# Reconstructing Dense Light Field from Multi-Focus Images Array for Virtual View Synthesis

Akira Kubota, *Member, IEEE,* Kiyoharu Aizawa, *Member, IEEE,* and Tsuhan Chen, *Member, IEEE*

***Abstract*— This paper presents a novel method for synthesizing a novel view from two sets of differently focused images taken by an aperture camera array for a scene of two approximately constant depths. The proposed method consists of two steps. The first step is a view interpolation to reconstruct an all in-focus dense light field of the scene. The second step is to synthesize a novel view by light field rendering technique from the reconstructed dense light field. The view interpolation in the first step can be achieved simply by linear filters that are designed to shift different object regions separately without estimating the depth map of the scene. The proposed method can effectively create a dense array of pin-hole cameras (i.e., all-focused images), so that the novel view can be synthesized with better quality.**

***Index Terms*— Image based rendering, view interpolation, spatial invariant filter, blur, light field rendering**

## I. Introduction

Most of view synthesis methods using multiple view images involve a problem of estimating a scene geometry [1]. Although a number of methods (e.g., stereo matching [2]) have been investigated for solving this problem; however it is still difficult to obtain the accurate geometry for real and arbitrary scenes. Using the inaccurate geometry for the view synthesis would induce visible artifacts on the synthesized novel image. As an alternative approach to such a geometry-based approach, image based rendering (IBR) [3], [4] has been studied for view synthesis in recent years. It does not need to estimate the scene geometry and enables us to synthesize photo-realistic novel images, independent of the scene complexity. The idea of IBR is to sample light rays flowing in the scene by capturing multiple images densely enough to create the novel views without aliasing artifacts through resampling of the sampled light rays [5]–[7]. The sampling density (i.e., camera spacing density) required for non-aliasing resampling is impractically high. To reduce the required sampling density, however, the geometric information is needed in some degree [8]. A new problem arises from this fact; there is a tradeoff on quality of the novel image between the required sampling density and geometric information. It is practically difficult to realize both approaches, accurately obtaining the scene geometry and densely capturing light rays, unless some specific equipments such as a laser scanner and plenoptic camera [9] are available; hence it is necessary to seek some desirable solution for effectively solving this problem.

In most of recently presented methods in IBR, for solving this problem, much efforts have been made on how to accurately obtain the geometric information to reduce the required sampling density. One approach is to find feature correspondence between the reference images. It is a traditional approach mainly used in structure from stereo or motion problems, but it is improved for the purpose of IBR recently in such a way that it detects and matches confident feature points that are required to provide sufficient quality of the synthesized view. Aliaga et al. [10] presented a robust feature detection and tracking method in which potential features are redundantly labeled in every images along with possible multiple paths and only confident features are used for matching. Siu et al. [11] proposed an image registration method between sparse set of three reference images, allowing feature matching in a large search area. They introduced an iterative matching scheme based on how much confident the extracted feature points are in terms of topological constraints that hold between triangles composed of three feature points in the reference images. Another approach is to estimate a view-dependent depth map at the novel viewpoint based on a color consistency of corresponding light rays or pixel values between the reference images, adopting a concept used in volumetric techniques [12]–[14] such as space-sweeping for a scene reconstruction. In this approach, the color consistency is checked at every or selected pixels with respect to different hypothetical depths in the scene and the depth value is estimated to be the depth that gives the highest consistency. The averaged value of the color values with the highest consistency is rendered at the pixel of the novel image. This approach is much suitable for real time processing [15]–[18].

In this paper, we present a novel approach to tackle this tradeoff problem in different way from the conventional ones. The concept of our approach is illustrated in Fig. 1. We deal with a scene consisting of foreground and background objects at approximately constant two depths and capture two sets of images with different focuses, as reference images, with an 1D array of real aperture cameras. Unlike the conventional methods using pin-hole cameras, the proposed method uses aperture cameras to capture differently focused images at each camera, one focused on the foreground and the other focused on the background. The proposed method consists of two steps. In the first step, we interpolate the intermediate images that are focused on the both objects at densely sampled positions among all the capturing cameras. In the second step,

Akira Kubota is with the Department of Information Processing, Tokyo Institute of Technology, 4259-G2-31 Nagatsuta, Midori-ku, Yokohama 226-8502, Japan (email: kubota@ip.titech.ac.jp)

Kiyoharu Aizawa is with the Department of Electrical and Electronic Engineering, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan (email: aizawa@hal.t.u-tokyo.ac.jp)

Tsuhan Chen is with the Department of Electrical and Computer Engineering, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213-3890, USA (email: tsuhan@cmu.edu)

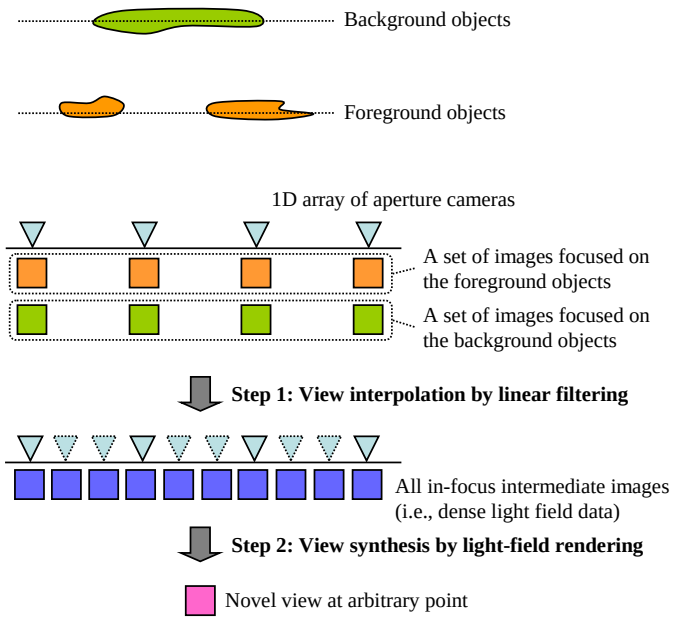



Fig. 1. Concept of our approach. Assuming a scene has two objects at different depths, we capture near and far-focused images at each camera position of 1D aperture camera array. In the first step, we interpolate all in-focus intermediate images densely between the cameras by linger filtering of the captured images. In the second step, we apply light field rendering to synthesizing novel views using the dense light filed data obtained in the first step.

we synthesize a novel view image by light field rendering (LFR) [6] using the intermediate views, i.e., dense light field data, obtained in the first step. View synthesize in the second step can be easily achieved with adequate quality, only if, in the first step, light field data sets are obtained correctly and densely enough for non-aliasing LFR.

