

Accurate Depth Map Estimation from a Lenslet Light Field Camera

Hae-Gon Jeon Jaesik Park Gyeongmin Choe Jinsun Park
Yunsu Bok Yu-Wing Tai In So Kweon

Korea Advanced Institute of Science and Technology (KAIST), Republic of Korea

[hgjeon, jspark, gmchoe, ysbok]@rcv.kaist.ac.kr
[zzangjinsun, yuwing]@gmail.com, iskweon@kaist.ac.kr

Abstract

This paper introduces an algorithm that accurately estimates depth maps using a lenslet light field camera. The proposed algorithm estimates the multi-view stereo correspondences with sub-pixel accuracy using the cost volume. The foundation for constructing accurate costs is threefold. First, the sub-aperture images are displaced using the phase shift theorem. Second, the gradient costs are adaptively aggregated using the angular coordinates of the light field. Third, the feature correspondences between the sub-aperture images are used as additional constraints. With the cost volume, the multi-label optimization propagates and corrects the depth map in the weak texture regions. Finally, the local depth map is iteratively refined through fitting the local quadratic function to estimate a non-discrete depth map. Because micro-lens images contain unexpected distortions, a method is also proposed that corrects this error. The effectiveness of the proposed algorithm is demonstrated through challenging real world examples and including comparisons with the performance of advanced depth estimation algorithms.

1. Introduction

The problem of estimating an accurate depth map from a lenslet light field camera, e.g. LytroTM [1] and RaytrixTM [19], is investigated. Different to conventional cameras, a light field camera captures not only a 2D image, but also the directions of the incoming light rays. The additional light directions allow the image to be re-focused and the depth map of a scene to be estimated as demonstrated in [12, 17, 19, 23, 26, 29].

Because the baseline between sub-aperture images from a lenslet light field camera is very narrow, directly applying the existing stereo matching algorithms such as [20] cannot produce satisfying results, even if the applied algorithm is a top ranked method in the Middlebury stereo matching benchmark. As reported in Yu *et al.* [29], the disparity range

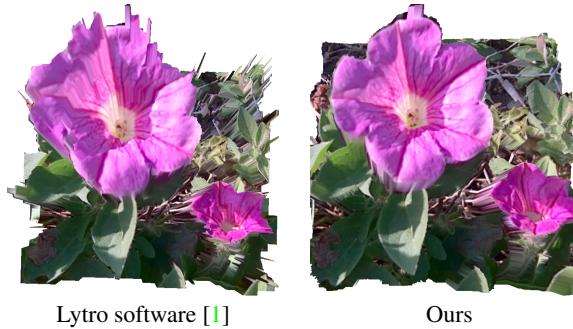


Figure 1. Synthesized views of the two depth maps acquired from Lytro software [1] and our approach.

of adjacent sub-aperture images in Lytro is between -1 to 1 pixels. Consequently, it is very challenging to estimate an accurate depth map because the one pixel disparity error is already a significant error in this problem.

In this paper, an algorithm for stereo matching between sub-aperture images with an extremely narrow baseline is presented. Central to the proposed algorithm is the use of the phase shift theorem in the Fourier domain to estimate the sub-pixel shifts of sub-aperture images. This enables the estimation of the stereo correspondences at sub-pixel accuracy, even with a very narrow baseline. The cost volume is computed to evaluate the matching cost of different disparity labels, which is defined using the similarity measurement between the sub-aperture images and the center view sub-aperture image shifted at different sub-pixel locations. Here, the gradient matching costs are adaptively aggregated based on the angular coordinates of the light field camera.

In order to reduce the effects of image noise, the weighted median filter was adopted to remove the noise in the cost volume, followed by using the multi-label optimization to propagate reliable disparity labels to the weak texture regions. In the multi-label optimization, confident matching correspondences between the center view and other views are used as additional constraints, which assist in preventing oversmoothing at the edges and texture regions. Finally, the estimated depth map is iteratively refined using quadratic polynomial interpolation to enhance the estimated depth map with sub-label precision.

In the experiments, it was found that a micro-lens image of lenslet light field cameras contains depth distortions. Therefore, a method of correcting this error is also presented. The effectiveness of the proposed algorithm is demonstrated using challenging real world examples that were captured by a Lytro camera, a Raytrix camera, and a lab-made lenslet light field camera. A performance comparison with advanced methods is also presented. An example of the results of the proposed method are presented in Fig. 1.

2. Related Work

Previous work related to depth map (or disparity map¹) estimation from a light field image is reviewed. Compared with conventional approaches in stereo matching, lenslet light field images have very narrow baselines. Consequently, approaches based on correspondence matching do not typically work well because the sub-pixel shift in the spatial domain usually involves interpolation with bluriness, and the matching costs of stereo correspondence are highly ambiguous. Therefore, instead of using correspondence matching, other cues and constraints were used to estimate the depth maps from a lenslet light field image.

Georgiev and Lumsdaine [7] computed a normalized cross correlation between microlens images in order to estimate the disparity map. Bishop and Favaro [4] introduced an iterative method for a multi-view stereo image for a light field. Wanner and Goldluecke [26] used a structure tensor to compute the vertical and horizontal slopes in the epipolar plane of a light field image, and they formulated the depth map estimation problem as a global optimization approach that was subject to the epipolar constraint. Yu *et al.* [29] analyzed the 3D geometry of lines in a light field image and computed the disparity maps through line matching between the sub-aperture images. Tao *et al.* [23] introduced a fusion method that uses the correspondences and defocus cues of a light field image to estimate the disparity maps. After the initial estimation, a multi-label optimization is applied in order to refine the estimated disparity map. Heber and Pock [8] estimated disparity maps using the low-rank structure regularization to align the sub-aperture images.

