODUCTION

Light field cameras capture the 4D light field of a scene by decoding 2D images. Light field cameras are categorized into two types, camera-array type and lense-array type, with respect to their system structures. The camera-array type consists of an array of multiple-cameras and recovers the 4D light field from images captured by the cameras [1]. On the other hand, the lense-array type puts an array of micro-lenses between the single main lens and the image sensor and recovers the 4D light field from the single image [2]. The single image can be decomposed into sub-images, called sub-aperture images. The lense-array type has the advantage of system portability. Some portable products have been released with reasonable price such as Lytro [3] and Raytrix [4]. Once the 4D light field is obtained, we can refocus the captured scene and synthesize arbitrary view point images. To realize those applications, we need to estimate scene depth from the 4D light field.

Depth estimation for the lense-array type cameras is a challenging problem [5, 2, 4, 6, 7, 8]. Unlike the camera-array type cameras, baseline between sub-aperture image pair is extremely shorter. Furthermore, lense vignetting effect on the sub-aperture images is non-negligible. Therefore, standard stereo matching methods [9] are not directly applicable as mentioned by Jeon $et\ al.$ [10].

Recent studies [10, 11, 6] have addressed the problem caused by extremely short baselines. Jeon $et\ al.$ [10] used the phase shift theorem in the Fourier domain to estimate the sub-pixel shifts of sub-aperture images. Kim $et\ al.$ [11] aggregated matching costs among all the sub-aperture images on cost volume to alleviate noise effects.

(a) (b) (c)

Fig. 1. (a) Center view of sub-aperture images taken by the Lytro camera [3]. (b) and (c) Initial depth map and output depth map by our method.

The depth estimation framework proposed by Tao $et\ al.$ [6] is simple and requires relatively lower computational cost. Their method combines defocus and correspondence cues from light fields. Since these cues are calculated using a pixel-based measure instead of a window-based metric, their method has less robustness against noise. In addition, these cues are affected by radiometric distortion of sub-aperture images caused by a vignetting effect of the micro-lenses.

In this paper, we propose a depth map estimation method for lense-array type cameras that has robustness against the radiometric distortion and the sensor noise. Our method consists of initial depth map estimation and optimization process. To reduce the influence of the radiometric distortion caused by a vignetting effect of the micro-lenses, the proposed method computes the matching cost in window-based measure for sub-aperture images transformed by the census transform. In addition, our cost calculation uses the majority operations to reduce the influence of noise. After initial depth map estimation, we refine an initial depth map by the optimization which uses the assumption that the variations of the depth values in the depth map and of the pixel values in the texture-less objects are similar. An example of the results of the proposed method is shown in Fig. 1.

2. REMAPPING SUB-APERTURE IMAGES FOR DEPTH ESTIMATION

We review depth estimation theory based on Epipolar plane image (EPI) analysis [12], which is the basic theory behind the proposed method, for lense-array type cameras in this section.

A lense-array type camera has a micro-lenses array placed on a regular grid and each micro-lense covers $n \times n$ pixels on the image sensor. A raw image captured by such camera can be decomposed into n^2 sub-aperture images [2]. Let f_s denote a sub-aperture image

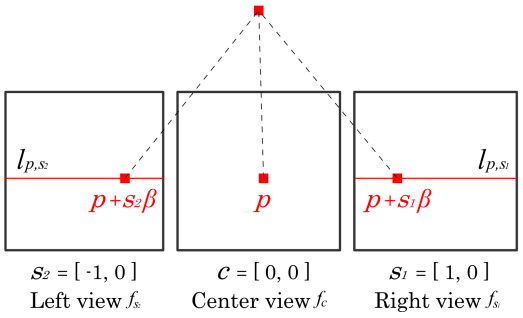

Fig. 2. Example of point correspondence in sub-aperture images. A point corresponding to pixel p in the center view can be written by $p + s\beta$ for all sub-aperture images.

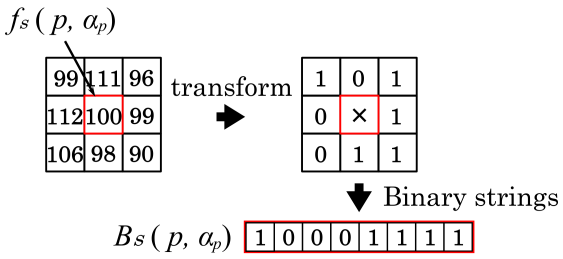

Fig. 3. Examples of census transform.

and $s = [u, v]^\top$ the relative position of the corresponding micro-lense on the array relative to the center view image $f_c, c = [0, 0]^\top$. Figure 2 shows an example of point correspondence in sub-aperture images.