This paper mainly addresses the problem of view interpolation between the capturing cameras in the first step and presents an efficient view interpolation method using linear and spatially invariant filters, avoiding problems of feature correspondence and depth map estimation. The reconstruction filters we present in this paper make possible shifting each object region according to the parallax required for the view interpolation without region segmentation. For the simple scene we assume here, the conventional vision-based methods such as stereo-matching [2] or depth from focus/defocus [19]–[21] can estimate the scene geometry, but they depend on the scene complexity and need much computations. In contrast, our approach needs only filtering and it is much simpler than such vision-based approaches, independent of the scene complexity as long as the scene consists of two approximately constant depths.

## II. Reconstructing Dense Light Field

### A. Problem description

In the first step of our method, we interpolate intermediate images densely between every camera pairs. The view interpolation problem we deal with here is illustrated in Fig. 2. Our goal is to generate an all in-focus intermediate images $f$ that would be captured at an arbitrary position (at a virtual camera in Fig. 2) between the adjacent cameras $C_i$ and $C_{i+1}$ (where $i$ is referred as to the camera index number) nearest to the view position. We use the four reference images captured with the cameras: the image $g_1^i$ at the camera $C_i$ focused on the foreground; the image $g_2^i$ at $C_i$ focused on the background; the image $g_1^{i+1}$ at $C_{i+1}$ focused on the foreground; and the image $g_2^{i+1}$ at $C_{i+1}$ focused on the background (see example in Fig. 6). The view position of the intermediate image is represented as an internally divided position between the cameras, parameterized by $\alpha$ for $0 \leq \alpha \leq 1$, and the distance between the cameras is $b$. We assume that the focal lengths of every cameras including the virtual camera are the same.

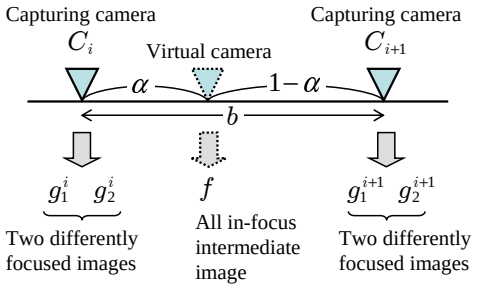



Fig. 2. View interpolation problem in the first step of our approach. We interpolate an all in-focus intermediate image, $f$, that would be seen from virtual camera position between the two nearest cameras. We use four reference images for this interpolation: the near-focused image at $C_i$, $g_1^i$; the far-focused image at $C_i$, $g_2^i$; the near-focused image at $C_{i+1}$, $g_1^{i+1}$; and the far-focused image at $C_{i+1}$, $g_2^{i+1}$.

### B. Imaging model

We model four reference images and the desired intermediate image by a linear combination of foreground and background textures. Consider the model of two differently focused images $g_1^i$ and $g_2^i$ at camera $C_i$. First, we introduce the foreground texture $f_1^i(u,v)$ and the background texture $f_2^i(u,v)$ that are visible from the camera $C_i$ and define them as the same as those in [22]:

$$f_1^i(u,v) \stackrel{\text{def.}}{=} \begin{cases} f(u,v), & d^i(u,v) = Z_1 \\ 0, & d^i(u,v) = Z_2 \end{cases} \quad (1)$$

$$f_2^i(u,v) \stackrel{\text{def.}}{=} \begin{cases} 0, & d^i(u,v) = Z_1 \\ f(u,v), & d^i(u,v) = Z_2 \end{cases} \quad (2)$$

where $f^i(u,v)$ is the ideal all in-focus image that is supposed to be captured with camera $C_i$. $d^i(u,v)$ is the depth map at the camera position, denoting the depth value corresponding to the pixel coordinate $(u,v)$. $Z_1$ and $Z_2 (> Z_1)$ are depths of the foreground and the background objects, respectively. Note that these textures and the depth map are unknown. The two differently focused images $g_1^i$ and $g_2^i$ can be modeled by the following linear combination of the defined textures:

$$\begin{cases} g_1^i(u,v) = & f_1^i(u,v) & + h(u,v) * f_2^i(u,v) \\ g_2^i(u,v) = h(u,v) * f_1^i(u,v) & + & f_2^i(u,v) \end{cases}, \quad (3)$$

where $h(u,v)$ is a point spread function (PSF) and $*$ denotes a 2D convolution operation. We assume that PSF can be

modeled as a Gaussian function

$$h(u,v) = \frac{1}{\pi R^2} \exp\left(-\frac{u^2+v^2}{R^2}\right), \quad (4)$$

where $R$ is an amount of blur, which is related to the corresponding standard deviation $\sigma$ of the Gaussian function [23]: $R = \sqrt{2}\sigma$.

In (3), the same PSF is used for both defocus regions because of the fact that the amount of blur on both regions become same after correction of image magnification due to the difference of the imaging plane position, if the imaging system is based on the thin-lens geometrical model [24]. The PSF does not depend on the camera position and can be commonly used for all the reference images, since we assume that depths of the two objects are constant with respect to every camera positions. In addition, the amount of blur $R$ is estimated in the pre-processing step using our previously presented method [24]; therefore the PSF is given.

The linear imaging model in (3) is not correct for the occluding boundaries as reported in [25]. However, it has an advantage that it enables us to use convolution operations, which can be represented by product operations in the frequency domain, resulting in the simpler imaging model (see (8)). Moreover, this model is adequate for obtaining satisfactory result, as shown in our experimental result. The limitation of using this model and its effect on the quality of the intermediate image will be tested on acquired real images in the session IV.