In addition to the aforementioned approaches, there have been recent studies that have estimated depth maps from light field images. For example, Kim *et al.* [10] estimated depth maps from a DSLR camera with movement, which simulated the multiple viewpoints of a light field image. Chen *et al.* [6] introduced a bilateral consistency metric on the surface camera in order to estimate the stereo correspondence in a light field image in the presence of occlusion. However, it should be noted that the baseline of the light field images presented in Kim *et al.* [10] and Chen *et al.* [6] are significantly larger than the baseline of the light field

images captured using a lenslet light field camera.

Compared with previous studies, the proposed algorithm computes the cost volume that is based on sub-pixel multi-view stereo matching. Unique in the proposed algorithm is the usage of the phase shift theorem when performing the sub-pixel shifts of sub-aperture image. The phase shift theorem allows the reconstruction of the sub-pixel shifted sub-aperture images without introducing blurriness in contrast to spatial domain interpolation. As is demonstrated in the experiments, the proposed algorithm is highly effective and outperforms the advanced algorithms in depth map estimation using a lenslet light field image.

3. Sub-aperture Image Analysis

First, the characteristics of sub-aperture images obtained from a lenslet-based light field camera are analyzed, and then the proposed distortion correction method is described.

3.1. Narrow Baseline Sub-aperture Image

Narrow baseline. According to the lenslet light field camera projection model proposed by Bok *et al.* [5], the viewpoint (S, T) of a sub-aperture image with an angular direction $s = (s, t)$ ² is as follows:

$$\begin{bmatrix} S \\ T \end{bmatrix} = \frac{D}{d} (D + d) \begin{bmatrix} s/f_x \\ t/f_y \end{bmatrix}, \quad (1)$$

where D is the distance between the lenslet array and the center of main lens, d is the distance between the lenslet array and imaging sensor, and f is the focal length of the main lens. With the assumption of a uniform focal length (i.e. $f_x = f_y = f$), the baseline between two adjacent sub-aperture images is defined as $\text{baseline} := \frac{(D+d)D}{df}$.

Based on this, we need to shorten f , shorten d , or lengthen D for a wider baseline. However, f cannot be too short because it is proportional to the angular resolution of the micro-lenses in a lenslet array. Therefore, the maximum baseline that is multiplication of the *baseline* and angular resolution of sub-aperture images remains unchanged even if the value of f varies. If the physical size of the micro-lenses is too large, the spatial resolution of the sub-aperture images is reduced. Shortening d enlarges the angular difference between the corresponding rays of adjacent pixels and might cause radial distortion of the micro-lenses. Finally, lengthening D increases the baseline, but the field of view is reduced. Due to these challenges, the disparity range of sub-aperture images is quite narrow. For example, the disparity range between adjacent sub-aperture views of the Lytro camera is smaller than ± 1 pixel [29].

²The 4D parameterization [7, 17, 26] is followed where the pixel coordinates of a light field image I are defined using the 4D parameters of (s, t, x, y) . Here, $s = (s, t)$ denotes the discrete index of the angular directions and $x = (x, y)$ denotes the Cartesian image coordinates of each sub-aperture image.

¹We sometimes use *disparity map* to represent *depth map*.

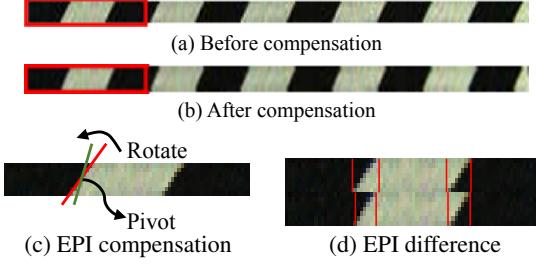


Figure 2. (a) and (b) EPI before and after distortion correction. (c) shows our compensation process for a pixel. (d) shows slope difference between two EPIS.

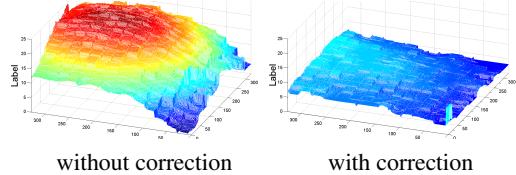


Figure 3. Disparity map before and after distortion correction (Sec. 3.2). Real-world planar scene is captured and the depth map is computed using our approach (Sec. 4).

Sub-aperture image distortion. From the analyses conducted in this study, it is observed that the lenslet light field images contain optical distortions that are caused by both the main lens (thin lens model) and micro-lenses (pinhole model). Although the radial distortion of the main lens can be calibrated using conventional methods, it is imperfect, particularly for light rays that have large angular differences from the optical axis. The distortion caused by these rays is called astigmatism [22]. Moreover, because the conventional distortion model is based on a pinhole camera model, the rays that do not pass through the center of the main lens cannot fit well to the model. The distortion caused by those rays is called field curvature [22]. Because they are the primary causes of the depth distortion, the two distortions are compensated in the following subsection.

3.2. Distortion Estimation and Correction

During the capture of a light field image of a planar object, spatially variant epipolar plane image (EPI) slopes (i.e. non-uniform depths) are observed that result from the distortions mentioned in Sec. 3.1 (see Fig. 3). In addition, the degree of distortion also varies for each sub-aperture image.

To solve this problem, an energy minimization problem is formulated under a constant depth assumption as follows:

$$\hat{G} = \underset{G}{\operatorname{argmin}} \sum_{\mathbf{x}} |\theta(I(\mathbf{x})) - \theta_o - G(\mathbf{x})| \quad (2)$$

where $|\cdot|$ denotes the absolute operator. θ_o , $\theta(\cdot)$, and $G(\cdot)$ denote the slope without distortion, the slope of EPI, and the amount of distortion at point \mathbf{x} , respectively.