To estimate depth from sub-aperture images, we perform a window-based matching method, as described in Sec. 3.1. Due to a large number of sub-aperture images, a computationally efficient matching method is required. Therefore, we perform depth estimation based on EPI analysis. EPI analysis estimates the depth of a point from orientation patterns observed on an epipolar plane image. As mentioned above, all the lenses are aligned on a grid and have the same optical axis. This indicates that the sub-aperture images can be regarded as images taken by parallel cameras located densely on a regular grid. This is the reason why EPI analysis is one of the well-used method to estimate depth map from light field camera images.

Suppose a point is projected on the pixel p in the center view image f_c. The corresponding point p_s on another sub-aperture image f_s is known to lie on the epipolar line $l_{p,s}$. Since these two images f_c and f_s are taken by parallel cameras, their epipoles are at infinity. Hence, the epipolar line $l_{p,s}$ is parallel to their baseline as

$$l_{p,s}(\beta) = p + s\beta, \tag{1}$$

where β is a disparity. Since the equation (1) holds for any sub-aperture image, the parameter β for a 3D point is same among all sub-aperture images. Thus, we can use the consistency of the parameter β to define energy function for depth estimation.

To achieve the search for the optimal value, we remap sub-aperture images as follows,

$$f_s(p, \alpha_p) = f_s(p + s\alpha_p), \tag{2}$$

where $f_s(p)$ is a pixel intensity of pixel p, α_p is a remap parameter corresponding to each pixel p and α_p corresponds to the above-mentioned parameter β. The remapping (2) is similar to [6]. Whereas [6] uses remapped images to shear the epipolar images [12] for defocus and correspondence analysis, we use them for window-based matching cost calculation. When all pixels on all sub-aperture images are remapped with the same α_p, we compute a matching cost among windows of the same coordinates for all sub-aperture images. Because α_p controls the disparity among sub-aperture images, α_p corresponds to a relative depth value. We calculate a matching cost while changing α_p as follows:

$$\alpha_p = 1 - \frac{1}{\alpha}, \tag{3}$$

where α changes from α_{min} to α_{max} at an interval of α_{step}.

3. DEPTH MAP ESTIMATION

Our method first estimates an initial depth map (Sec. 3.1). Next, we refine an initial depth map by the optimization (Sec. 3.2).

3.1. Initial depth estimation using census transform

Our method first binarizes remapped sub-aperture images by applying the census transform. Next, a matching cost among these images is computed using the majority operations between corresponding bits. Finally, we find the optimal value of remapped parameter α_p that minimizes the matching cost to estimate an initial depth map.

The census transform is a non-parametric local transform, which was first proposed by Zabih and Woodfill [13]. It computes a binary string by comparing a center pixel and its neighborhood pixels within the local window, as illustrated in Fig. 3. In this paper, we use a window of size 3×3. Since the census transform relies only on magnitude relationship among pixels, it is invariant under global brightness change. Thus the census transform is suitable for matching among the sub-aperture images which has the radiometric distortion caused by a vignetting effect of the micro-lenses.

We transform the target pixel p on remapped sub-aperture image into binary strings B_s as follows,

$$B_s(p, \alpha_p) = \bigotimes_{q \in N_p} \xi(f_s(p, \alpha_p), f_s(q, \alpha_p)), \tag{4}$$

where N_p is the neighborhood of p within the census window, $\otimes$ is a bit-wise concatenation operator and ξ is the step function defined as:

$$\xi(f_s(p, \alpha_p), f_s(q, \alpha_p)) = \begin{cases} 0 & \text{if } f_s(p, \alpha_p) \leq f_s(q, \alpha_p), \\ 1 & \text{otherwise.} \end{cases} \tag{5}$$

We apply the census transform to the gray image converted from the RGB color sub-aperture image.