For modeling the all in-focus intermediate image, we have to consider two things: how to model the parallax and how to model occluded background texture. To model the parallax, we use a combination of textures that are appropriately shifted according to the intermediate position parameterized by $\alpha$. When the textures $f_1^i(u,v)$ and $f_2^i(u,v)$ are used, the model of the intermediate image, say $f'$, is modeled by

$$f'(u,v;\alpha) = f_1^i(u - \alpha d_1,\ v) + f_2^i(u - \alpha d_2,\ v), \quad (5)$$

where $d_1$ and $d_2$ are disparities of the foreground and the background objects between the adjacent cameras, respectively. These disparities can be estimated in a pre-processing step or known through camera calibration. The shift amounts we have to provide on the foreground and the background textures be $\alpha d_1$ and $\alpha d_2$, respectively.

Similarly, when the textures $f_1^{i+1}(u,v)$ and $f_2^{i+1}(u,v)$ are used, the intermediate image (say $f''$) at the same position is modeled by

$$\begin{aligned} f''(u,v;\alpha) = {} & f_1^{i+1}(u + (1-\alpha)d_1,\ v) \\ & + f_2^{i+1}(u + (1-\alpha)d_2,\ v), \end{aligned} \quad (6)$$

where $-(1-\alpha)d_1$ and $-(1-\alpha)d_2$ are the shift amounts to be provided on the foreground and the background textures, respectively.

To fill in occluded background in either one of two images $f'$ and $f''$ modeled by equations (5) and (6), we simply take a weighted average of them with weighting values $(1-\alpha)$ and $\alpha$ for $f'$ and $f''$, respectively, and finally model the desired intermediate image $f$ as

$$f(u,v;\ \alpha) = (1-\alpha)f'(u,v;\ \alpha) + \alpha f''(u,v;\ \alpha). \quad (7)$$

### C. View interpolation with linear filters in the frequency domain

Our goal is to generate the intermediate image $f$ in (7) from the four reference images, $g_1^i$, $g_2^i$, $g_1^{i+1}$, and $g_2^{i+1}$, modeled in (3). Note that PSF $h$, the disparities $d_1$ and $d_2$ are given, but the textures $f_1^i$, $f_2^i$, $f_1^{i+1}$, and $f_2^{i+1}$ are unknown. In this section, we derive the reconstruction filters that can generate the desired image $f$ directly from the reference images without region segmentation or depth map estimation.

First, consider the problem of generating $f'$ in (5) from $g_1^i$ and $g_2^i$ in (3). The same problem was already dealt with in our previous paper [22] in which an iterative reconstruction was presented in the spatial domain. In this paper, we present a much efficient method using filters in the frequency domain. We take the Fourier transform of equations (3) and (5) to obtain those imaging models in the frequency domain as follows:

$$\begin{cases} G_1^i(\xi,\eta) = & F_1^i(\xi,\eta) & + H(\xi,\eta)F_2^i(\xi,\eta) \\ G_2^i(\xi,\eta) = H(\xi,\eta)F_1^i(\xi,\eta) & + & F_2^i(\xi,\eta) \end{cases}, \quad (8)$$

and

$$F'(\xi,\eta;\alpha) = F_1^i(\xi,\eta)e^{-j2\pi\xi\alpha d_1} + F_2^i(\xi,\eta)e^{-j2\pi\xi\alpha d_2}, \quad (9)$$

where $\xi$ and $\eta$ denote the horizontal and vertical frequency. Capital letter function is used as the Fourier transform of the corresponding small letter function. These imaging models in the frequency domain are simpler than those in the spatial domain because the convolution and shifting operations are transformed into product operations. By eliminating $F_1^i$ and $F_2^i$ from equations (8) and (9), we then derive the following sum-of-products formula:

$$F'(\xi,\eta;\alpha) = K_1'(\xi,\eta;\alpha)G_1^i(\xi,\eta) + K_2'(\xi,\eta;\alpha)G_2^i(\xi,\eta) \quad (10)$$

where $K_1'$ and $K_2'$ can be considered as the frequency characteristics of linear filters that are applied to $G_1^i$ and $G_2^i$ in the frequency domain, respectively. The forms of $K_1'$ and $K_2'$ are given as follows:

$$\begin{cases} K_1'(\xi,\eta;\alpha) = \dfrac{e^{-j2\pi\xi\alpha d_1} - e^{-j2\pi\xi\alpha d_2}H}{1-H^2} \\ \\ K_2'(\xi,\eta;\alpha) = \dfrac{e^{-j2\pi\xi\alpha d_2} - e^{-j2\pi\xi\alpha d_1}H}{1-H^2} \end{cases}. \quad (11)$$

These filter values can not be determined in stable at $(\xi,\eta) = (0,0)$ (i.e., DC), since the denominator, $1-H^2$, equals 0 and either of the limit values of eq. (11) to the DC diverges. This divergence causes visual artifacts on the interpolated image $f'$ when there is noise or modeling error in low frequency components of $G_1^i$ and $G_2^i$. To avoid this problem, regularization is needed based on a constraint or prior information on the intermediate image. In this paper, we propose a frequency-dependent shifting method that is designed to have the shift amounts gradually decreased to zero at DC, as shown in Fig. 3. We do not interpolate the DC component of the intermediate image. This is reasonable from the fact that the low frequency components including DC do not cause much visual artifacts in the quality of the image, even if they are not shifted.

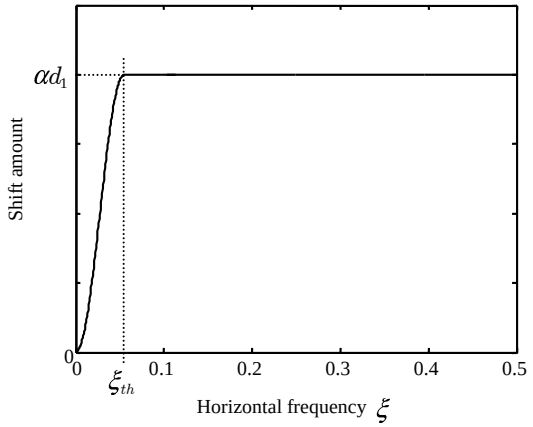



Fig. 3. Frequency-dependent shift amount of $\alpha d_1$. The shift amount is designed so that it gradually increases from zero at the DC to $\alpha d_1$ at preset threshold $\xi_{th}$ of frequency. For the frequency larger than $\xi_{th}$, it has a constant value of $\alpha d_1$.