The amount of field curvature distortion is estimated for

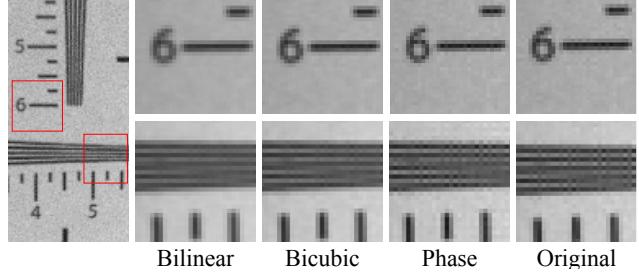


Figure 4. An original sub-aperture image is shifted with bilinear, bicubic and phase shift theorem.

each pixel. An image of a planar checkerboard is captured and compared with the observed EPI slopes with θ_o ³. Points with strong gradients in the EPI are selected and the difference ($\theta(\cdot) - \theta_o$) is calculated in Eq. (2). Then, the entire field curvature G is fitted to a second order polynomial surface model.

After solving Eq. (2), each point's EPI slope is rotated using \hat{G} . The pixel of reference view (i.e. center view) is set as the pivot of the rotation (see Fig. 2 (c)). However, due to the astigmatism, the field curvature varies according to the slice direction. In order to consider this problem, Eq. (2) is solved twice: once each for the horizontal and vertical directions. The correction order does not affect the compensation result. In order to avoid chromatic aberrations, the distortion parameters are estimated for each color channel. Figure 2 and Fig. 3 present the EPI image and estimated depth map before and after the proposed distortion correction, respectively⁴.

The proposed method is classified as a low order approach that targets the astigmatism and field curvature. A more generalized technique for correcting the aberration has been proposed by Ng and Hanrahan [16], and it is currently used for real products [2].

4. Depth Map Estimation

Given the distortion-corrected sub-aperture images, the goal is to estimate accurate dense depth maps. The proposed depth map estimation algorithm is developed using a cost-volume-based stereo [20]. In order to manage the narrow baseline between the sub-aperture images, the pipeline is tailored with three significant differences. First, instead of traversing the local patches to compute the cost volume, the sub-aperture images were directly shifted using a phase shift theorem and the per-pixel cost volume was computed. Second, in order to effectively aggregate the gradient costs computed from dozens of sub-aperture image pairs, a

³A tilt error might exist if the sensor and calibration plane are not parallel. In order to avoid this, an optical table is used.

⁴It is observed that altering the focal length and zooming parameters affect the correction. This is a limitation of the proposed method. However, it is also observed that the distortion parameter is not scene dependent.

weight term that considers the horizontal/vertical deviation in the st coordinates between the sub-aperture image pairs is defined. Third, because small viewpoint changes of sub-aperture images allow feature matching to be more reliable, a guidance of confident matching correspondences is also included in the discrete label optimization [11]. The details are described in following sub-sections.

4.1. Phase Shift based Sub-pixel Displacement

A key contribution of the proposed depth estimation algorithm is matching the narrow baseline sub-aperture images using sub-pixel displacements. According to the phase shift theorem, if an image I is shifted by $\Delta\mathbf{x} \in \mathbb{R}^2$, the corresponding phase shift in the 2D Fourier transform is:

$$\mathcal{F}\{I(\mathbf{x} + \Delta\mathbf{x})\} = \mathcal{F}\{I(\mathbf{x})\} \exp^{2\pi i \Delta\mathbf{x}}, \quad (3)$$

where $\mathcal{F}\{\cdot\}$ denotes the discrete 2D Fourier transform. In Eq. (3), multiplying the exponential term in the frequency domain is the same as convolving a Dirichlet kernel (or periodic sinc) in the spatial domain. According to the Nyquist-Shannon sampling theorem [21], a continuous band-limited signal can be perfectly reconstructed through convolving it with a sinc function. If the centroid of the sinc function is deviated from the origin, precisely shifted signals can be obtained. In the same manner, Eq. (3) generates a precisely shifted image in the spatial domain if the sub-aperture image is band-limited. Therefore, the sub-pixel shifted image $I'(\mathbf{x})$ is obtained using:

$$I'(\mathbf{x}) = I(\mathbf{x} + \Delta\mathbf{x}) = \mathcal{F}^{-1}\{\mathcal{F}\{I(\mathbf{x})\} \exp^{2\pi i \Delta\mathbf{x}}\}. \quad (4)$$

In practice, the light field image is not always a band-limited signal. This results from the weak pre-filtering that fits the light field into the sub-aperture image resolution [13, 24]. However, the artifact is not obvious for regions where the texture is obtained from the source surface in the scene. For example, a sub-aperture image of a resolution chart captured by Lytro camera is presented in Fig. 4. This image is shifted by $\Delta\mathbf{x} = [2.2345, -1.5938]$ pixels. Compared with the displacement that results from the bilinear and bicubic interpolations, the sub-pixel shifted image using the phase shift theorem is sharper and does not contain blurriness. Note that having an accurate reconstruction of sub-pixel shifted images is significant for accurate depth map estimations, particularly when the baseline is narrow. The effect of the interpolation method and depth accuracy is analyzed in Sec. 5.

In this implementation, the fast Fourier transform with a circular boundary condition is used to manage the non-infinite signals. Because the proposed algorithm shifts the entire sub-aperture image instead of local patches, the artifacts that result from periodicity problems only appear at the boundary of the image within a width of a few pixels (less than two pixels), which is negligible.