Using binarized images, we compute a matching cost as follows:

$$C(p, \alpha_p) = \sum_{s \in S} H(G_S(p, \alpha_p), B_s(p, \alpha_p)), \tag{6}$$

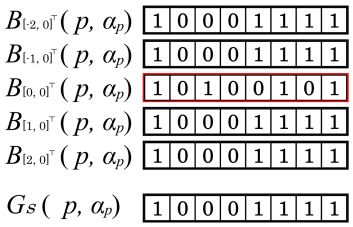

Fig. 4. Effect of majority operators. When an image of the center view is degraded with noise, its corresponding binary strings $B_{[0,0]^\top}$ are unreliable for matching. Using binary string G_S computed by majority operations is more reliable than using $B_{[0,0]^\top}$ because it can reduce noise effect.

where S is a set of all sub-aperture images, H is the Hamming distance function and $G_S(p, \alpha_p) = \phi(B_1(p, \alpha_p), ..., B_n(p, \alpha_p))$. Here, ϕ is a majority operation which outputs a binary string where the i-th bit is 1 if more than half of i-th bits of input binary strings are 1, and is 0 otherwise. Our proposed cost calculation is computationally effective because it simply computes the sum of the Hamming distance. In addition, it has robustness against noise. Let us consider the case where five sub-aperture images including the center view $f_{[0,0]^\top}$ degraded with noise. Figure 4 shows an example of binary strings B of a pixel by applying the census transform to sub-aperture images remapped with α_p which corresponds to true depth. Because the sub-aperture images are remapped with true α_p, a matching cost should be small. When the Hamming distance is calculated between the center view and another sub-aperture image, the matching cost becomes very high, which results in matching failure. One possible approach to avoid such failure is to compute matching costs among all possible pairs of sub-aperture images instead of the pairs of the center view and another sub-aperture image. This approach, however, takes much computation time due to a large number of sub-aperture images. The cost calculation method expressed in (6) is computationally efficient and has robustness against noise thanks to using a majority binary string for computation of the Hamming distance.

After aggregating the matching cost over α_p, we select the optimal value of α_p that minimizes the matching cost as an initial depth value:

$$Z_{init}(p) = \underset{\alpha_p}{\operatorname{argmin}} C(p, \alpha_p). \tag{7}$$

An example of the initial depth map is presented in Fig. 1 (b).

3.2. Depth map optimization

An initial depth map obtained by (7) has ambiguity caused by extremely short baselines among sub-aperture images. The depth values for real scenes have the following two features. First, the values inside of objects are constant or smoothly change. Second, the values at object boundary regions may greatly change. We refine an initial map by the optimization as follows,

$$\begin{aligned} Z^* = \underset{Z}{\operatorname{argmin}} \Big\{ & \sum_p \big\{ Z_{init}(p) - Z(p) \big\}^2 \\ & + \lambda_S \sum_p \big\{ D_x(Z,p)^2 + D_y(Z,p)^2 \big\} + \lambda_B \sum_p W(Z,p)^2 \Big\}, \end{aligned} \tag{8}$$

where λ_S and λ_B control the weights, $D_x(Z,p)$ and $D_y(Z,p)$ calculate the finite difference at pixel p on Z in horizontal and vertical directions, respectively, and $W(Z,p)$ is a difference filter as follows:

$$W(Z,p) = \sum_{q \in N_p} w(p,q) Z(q), \tag{9}$$

where N_p is the neighborhood of p within a local window and

$$w(p,q) = \begin{cases} \frac{1}{M} \exp(-d_{\text{geo}} - d_{\text{photo}}) & \text{if } p \neq q, \\ -1 & \text{otherwise,} \end{cases} \tag{10}$$

$$d_{\text{geo}} = \frac{\|p-q\|_2^2}{\sigma_s^2}, \tag{11}$$

$$d_{\text{photo}} = \frac{\|f_c(p) - f_c(q)\|_2^2}{\sigma_c^2}, \tag{12}$$

where M is the normalization factor satisfying $\sum w(p,q) = 0$, f_c is the sub-aperture image of the center view, and σ_s and σ_c are geometric and photometric spreads, respectively. Since the optimization problem in (8) is quadratic in Z, it yields a sparse system of linear equations.

The optimization of (8) consists of the first fidelity term and the second and third smoothness terms. The third term derives from the assumption that the variations of the depth values in the depth map and of the pixel values in the texture-less objects are similar.

4. EXPERIMENTAL RESULTS

To evaluate output depth map, we perform numerical evaluation experiment. We compared our method with two state-of-the-arts depth estimation methods, one for stereo camera by Rhemann *et al.* [9] and the other for lense-array type cameras by Tao *et al.* [6]. Since [9] is standard stereo matching method, we use two sub-aperture images at the left and right side of the center view as input images. We used the source code provided by the authors and default parameter setting was used. We use the first generation Lytro camera [3] which providing 81 sub-aperture images and take scenes consist of a main object in the foreground regions and the background region. In our implementation, the parameters α in (3) was $\alpha_{min} = 0.2$, $\alpha_{max} = 2$, and $\alpha_{step} = 0.002$. We performed (8) with $\lambda_S = 10$, $\lambda_B = 10$. The parameters in (10) was with $\sigma_d = 2$, $\sigma_c = 0.02$ and window size 7×7.