Let the amounts of the frequency-dependent shifting on $f_1^i$ and $f_2^i$ for $f'$ be $\tau_1'$ and $\tau_2'$, respectively. Using a Cosine function, we design $\tau_1'$ as follows:

$$\tau_1'(\xi;\alpha) = \begin{cases} \alpha d_1 \left\{ \frac{1}{2} - \frac{1}{2}\cos\left(\frac{\pi\xi}{\xi_{th}}\right) \right\}, & \xi \le \xi_{th} \\ \alpha d_1, & \xi > \xi_{th} \end{cases} \quad (12)$$

where $\xi_{th}$ is a threshold frequency up to which the shift amount is changed depending on the frequency. We formulate $\tau_2$ similarly to the above equation using $d_2$ instead of $d_1$. By using these frequency-dependent shifting, we can make reconstruction filters stable and design them to be

$$\begin{cases} K_1'(\xi,\eta;\alpha) = \dfrac{e^{-j2\pi\xi\tau_1'} - e^{-j2\pi\xi\tau_2'}H}{1-H^2} \\ \\ K_2'(\xi,\eta;\alpha) = \dfrac{e^{-j2\pi\xi\tau_2'} - e^{-j2\pi\xi\tau_1'}H}{1-H^2} \end{cases}. \quad (13)$$

It is shown that both of the limit values of eq. (13) to the DC converge to 0.5.

Second, consider the problem of generating $F''$, which is the Fourier transform of (6), using $G_1''$ and $G_2''$. Similarly to the first case, we can derive the following formula for generating $F''$:

$$F''(\xi,\eta;\alpha) = K_1''(\xi,\eta;\alpha)G_1^{i+1}(\xi,\eta) + K_2''(\xi,\eta;\alpha)G_2^{i+1}(\xi,\eta) \quad (14)$$

with the stable reconstruction filters $K_1''$ and $K_2''$ that are designed to be

$$\begin{cases} K_1''(\xi,\eta;\alpha) = \dfrac{e^{-j2\pi\xi\tau_1''} - e^{-j2\pi\xi\tau_2''}H}{1-H^2} \\ \\ K_2''(\xi,\eta;\alpha) = \dfrac{e^{-j2\pi\xi\tau_2''} - e^{-j2\pi\xi\tau_1''}H}{1-H^2} \end{cases} \quad (15)$$

using the frequency-dependent shift amounts $\tau_1''$ and $\tau_2''$. In this case, $\tau_1''$ is formulated to be

$$\tau_1''(\xi;\alpha) = \begin{cases} -(1-\alpha)d_1 \left\{ \frac{1}{2} - \frac{1}{2}\cos\left(\frac{\pi\xi}{\xi_{th}}\right) \right\}, & \xi \le \xi_{th} \\ -(1-\alpha)d_1, & \xi > \xi_{th} \end{cases} \quad (16)$$

and $\tau_2''$ is formulated in the same manner. Note that both of the limit values of (15) to the DC also converge to 0.5.

Finally, by substituting resultant equations (10) and (14) using the stable reconstruction filters in (13) and (15) into the Fourier transform of (7), we derive the filtering method for generating the desired intermediate image $F$ in the frequency domain:

$$\begin{aligned} &F(\xi,\eta;\ \alpha) = \\ &\quad (1-\alpha)K_1'(\xi,\eta;\alpha)G_1^i(\xi,\eta) + (1-\alpha)K_2'(\xi,\eta;\alpha)G_2^i(\xi,\eta) \\ &\quad +\alpha K_1''(\xi,\eta;\alpha)G_1^{i+1}(\xi,\eta) + \alpha K_2''(\xi,\eta;\alpha)G_2^{i+1}(\xi,\eta). \end{aligned} \quad (17)$$

Taking the inverse Fourier transform of $F$, we can obtain the all in-focus intermediate image $f$. This suggests that interpolation of the desired intermediate image $f$ can be achieved simply by linear filtering of the reference images. The reconstruction filters we designed in equations (13) and (15) consist of PSF and shifting operators (i.e., exponential functions) that are spatially invariant; therefore the view interpolation without depth map estimation is possible. We can form the filters, because the amount of blur and disparities are known or estimated.

Figure 4 shows the block diagram of the proposed view interpolation method based on (17). Because of faster calculation, we use filtering in the frequency domain. First, we take Fourier transform (FT) of four reference images and multiply them by reconstruction filters in (13) and (15). We then sum up all the images after the multiplication and finally take inverse FT of the summed image. It should be noted that neither region segmentation nor depth map estimation is performed in the proposed method.

Using the proposed view interpolation method, we can generate an all in-focus intermediate image at any position between two adjacent cameras. Therefore, we can virtually create images that would be captured with densely arranged pin-hole cameras. The created set of images is considered as the dense light field, from which novel views from arbitrary positions can be rendered with sufficient quality by LFR. This will be tested in the next section.

## III. Experiment

### A. Experimental setup

Our experimental setup is shown in Fig. 5. We assume a $X$-$Z$ coordinate as a 2D world coordinate system. In our experiment, we captured 10 sets of two differently focused images (totally 20 reference images) at 10 different positions on the horizontal $X$ axis with F-number fixed at 2.4. When capturing the reference images, we used a x-stage to move a single digital camera (Nikon D1) horizontally with equal distance of $b$=8 [mm] from 0 to 72 [mm] in the horizontal position. Let us refer the camera at each position as $C_1$, $C_2$,..., and $C_{10}$. The test scene we set in this experiment

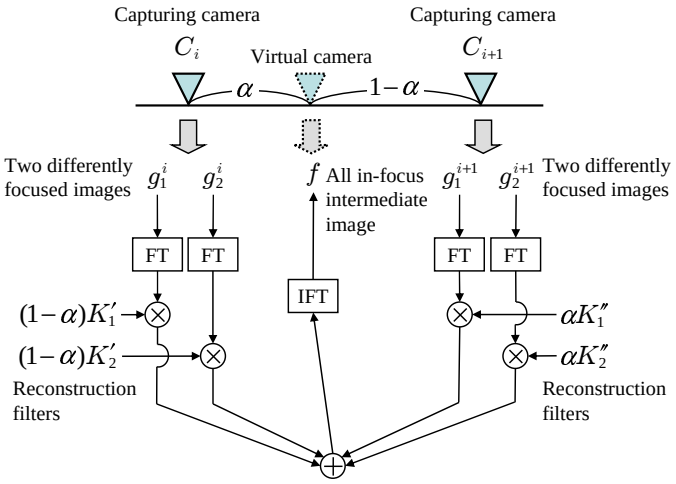



Fig. 4. Block diagram of the proposed view interpolation method using filters.