4.2. Building the Cost Volume

In order to match sub-aperture images, two complementary costs were used: the sum of absolute differences (SAD) and the sum of gradient differences (GRAD). The cost volume C is defined as a function of \mathbf{x} and cost label l :

$$C(\mathbf{x}, l) = \alpha C_A(\mathbf{x}, l) + (1-\alpha) C_G(\mathbf{x}, l), \quad (5)$$

where $\alpha \in [0, 1]$ adjusts the relative importance between the SAD cost C_A and GRAD cost C_G . C_A is defined as

$$C_A(\mathbf{x}, l) = \sum_{\mathbf{s} \in V} \sum_{\mathbf{x} \in R_{\mathbf{x}}} \min(|I(\mathbf{s}_c, \mathbf{x}) - I(\mathbf{s}, \mathbf{x} + \Delta\mathbf{x}(\mathbf{s}, l))|, \tau_1), \quad (6)$$

where $R_{\mathbf{x}}$ is a small rectangular region centered at \mathbf{x} ; τ_1 is a truncation value of a robust function; and V contains the st coordinate pixels \mathbf{s} , except for the center view \mathbf{s}_c . Equation (3) is used for precise sub-pixel shifting of the images. Equation (6) builds a matching cost through comparing the center sub-aperture image $I(\mathbf{s}_c, \mathbf{x})$ with the other sub-aperture images $I(\mathbf{s}, \mathbf{x})$ to generate a disparity map from a canonical viewpoint. The 2D shift vector $\Delta\mathbf{x}$ in Eq. (6) is defined as follows:

$$\Delta\mathbf{x}(\mathbf{s}, l) = lk(\mathbf{s} - \mathbf{s}_c), \quad (7)$$

where k is the unit of the label in pixels. $\Delta\mathbf{x}$ linearly increases as the angular deviations from the center viewpoint increase. Another cost volume C_G is defined as follows:

$$C_G(\mathbf{x}, l) = \sum_{\mathbf{s} \in V} \sum_{\mathbf{x} \in R_{\mathbf{x}}} \beta(\mathbf{s}) \min(\text{Diff}_x(\mathbf{s}_c, \mathbf{s}, \mathbf{x}, l), \tau_2) + (1 - \beta(\mathbf{s})) \min(\text{Diff}_y(\mathbf{s}_c, \mathbf{s}, \mathbf{x}, l), \tau_2) \quad (8)$$

where $\text{Diff}_x(\mathbf{s}_c, \mathbf{s}, \mathbf{x}, l) = |I_x(\mathbf{s}_c, \mathbf{x}) - I_x(\mathbf{s}, \mathbf{x} + \Delta\mathbf{x}(\mathbf{s}, l))|$ denotes the differences between the x -directional gradient of the sub-aperture images. Diff_y is defined similarly on the y -directional gradients. τ_2 is a truncation constant that suppresses outliers. $\beta(\mathbf{s})$ in Eq. (8) controls the relative importance of the two directional gradient differences based on the relative st coordinates. $\beta(\mathbf{s})$ is defined as follows:

$$\beta(\mathbf{s}) = \frac{|s - s_c|}{|s - s_c| + |t - t_c|}. \quad (9)$$

According to Eq. (9), β increases if the target view \mathbf{s} is located at the horizontal extent of the center view \mathbf{s}_c . In this case, only the gradient costs in the x direction are aggregated to C_G . Note that β is independent of the scene because it is determined purely using the relative position between \mathbf{s} and \mathbf{s}_c .

As a sequential step, every cost slice is refined using an edge-preserving filter [15] to alleviate the coarsely scattered unreliable matches. Here, the central sub-aperture image is used to determine the weights used for the filter. They are

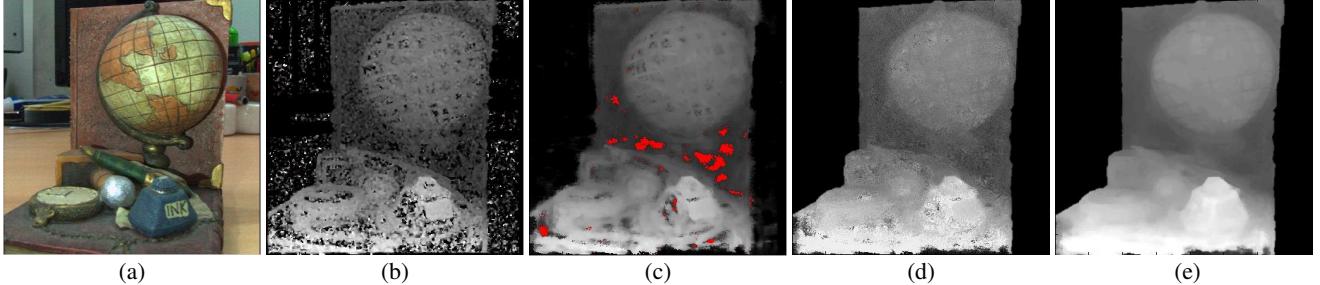


Figure 5. Estimated disparity maps at different step of our algorithm. (a) The center view sub-aperture image. (b)-(e) Disparity maps (b) based on the initial cost volume (winner-takes-all strategy), (c) after weighted median filter refinement (The red pixels indicates detected outlier pixels), (d) after the multi-label optimization, and (e) after the iterative refinement. The processes in (b) and (c) are described in Sec. 4.2, and the processes in (d) and (e) are described in Sec. 4.3 respectively.

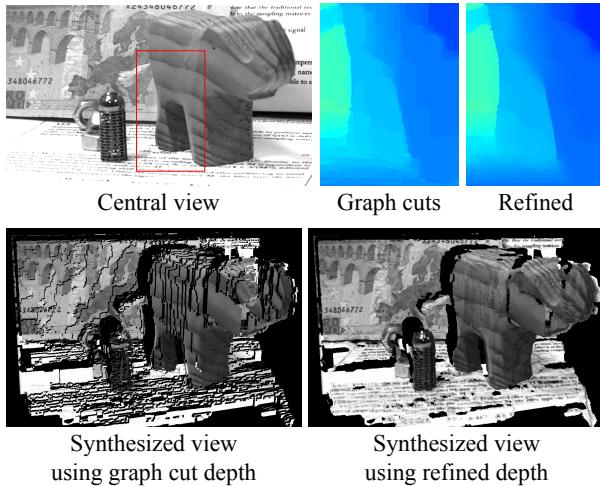


Figure 6. The effectiveness of the iterative refinement step described in Sec. 4.3.

determined using the Euclidean distances between the RGB values of two pixels in the filter, which preserves the discontinuity in the cost slices. From the refined cost volume C' , a disparity map l_a is determined using the winner-takes-all strategy. As depicted in Figs. 5 (b) and (c), the noisy background disparities are substituted with the majority value (almost zero in this example) of the background disparity. In each pixel, if the variance over the cost slices is smaller than a threshold τ_{reject} , this pixel is regarded as an outlier because it does not have distinctive minimum values. The red pixels in Fig. 5 (c) indicate these outlier pixels.