We evaluated occlusion boundary instead of the estimated depth values because it is difficult to obtain the true values of depth in real scenes. We manually marked occlusion boundary between the foreground and the background in the center view image as the ground truth of occlusion boundaries. As occlusion boundary detection, we first take derivative on the estimated depth map and then threshold the gradient map with a constant thresholding value. As evaluation metrics, we computed precision and recall between the ground truth and the detected occlusion boundary.

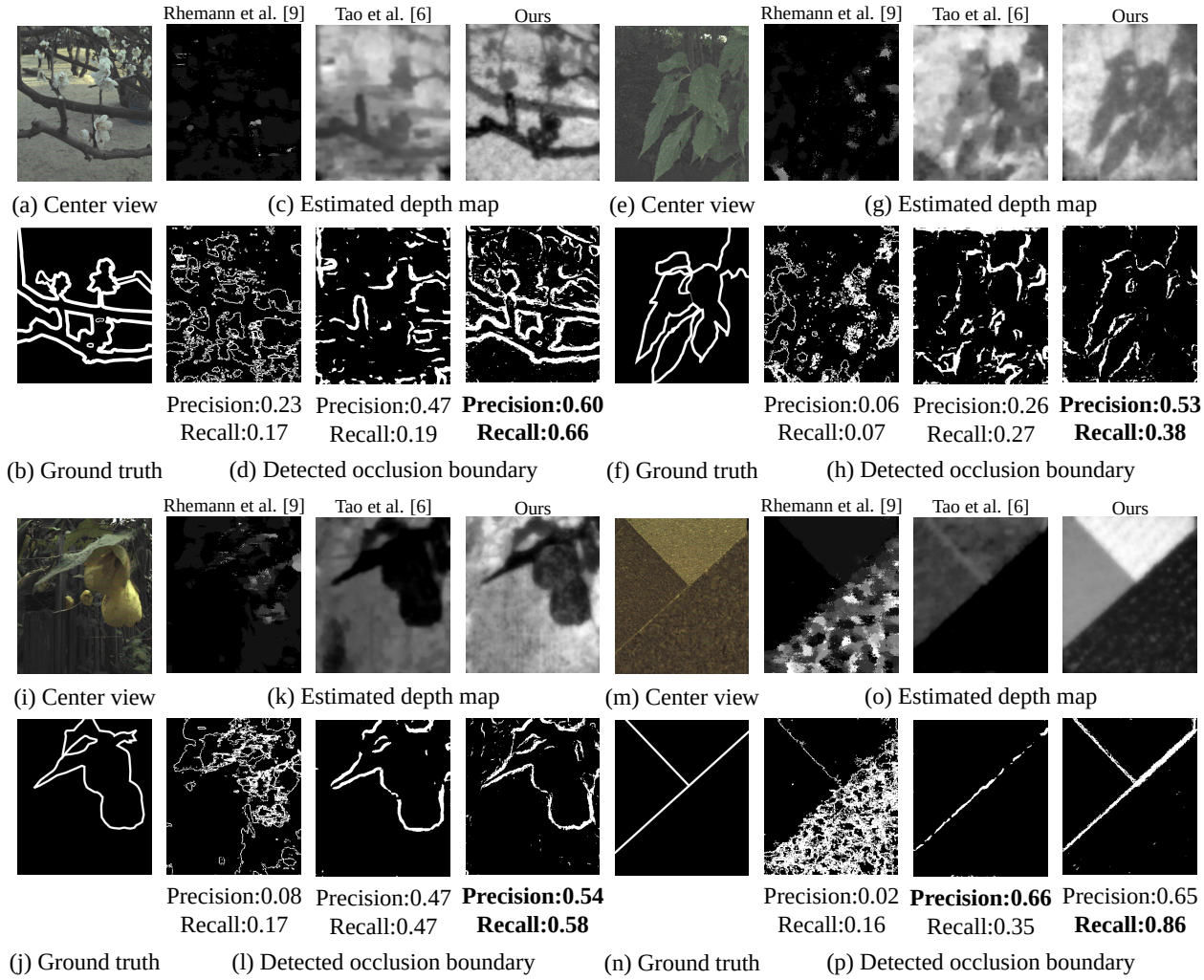

Fig. 5. Depth map result comparison.