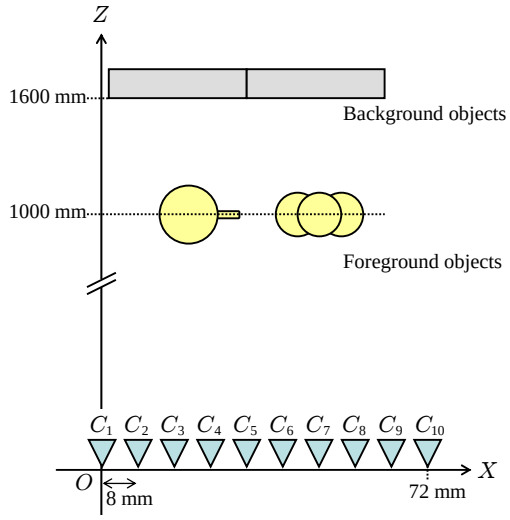



Fig. 5. Configurations of the scene and camera array used in the experiment.

consists of foreground objects (a cup with pencil and balls) and background objects (textbooks) that are located roughly at $Z_1$=1000 and $Z_2$=1600 [mm] of the $Z$ axis, respectively.

### *B. Preprocessing steps*

As preprocessing steps, we need three steps: (1) image registration; (2) estimation of blur parameter, $R$; and (3) estimation of disparities, $d_1$ and $d_2$. Image registration is needed to correct the difference in displacements and magnification between acquired two differently focused images at every camera positions. Magnification is caused by the difference of focus between the images. We used the image registration method presented in [24] that is based on a hierarchical matching technique. Parameters estimation in (2) and (3) can be performed via camera calibration using images captured for a test chart. In our experiment, we estimate these parameters from the captured reference images for the test scene. For the blur estimation using a test chart, see the reference [22].

TABLE I
ESTIMATED AMOUNT OF BLUR AND DISPARITIES

| Camera # | $R_1$ | $R_2$ | $d_1$ | $d_2$ |
|---|---|---|---|---|
| 1 | 4.1 | 3.5 | 18.6 | 11.5 |
| 2 | 4.0 | 3.6 | 18.2 | 11.0 |
| 3 | 4.0 | 3.4 | 18.9 | 11.8 |
| 4 | 3.7 | 3.7 | 18.4 | 11.2 |
| 5 | 3.8 | 3.5 | 18.5 | 11.5 |
| 6 | 3.6 | 3.8 | 18.2 | 11.1 |
| 7 | 3.4 | 3.6 | 18.9 | 11.9 |
| 8 | 3.3 | 3.6 | 18.4 | 11.3 |
| 9 | 3.3 | 3.6 | 18.7 | 11.9 |
| 10 | 3.3 | 3.6 | | |
| ave. | 3.65 | 3.58 | 18.5 | 11.4 |
| std. | 0.29 | 0.10 | 0.28 | 0.34 |

Unit: [pixels]

For the blur estimation based on the captured images, we also used our previously presented method [24] that estimates blur amounts on foreground and background regions independently, say $R_1$ and $R_2$, respectively. The basic idea used in the method is as follows. Blurred version of near-focused image $g_1^i$ by PSF $h$ with blur amount of $R$ is created and compared with far-focused image $g_2^i$ to evaluate the similarity (here we used absolute difference) between them at each pixel. After these evaluations for various amounts of $R$, the blur amount $R$ that gives the maximum similarity is estimated to be $R_1$ at the pixel. This pixel belongs to the region that is a part of the foreground region in principal. Finally, the estimated blur amounts are averaged in this region. The same idea is applied to estimating of $R_2$. Note that we do not need to identify each region in this blur estimation.

Estimated blur amounts at every camera positions and their statistics, averaged values (ave.) and standard deviations (std.), are shown in Table I. Both estimated values were not exactly the same for the most cases and the estimated values of $R_1$ have a variation (their std. of 0.29 [pixels]). This is because of depth variation on the surface of the foreground objects other than the estimation error. Nevertheless, we can use 3.6 [pixels] that is the averaged values of all the estimates of $R_1$ and $R_2$ as the representative blur amount for forming the PSF $h$. This approximation is possible because the proposed view interpolation method is robust to the blur estimation error, followed by experimental evaluation in the session IV-B.

Disparities estimation was performed by a template matching in sub-pixel accuracy. In this experiment, we specified a block region as a template by hand. When estimating the disparity of the foreground, $d_1$, we specified a region of face pattern drown at the center of the cup in the near-focused image $g_1^1$ at $C_1$ and find the horizontal position of the template that yields the best approximation of the corresponding region in every other near-focused images, $g_1^i$ for $i$=2,3,...,10. When estimating the disparity of the background, $d_2$, we specified the letter region of "SIGNAL" on the textbook and carried out the template matching with every other far-focused images. Estimated disparities $d_1$ and $d_2$ between adjacent cameras are

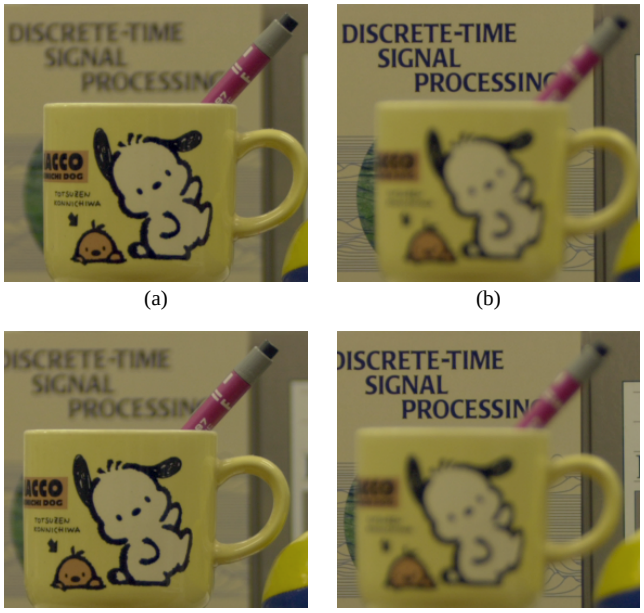



Fig. 6. Examples of captured real images: (a) A near-focused image at $X$=16 [mm] (b) A far-focused image at $X$=16 [mm] (c) A near-focused image at $X$=24 [mm] (d) A far-focused image at $X$=24 [mm]

shown in the last two columns in Table I. Since the maximum deviation is within 0.5 [pixels] for both estimated values for all the cases of camera pairs, we determined the final estimates $d_1$ and $d_2$ to be the averaged values, 18.5 and 11.4 [pixels], respectively. Note that region segmentation is not required in this disparity estimation.