4.3. Disparity Optimization and Enhancement

The disparity map from the previous step is enhanced through discrete optimization and iterative quadratic fitting.

Confident matching correspondences. Besides the cost volume, the correspondences are also matched at salient feature points as strong guides for multi-label optimization. In particular, local feature matching is conducted between the center sub-aperture image and other sub-aperture im-

ages. Here, the SIFT algorithm [14] is used for the feature extraction and matching. From a pair of matched feature positions, the positional deviation $\Delta f \in \mathbb{R}^2$ in the xy coordinates is computed. If the amount of deviation $\|\Delta f\|$ exceeds the maximum disparity range of the light field camera, they are rejected as outliers. For each pair of matched positions, given $s, s_c, \Delta f$, and k , an over-determined linear equation $\Delta f = lk(s - s_c)$ is solved for l . This is based on the linear relationship depicted in Eq. (7). Because the feature point in the center view is matched with that of multiple images, it has several candidates for disparities. Therefore, their median value is obtained and used to compute the sparse and confident disparities l_c .

Multi-label optimization. Multi-label optimization is performed using graph cuts [11] to propagate and correct the disparities using neighboring estimation. The optimal disparity map is obtained through minimizing

$$l_r = \operatorname{argmin}_l \sum_x C'(x, l(x)) + \lambda_1 \sum_{x \in \mathcal{I}} \|l(x) - l_a(x)\| + \lambda_2 \sum_{x \in \mathcal{M}} \|l(x) - l_c(x)\| + \lambda_3 \sum_{x' \in \mathcal{N}_x} \|l(x) - l(x')\|, \quad (10)$$

where \mathcal{I} contains inlier pixels that are determined in the previous step in Sec. 4.2, and \mathcal{M} denotes the pixels that have confident matching correspondences. Equation (10) has four terms: matching cost reliability ($C'(x, l(x))$), data fidelity ($\|l(x) - l_a(x)\|$), confident matching cost ($\|l(x) - l_c(x)\|$), and local smoothness ($\|l(x) - l(x')\|$). Figure 5 (d) presents a corrected depth map after the discrete optimization. Note that even without the confident matching cost, the proposed approach estimates a reliable disparity map. The confident matching cost further enhances the estimated disparity at regions with salient matching.

Iterative refinement. The last step refines the discrete disparity map after the multi-label optimization into a continuous disparity with sharp gradients at depth discontinuities. The method presented by Yang *et al.* [28] is adopted. A new cost volume \hat{C} that is filled with one is computed. Then, for

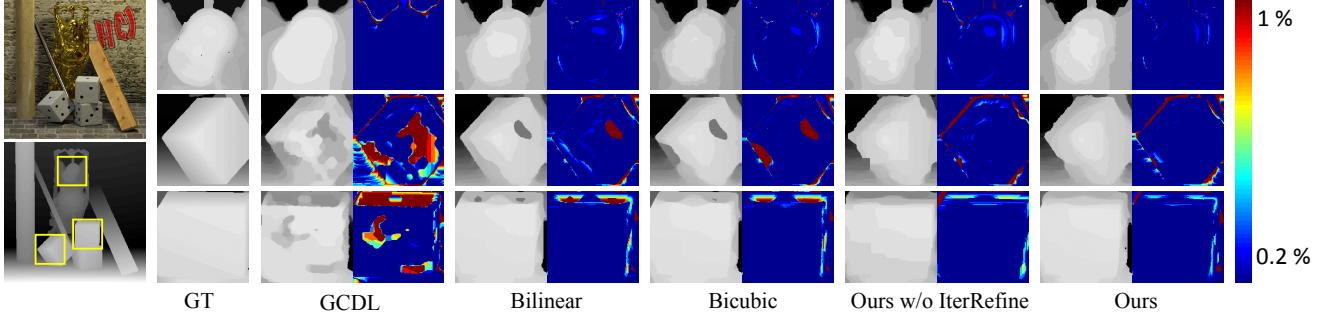


Figure 8. Zoom-up images of the Buddha2 dataset [27]. The error maps correspond to a relative depth error of less than 1%.

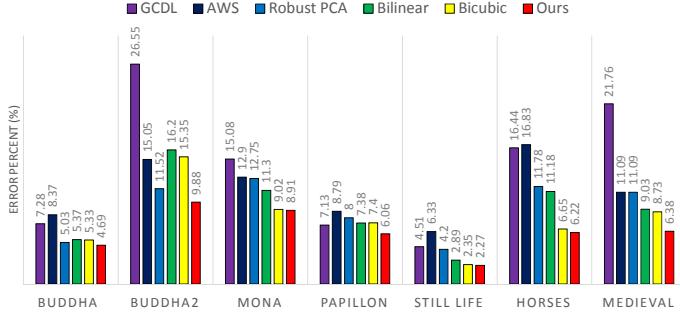


Figure 7. Relative depth measures on the results of GCDL [26], AWS [9], Robust PCA [8], and our approach under synthetic light fields benchmark [27]. The values indicate the percentage of erroneous pixels exceed a relative depth error of more than 0.2%. The error values of AWS and Robust PCA are from [8].

every \mathbf{x} , $\hat{C}(\mathbf{x}, l_r(\mathbf{x}))$ is set to 0, followed by weighted median filtering [15] of the cost slices. Finally, a non-discrete disparity l^* is obtained via

$$l^* = l_r - \frac{C(l_+) - C(l_-)}{2(C(l_+) + C(l_-) - 2C(l_r))}, \quad (11)$$

where $l_+ (= l_r + 1)$ and $l_- (= l_r - 1)$ are the adjacent cost slices of l_r . Here, \mathbf{x} in Eq. (11) is omitted for simplicity. l^* is the disparity map with the minimum cost, and it is derived from the least square quadratic fitting over three costs: $C(l_r)$, $C(l_+)$, and $C(l_-)$. Using the refined disparity, the overall procedure is applied again for better results. Figure 6 presents a discrete disparity map that was obtained from Eq. (10) and a continuous disparity map after the refinement. It can be seen that four iterations are sufficient for appropriate results.