Figure 5 shows the results with four scenes, TREE, LEAF, FLOWER and PLATE cases. Each scene consists of a main object in its foreground region as shown in Fig. 5 (a), (e), (i) and (m). Note that Figure 5 (a), (e), (i) and (m) is the image amplified the brightness of the original image to clearly display it. The sub-aperture images of LEAF case include signal noise caused by high ISO sensitivity. The sub-aperture images of PLATE case consist of three plates which have different depth. Figure 5 (b), (f), (j) and (n) show the manually obtained ground truth that has occlusion boundary on the border of the foreground object. The estimated depth map and occlusion boundary are shown in Fig. 5 (c), (g), (k) and (o) and Fig. 5 (d), (h), (l) and (p) respectively.

Contrast to the conventional methods [9, 6], our depth map exhibits higher performance visually. The method of Rhemann *et al.* [9] exhibits lower performance because of the short baselines of sub-aperture images. Although the method of Tao *et al.* [6] exhibits a high precision by combining advantage of the correspondence and defocus cues, it exhibits a low recall as compared to the our method. Because [6] has less robustness against the sensor noise and the radiometric distortion. As shown in LEAF case of high ISO condition, our method produces satisfactory result.

5. CONCLUSIONS

In this paper, we proposed a depth map estimation method which has robustness against the sensor noise and the radiometric distortion. To reduce the influence of the radiometric distortion caused by a vignetting effect of the micro-lenses, we used sub-aperture images binarized by the census transform for matching. Since we used majority operators in the cost calculation, our method has robustness against the sensor noise. Experiments showed that our method outperforms the conventional methods.

6. REFERENCES

[1] Bennett Wilburn, Neel Joshi, Vaibhav Vaish, Eino-Ville Talvala, Emilio Antunez, Adam Barth, Andrew Adams, Mark Horowitz, and Marc Levoy, "High performance imaging using large camera arrays," in *ACM SIGGRAPH 2005 Papers*, New York, NY, USA, pp. 765–776, ACM.

[2] Ren Ng, Marc Levoy, Mathieu Brédif, Gene Duval, Mark Horowitz, and Pat Hanrahan, "Light field photography with a hand-held plenoptic camera," Tech. Rep., Standord University, 2005.

[3] LYTRO, "The lytro camera," http://www.lytro.com/.

[4] Raytrix, "3d light field camera technology," http://www.raytrix.de/.

[5] Chia-Kai Liang, Tai-Hsu Lin, Bing-Yi Wong, Chi Liu, and Homer H. Chen, "Programmable aperture photography: Multiplexed light field acquisition," *ACM Trans. Graph.*, vol. 27, no. 3, pp. 55:1–55:10, Aug. 2008.

[6] Michael W. Tao, Sunil Hadap, Jitendra Malik, and Ravi Ramamoorthi, "Depth from combining defocus and correspondence using light-field cameras," in *International Conference on Computer Vision (ICCV)*, 2013.

[7] S. Wanner and B. Goldluecke, "Variational light field analysis for disparity estimation and super-resolution," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 36, no. 3, pp. 606–619, 2014.

[8] Zhan Yu, Xinqing Guo, Haibin Ling, Andrew Lumsdaine, and Jingyi Yu, "Line assisted light field triangulation and stereo matching," in *IEEE International Conference on Computer Vision, ICCV*, 2013, pp. 2792–2799.

[9] Christoph Rhemann, Asmaa Hosni, Michael Bleyer, Carsten Rother, and Margrit Gelautz, "Fast cost-volume filtering for visual correspondence and beyond," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[11] M.J. Kim, T.H. Oh, and I.S. Kweon, "Cost-aware depth map estimation for lytro camera," in *International Conference on Image Processing (ICIP)*, 2014, pp. 36–40.

[12] Robert C. Bolles, H. Harlyn Baker, and David H. Marimont, "Epipolar-plane image analysis: An approach to determining structure from motion," *International Journal of Computer Vision*, vol. 1, no. 1, pp. 7–55, 1987.

[13] Ramin Zabih and John Woodfill, "Non-parametric local transforms for computing visual correspondence," in *European Conference on Computer Vision (Vol. II)*, Secaucus, NJ, USA, 1994, ECCV '94, pp. 151–158, Springer-Verlag New York, Inc.