### C. Reconstructed dense intermediate images

We constructed the reconstruction filters using the estimated parameters of blur amount and disparities in the pervious subsection. Based on the proposed filtering method in Fig. 4, we interpolated 9 all in-focus intermediate images between every adjacent cameras, totally 91 images including all in-focus images at the same position of all the cameras (this is the case of $\alpha$=0 or 1). Parameter $\alpha$ was set from 0 to 1 with equal increment of 0.1 for each interpolation. The threshold frequency $\xi_{th}$ was set at 0.01 for every experiments, which was empirically determined.

Example of captured images are shown in Fig. 6: (a) and (b) are respectively near-focused and far-focused images captured with camera $C_3$ at $X$ = 16 [mm]; (c) and (d) are those captured with camera $C_4$ at $X$ = 24 [mm]. These images are 24 bit color images of 280x256 [pixels]. We can observe different disparities (which were measured to be 18.9 and 11.8 [pixels] as shown in Table I) on the foreground and background objects regions. Although the distance between camera was set 8 [mm], this is not enough small for interpolating the intermediate image with adequate quality by LFR. This is because the difference of the disparities is about 8 [pixel] and is larger than 2 [pixel] that is the maximum difference ideally allowed for non-aliasing LFR [8].

Figure 7 shows all in-focus intermediate images interpolated by our proposed method from reference images in Fig. 6 for


Fig. 7. Intermediate images interpolated by the proposed method from four reference images in Fig. 6

the cases of $\alpha$=0, 0.2, 0.4, 0.6, 0.8, and 1, which correspond to the camera positions of 16, 17.6, 19.2, 20.8, 22.4, and 24 [mm], respectively. In every interpolated images, it can be seen that both foreground and background regions appear in-focus and are properly shifted. Expanded regions (64x64 [pixels] in size) including occluded boundaries in these images are shown in the right column of Fig. 7. In the occluded boundaries, although the imaging model we used is not correct, crucial artifacts are not visibly caused. Instead of artifacts, background textures (letters) that should have been unseen are partially visible because object regions are shifted and blended in our method. This transparency of the boundaries does not also produce much noticeable artifacts on the final result of novel view synthesis, seen in Fig. 9. Filling unseen background in precise is generally a difficult problem even for state-of-the-art vision-based approaches. Estimation error of the segmentation causes unnaturally distorted or broken boundaries, which would appear much visible in the novel image sequence created when the novel viewpoint was successively changed. Either of the presented and the conventional vision-based approaches needs a smoothing process on the boundaries to prevent visible artifacts. Blending texture in our approach acts as this effect.

Figure 8 (a) and (b) show epipolar plane images (EPIs) constructed from captured reference images $g_1^i$ and $g_2^i$, respectively, for $i$ = 1,2,...,10. Each horizontal line corresponds to a scan-line image (here $v$=184 is chosen) of each reference image and its vertical coordinate indicates the corresponding camera position $X$. These EPIs are very sparse and each EPI has only 10 scan lines according to the number of cameras. EPI of intermediate images interpolated by the proposed method is shown in Fig. 8 (c), where each horizontal line in the EPI corresponds to a scan-line image of each interpolated image. The interpolated EPI is much denser (91 horizontal scan lines) compared with those EPIs of the reference images, and it contains strait and sharper texture of lines. This means all in-focus intermediate images were generated accurately by the proposed method. The slope of line indicates the depth of objects. The stripe region with larger slop corresponds to the foreground regions and the stripe region with smaller slop, which are partially occluded with the foreground regions, corresponds to the background regions. Some conventional view interpolation methods exploiting interpolation of EPI [26]–[28] attempt to detect stripe regions with the same slop, whereas our method needs not such a region detection but simply spatial-invariant filtering, under our specific case of capturing two differently focused images at each camera position for the scene of two depths.

### D. Synthesizing a novel view by light field rendering

This section describes the second step of our approach, i.e., a novel view synthesis by LFR using the interpolated intermediate images obtained in the first step. A resultant set of densely interpolated intermediate images is considered as dense light field. The obtained light field is a collection of light rays specified by three parameters, camera position $X$ and pixel position $(u, v)$. Once the dense light field is obtained, a novel view image can be created by LFR. Since the novel view image is another set of light rays passing through the novel viewpoint and the pixel position, each novel right ray can be approximately created by properly resampling (interpolating) the nearest light rays in the obtained light field according to the novel viewpoint. To find the nearest light rays, we set a reference plane called *a focal plane* [29] between the foreground and the background objects in the scene. The optimal depth of the plane [8] is determined so that the novel image has the least aliasing artifacts. It is given by

$$Z_{opt} = 2\{1/(Z_1 - Z_v) + 1/(Z_2 - Z_v)\}^{-1}, \qquad (18)$$

where $Z_v$ is the position of the novel viewpoint in $Z$ axis. We find the intersection of the novel light ray with the focal plane and project it on every imaging planes of interpolated images to obtain corresponding pixel positions, that is, corresponding light rays, from which we can determine two nearest light rays. The intensity of the novel light ray is finally synthesized as a weighted average of the intensity values of the nearest light rays based on linear interpolation.