5. Experimental Results

The performance of the proposed algorithm was evaluated using synthetic and real world datasets. The 4D Light Fields benchmark dataset [27] was used for synthetic evaluation. For the real world experiments, the images captured using three lenslet based light field cameras were used: Lytro™, Raytrix™, and the lab-made light field camera.

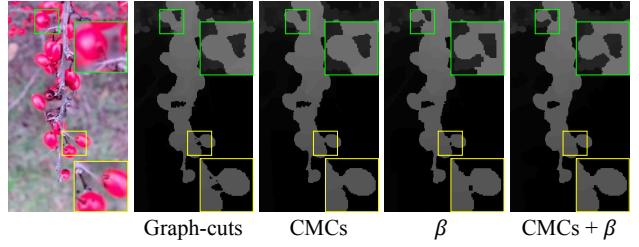


Figure 9. Evaluation on the role of β (in Sec. 4.2) and confidence matching correspondences (CMCs) (in Sec. 4.3).

The proposed algorithm required six minutes for the Lytro images and 25 minutes for the synthetic dataset. Among all computation steps, the building of the cost volume (Sec. 4.2) was the most time-consuming. The proposed algorithm is implemented in Matlab™, but it is expected that there would be a significant increase in speed if this step is parallelized using GPU. A machine equipped with an Intel i7 3.40 GHz CPU and 16 GB RAM was used for the computations. For the evaluation, the variables were empirically selected as $\alpha = 0.3$, $\tau_1 = 1$, $\tau_2 = 3$, $\tau_{reject} = 5$, $\lambda_1 = 0.5$, $\lambda_2 = 10$, and $\lambda_3 = 1$. Note that k varied according to the dataset, and this is described individually. The source code and dataset are released in our website⁵.

5.1. Synthetic Dataset Results

For the quantitative evaluation, the proposed method was compared with three advanced algorithms: active wavefront sampling based matching (AWS) [9], globally consistent depth labeling (GCDL) [26], and robust PCA [8]. The proposed algorithm was also evaluated through changing the sub-pixel shift methods (bilinear, bicubic, and phase, which are discussed in Sec. 4.1) while maintaining the other parameters consistent.

The benchmark dataset [27] used for validation was composed of a 9×9 angular resolution of sub-aperture images with $0.5 \sim 0.7$ mega pixels. The disparity between two adjacent sub-aperture images in the st domain was smaller than 3 pixels. For this dataset, $k = 0.03$ was used. As sug-

⁵https://sites.google.com/site/hgjeoncv/home/depthfromlf_cvpr15

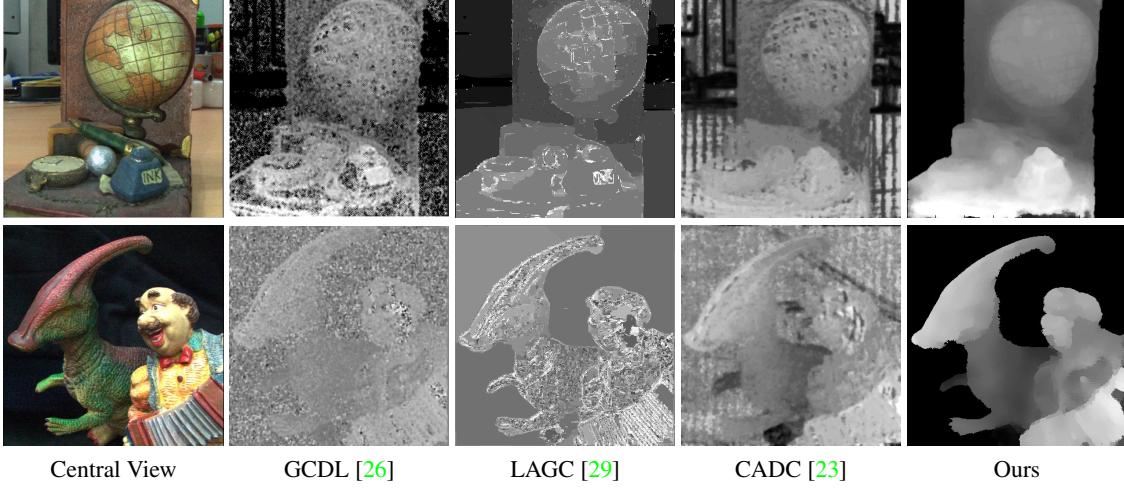


Figure 10. Qualitative comparison on the Lytro images.

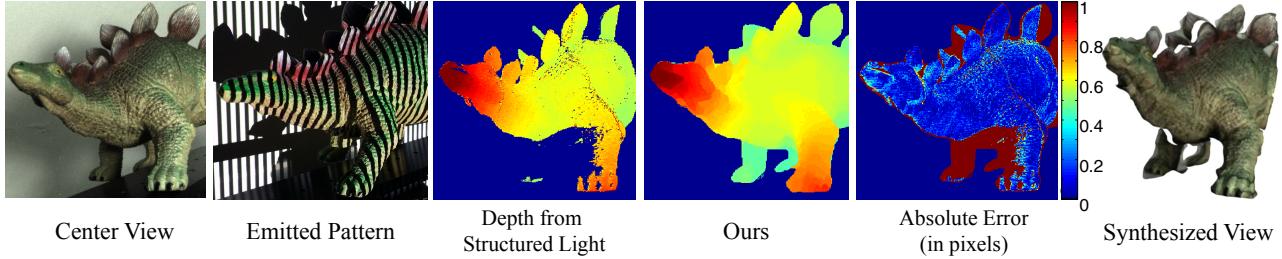


Figure 11. Evaluation of estimated depth map by using structured light based 3D scanner.

gested by Heber *et al.* [8], the *relative depth error*, which denotes the percentage of pixels whose disparity error is larger than 0.2%, was used. The author-provided implementation of GCDL was used, and parameter sweeps were conducted in order to achieve the best results. As the source codes of AWS and Robust PCA are not available, the error values reported in [8] are used.

Figure 7 presents a bar chart of the relative depth errors. The proposed method is compared with the other approaches, and it provided an accurate depth map for the seven datasets. Among the sub-pixel shift methods, the proposed phase-shift based approach exhibited the best results, which supports the importance of accurate sub-pixel shifting. Figure 8 presents a qualitative comparison of the proposed approach with GCDL. For the depth boundaries and homogeneous regions, the results of the proposed method do not have holes and they exhibit reliable accuracy.

The λ values in Eq. (10) were also altered in order to demonstrate the relative importance. The most significant term that influences the accuracy is the fourth term that adjusts the local smoothness. After the λ_3 is set to a reasonable value, λ_2 was altered in order to verify the confident matching correspondences (CMCs). Although the improvement was relatively small (from 9.32 to 8.91 in the Mona dataset), the third term assists in preserving the fine structures as depicted in Fig. 9.

5.2. Light-Field Camera Results

Lytro camera. Figure 10 presents a qualitative comparison of the proposed approach with GCDL, the line assisted graph cut (LAGC) [29], and the combinational approach of defocus and correspondence (CADC) [23] on two real-world Lytro images: a globe and a dinosaur. GCDL computes the elevation angles of lines in the epipolar plane images using structured tensors. Using challenging Lytro images, it may result in noisy depths even if the smoothness value is increased for the optimization. LAGC utilizes matched lines between the sub-aperture images. Its output is also noisy because low quality sub-aperture images affect accurate line matching.

Although Tao *et al.* [23] presented reasonable results through combining the defocus and correspondence cues, the correspondence was not robust to noisy and textureless regions, and it failed to clearly label the depth. The CADC exhibited reasonable results with the aid of the defocus and correspondence cues. These results also exhibited some holes and noisy disparities because its correspondence cue was not reliable in homogeneous regions. Because the proposed depth estimation method collects matching costs using robust clipping functions, it can tolerate significant outliers. In addition, the calculation of the exact sub-pixel shift using the phase shift theorem improves the matching quality as demonstrated in the synthetic experiments. The

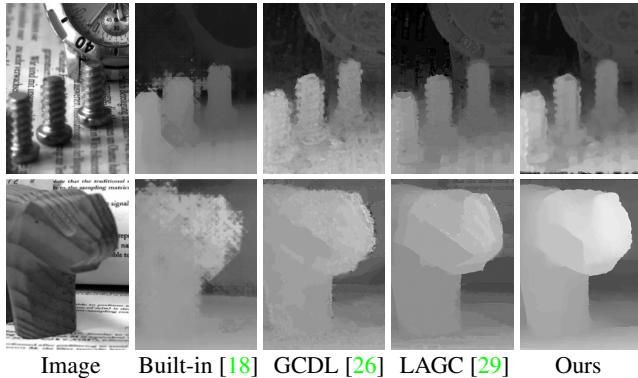


Figure 12. Comparisons of different disparity estimation techniques on Raytrix images.

Lytro software [1] also provided the depth map as presented in Fig. 1. However, the depth map quality was coarser than that obtained using the proposed method.

The proposed method was also verified using a structured light-based 3D scanner. A 16-bit gray code was emitted on the scene, and the disparity was computed through matching the observed code with the reference pattern. Figure 11 compares the two depths from the scanner and the proposed approach. Except for the occluded regions, the depth from the proposed approach exhibited less than 0.2 pixels of absolute error. Through using the geometric parameters acquired from [5], accurate depths are generated that can be used for the view synthesis.

Raytrix camera. A public Raytrix dataset [26] was used in this experiment. It has a 9×9 angular resolution and 992×628 pixels of spatial resolution. Its disparity between the adjacent sub-aperture images was less than 3 pixels, which is larger than that of Lytro. Then, $k = 0.05$ was set for a larger step size. Comparisons with GCDL, LAGC, and the built-in Raytrix algorithm [18] are presented in Fig. 12. The results of the built-in Raytrix were obtained from [25].

Because the built-in Raytrix algorithm only depends on a standard SAD cost volume, it fails to preserve the disparity edges. The GCDL exhibited a more reliable disparity using the Raytrix images than using the Lytro images. LAGC also exhibited well-preserved disparity edges, but it exhibited quantization errors as seen in Fig. 12. However, the proposed method exhibited a higher quality depth map that did not contain staircase artifacts (see Fig. 6).

Our simple lens camera. Inspired by [24], we constructed our own lab-made light field camera and tested our algorithm. A commercial mirrorless camera (Samsung™ NX1000) was modified through removing the cover glass on its CCD sensor and affixing a lenslet array.

Each lenslet in the array had a diameter of $52\mu\text{m}$, an inter-lenslet distance of $79\mu\text{m}$, and a focal length of $233\mu\text{m}$. In order to demonstrate its applicability for smaller devices,

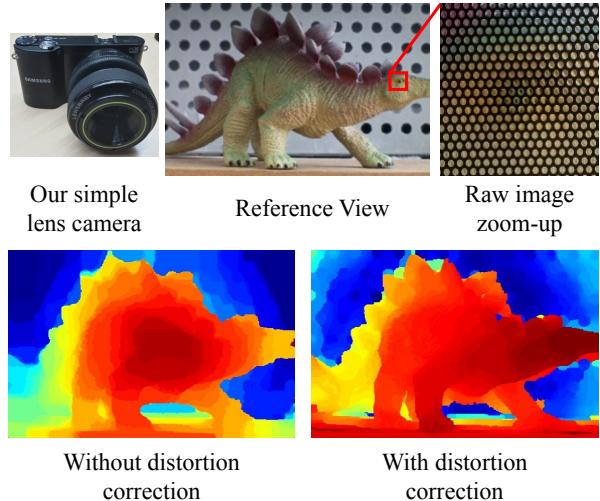


Figure 13. Result of our method with and without distortion correction. The input images for these result are captured by our simple lens light field camera.

the lab-made camera had only a single main lens. The focal length of the main lens was 50 mm and the F-number was 4. The camera was calibrated using an open geometric calibration toolbox [5]. The sub-aperture images had an 11×11 angular resolution and 392×256 pixels of spatial resolution. The disparity range was smaller than 1 pixel.