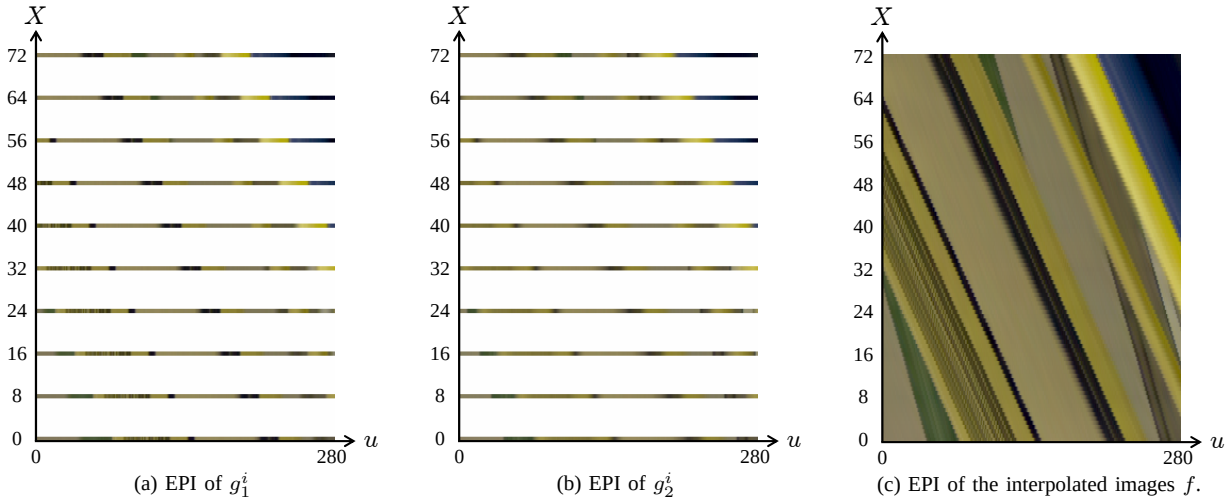
Figure 9 shows examples of synthesized novel view images from various viewpoints. The coordinate $(X, Z)$ written under each image is the novel viewpoint specified by the horizontal position $X$ and the depth $Z$. In order to clearly see perspective effects of the novel view images, we changed the viewpoint along with horizontal and depth axes independently. In Fig. 9 (a) and (b), zooming (close up) effect was demonstrated by changing only depth position of the viewpoint when horizontal position was fixed at 20 [mm]. We can see that the image in (b) is not a magnified version of the image in (a); the foreground objects (a cup and a pencil) was magnified larger than the background object in the image (b). Figure 9 (c) and (d) demonstrated parallax effect by changing only horizontal position of the viewpoint. Different shifting (displacement) effects were provided on the objects and unseen background texture in one image was visible in the other image.

## IV. Discussion

We have shown that we can effectively interpolate all in-focus intermediate images between neighboring two cameras simply by filtering the reference images captured with these cameras with different focuses. In this section, we experimentally test the performance of our view interpolation method in terms of its accuracy and robustness against estimation error of blur amount.

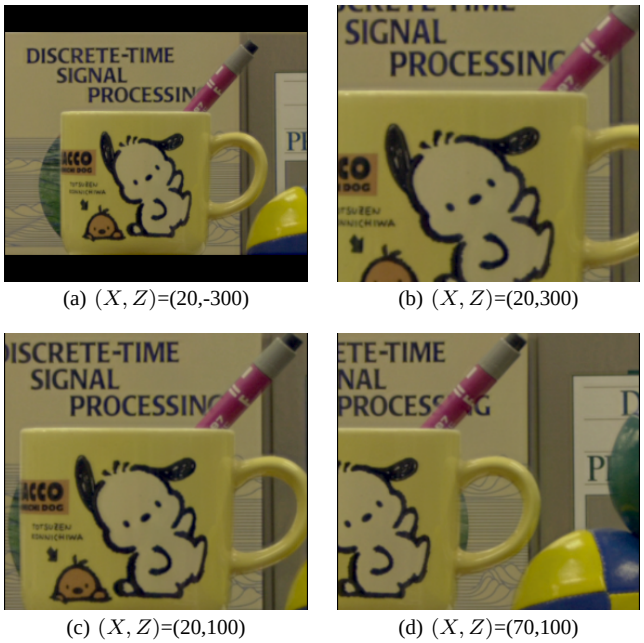
### A. Performance evaluation

We tested accuracy of the proposed method for the case of interpolating an image at the center of two camera positions. At each of two camera positions, near and far focused images were captured with F-number of 2.4 for the same scene used in the experiments. For measuring the accuracy of the interpolated image, the ground truth image is required. In this test, we captured an image at the center of camera positions with small aperture (F-number: 13) as the ground truth image so that the image is focused on both regions at different depths.




(a) EPI of $g_1^i$ (b) EPI of $g_2^i$ (c) EPI of the interpolated images $f$.

Fig. 8. Example of epipolar plane images (EPIs) of the reference images and the densely interpolated images by the presented method. Here write regions in (a) and (b) indicate regions with zero value. The presented method can generate dense EPI in (c) from sparse EPIs in (a) and (b) by linear filtering of them.



(a) $(X, Z)$=(20,-300) (b) $(X, Z)$=(20,300)

(c) $(X, Z)$=(20,100) (d) $(X, Z)$=(70,100)

Fig. 9. Examples of the novel views synthesized by LFR from the densely interpolated images. $(X, Z)$ is the coordinate of the novel viewpoint in horizontal and depth axes. In (a) and (b), zooming effect is demonstrated by changing only depth position of the viewpoint. In (c) and (d), parallax effect is demonstrated by changing only horizontal position of the viewpoint.

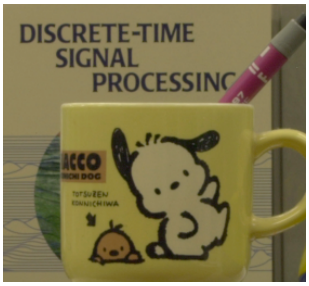
Figure 10 shows the ground truth image and interpolated images by the presented method with $\alpha$=0.5 for different distances between cameras of 4, 8, 12, and 16 [mm]. For comparison, we also generated the center image by LFR based on the focal plane at the optimal depth of 1230 [mm] calculated by (18). The reference images used for LFR were all in-focus images generated at two camera positions by the presented method with $\alpha$=0 and 1. The comparison between results in Fig. 10 shows an advantage of our method over LFR. In the images of our method, both object regions are properly in-focus and shifted without noticeable artifacts. In contrast, both regions in the images interpolated by LFR suffer from blur or ghosting artifacts. These artifacts are caused by incorrectness of pixel correspondences due to large distance between cameras. Our method can prevent pixel miscorrespondences by properly shifting each object region. In addition, this shifting operation can be achieved by spatially invariant filtering of the reference images, not requiring region segmentation.