The lab-made lenslet light field camera suffered from severe distortion, which negatively affected the depth map quality as seen in Fig. 13. Using the proposed distortion correction step, the proposed depth map algorithm could locate accurate correspondences.

6. Conclusion

A novel method of sub-pixel-wise disparity estimation was proposed for light field images captured using several representative hand-held light field cameras. The significant challenges of estimating the disparity using very narrow baselines was discussed, and the proposed method was found to be effective in terms of utilizing the sub-pixel shift in the frequency domain. The adaptive aggregation of the gradient costs and confident matching correspondences further enhanced the depth map accuracy. The effectiveness of the proposed method was verified for various synthetic and real-world datasets. The proposed method outperformed three existing advanced methods.

Acknowledgement

This work was supported by the Study on Imaging Systems for the next generation cameras funded by the Samsung Electronics Co., Ltd. (DMC R&D center) (IO130806-00717), and by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (No.2010- 0028680).

References

- [1] The lytro camera. <http://www.lytro.com/>.
- [2] Lytro illum™ features. <http://blog.lytro.com/post/89103476855/lens-design-of-lytro-illum-turning-the-camera>.
- [3] Project webpage of this paper. https://sites.google.com/site/hgjeoncv/home/depthfromlf_cvpr15.
- [4] T. E. Bishop and P. Favaro. The light field camera: Extended depth of field, aliasing, and superresolution. *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)*, 34(5):972–986, 2012.
- [5] Y. Bok, H.-G. Jeon, and I. S. Kweon. Geometric calibration of micro-lens-based light-field cameras. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2014.
- [6] C. Chen, H. Lin, Z. Yu, S. B. Kang, and J. Yu. Light field stereo matching using bilateral statistics of surface cameras. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
- [7] T. Georgiev and A. Lumsdaine. Reducing plenoptic camera artifacts. *Computer Graphics Forum*, 29(6):1955–1968, 2010.
- [8] S. Heber and T. Pock. Shape from light field meets robust pca. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2014.
- [9] S. Heber, R. Ranftl, and T. Pock. Variational shape from light field. In *Proceedings of Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 66–79, 2013.
- [10] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. Gross. Scene reconstruction from high spatio-angular resolution light fields. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH)*, 32(4):73:1–73:12, 2013.
- [11] V. Kolmogorov and R. Zabih. Multi-camera scene reconstruction via graph cuts. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2002.
- [12] C.-K. Liang, T.-H. Lin, B.-Y. Wong, C. Liu, and H. H. Chen. Programmable aperture photography: multiplexed light field acquisition. *ACM Transactions on Graphics (TOG)*, 27(3):55, 2008.
- [13] C.-K. Liang and R. Ramamoorthi. A light transport framework for lenslet light field cameras. *ACM Transactions on Graphics (TOG)*, 34(16):16:1–16:19, 2015.
- [14] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal on Computer Vision (IJCV)*, 60(2):91–110, 2004.
- [15] Z. Ma, K. He, Y. Wei, J. Sun, and E. Wu. Constant time weighted median filtering for stereo matching and beyond. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2013.
- [16] R. Ng and P. Hanrahan. Digital correction of lens aberrations in light field photography. In *Proc. SPIE 6342, International Optical Design Conference*, 2006.
- [17] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan. Light field photography with a handheld plenoptic camera. *Computer Science Technical Report CSTR*, 2(11), 2005.
- [18] C. Perwass and L. Wietzke. Single lens 3d-camera with extended depth-of-field. In *IS&T/SPIE Electronic Imaging*, 2012.
- [19] Raytrix. 3d light field camera technology. <http://www.raytrix.de/>.
- [20] C. Rhemann, A. Hosni, M. Bleyer, C. Rother, and M. Gelautz. Fast cost-volume filtering for visual correspondence and beyond. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.
- [21] C. E. Shannon. Communication in the presence of noise. *Proceeding of the IEEE*, 86(2):447–457, 1998.
- [22] H. Tang and K. N. Kutulakos. What does an aberrated photo tell us about the lens and the scene? In *Proceedings of International Conference on Computational Photography (ICCP)*, 2013.
- [23] M. W. Tao, S. Hadap, J. Malik, and R. Ramamoorthi. Depth from combining defocus and correspondence using light-field cameras. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2013.
- [24] K. Venkataraman, D. Lelescu, J. Duparré, A. McMahon, G. Molina, P. Chatterjee, R. Mullis, and S. Nayar. Picam: An ultra-thin high performance monolithic camera array. *ACM Transactions on Graphics (TOG)*, 32(6):166:1–166:13, 2013.
- [25] S. Wanner and B. Goldluecke. Globally consistent depth labeling of 4D lightfields. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
- [26] S. Wanner and B. Goldluecke. Variational light field analysis for disparity estimation and super-resolution. *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)*, 2013.
- [27] S. Wanner, S. Meister, and B. Goldluecke. Datasets and benchmarks for densely sampled 4d light fields. In *In Proceedings of Vision, Modelling and Visualization (VMV)*, 2013.
- [28] Q. Yang, R. Yang, J. Davis, and D. Nistér. Spatial-depth super resolution for range images. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.
- [29] Z. Yu, X. Guo, H. Ling, A. Lumsdaine, and J. Yu. Line assisted light field triangulation and stereo matching. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2013.