Mean square errors (MSEs) in green color channel were computed as a quality measure between the interpolated images and the ground truth image as shown in Fig. 11. MSEs in LFR are larger than those in the presented method. The quality of images interpolated by our method is sufficiently good for every cases, whereas that of LFR is much degraded with an increase of the distance between cameras.

In the interpolated images when the distance was 16 [mm] (bottom-left in the Fig. 10), occluded boundaries (a pencil in the foreground and letters in the background) look transparent due to shifting and blending of different textures by our method. Although the presented method can not prevent these effect of transparency, blending shifted textures in our approach has an effect of canceling out errors in $f'(u, v; 0.5)$ and $f''(u, v; 0.5)$, which are the intermediate images modeled in equations (5) and (6) before blending. These images are shown in Fig. 12. There are noticeable artifacts in color value, because the reconstruction filters for generating these images, $K_1'$, $K_2'$, $K_1''$, and $K_2''$, have much larger values at lower frequency and amplify error in the frequency. However, the finally interpolated image in the bottom-left in Fig. 10 that is a blended (average) of them has less visible error, showing the effect of canceling errors.

### B. Effect of blur estimation error

In this section, we examined robustness of the presented view interpolation method against estimation error of blur amount for the same scene. Setting various amounts of blur from 1.0 to 6.0 [pixels], we interpolated intermediate images at the center of the cameras in 8 [mm] apart and measured MSEs between interpolated images and the ground truth image. The




(a) The ground truth image that is the image actually captured at the center between cameras with a small aperture (F-number: 13).


(b) Interpolated images by the proposed method (left) and LFR method (right) at the center of two cameras for various distances between cameras of 4, 8, 12, and 16 [mm] (from top to bottom).

Fig. 10. Comparison between the proposed method and LFR method.

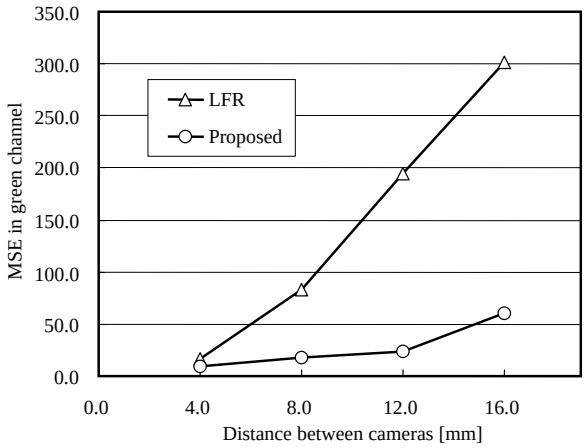



Fig. 11. Performance evaluations of the proposed method and the conventional light field rendering (LFR) for the middle view interpolation. Mean square errors (MSE) were evaluated between interpolated view images and actually captured all in-focused image (the ground truth image) at the middle position.


(a) (b)

Fig. 12. Generated $f'(u, v; 0.5)$ and $f''(u, v; 0.5)$ that are averaged to be the final interpolated image shown in bottom-left in Fig. 10.

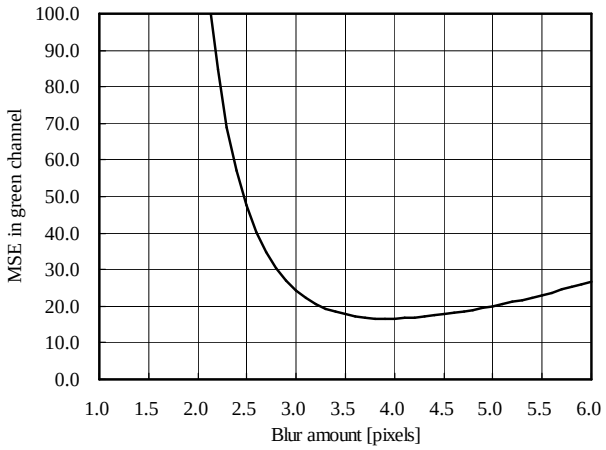



Fig. 13. Error evaluation of the proposed method for the middle view interpolation under various blur amounts. Mean square errors (MSE) were evaluated between interpolated view image and the ground truth image at the middle position.

measured MSEs in green channel are shown in Fig. 13. This result shows that MSE is smaller than 30 in the wide range of blur amounts from 2.9 to over 6.0 [pixels]; therefore our method is robust to blur estimation error. The result also shows that our method allows the scene to have depth variation in some extent as long as the amount of blur caused on the region is within that range.

## V. CONCLUSION

This paper has presented a novel two-steps approach for IBR. Unlike the conventional IBR methods, the presented method uses aperture cameras to capture the reference images with different focuses at different camera positions for a simple scene consisting of two depth layers. The first step is a view interpolation for densely generating all in-focus intermediate images among camera positions. The obtained set of intermediate images can be used as the dense light field data for quality view synthesis via LFR in the second step. This paper showed that the view interpolation can be achieved effectively by spatially invariant filtering of the reference images, not requiring estimation of the geometric information. The presented view interpolation method works well even for the case of camera spacing sparser than that required for non-aliasing LFR.

## REFERENCES


[1] M. Oliveira, "Image-Based Modeling and Rendering Techniques: A Survey," in *RITA - Revista de Informatica Teorica e Aplicada*, Volume IX, pp. 37–66, 2002.

[2] U. R. Dhond and J. K. Aggarwal "Structure from stereo: a review," *IEEE Trans. on System, Man, and Cybernetics*, vol. 19, no. 6, pp. 1489-1510, 1989.

[3] C. Zhang and T. Chen, "A survey on image-based rendering - representation, sampling and compression," *EURASIP Signal Processing: Image Communication*, vol. 19, pp. 1–28, 2004.

[4] H-Y. Shum, S. B. He, and S-C. Chan, "Survey of Image-Based Representations and Compression Techniques", *IEEE Trans. on Circuit and System for Video Technology*, vol. 13, no. 11, pp. 1020 – 1037, 2003.

[5] E. H. Adelson, J. R. Bergen, "The Plenoptic Function and the Elements of Early Vision," in in *M. S. Landy & J. A. Movshon (Eds.), Computational Models of Visual Processing (pp. 3–20). Cambridge, Massachusetts, MIT Press*, 1991.

[6] M. Levoy and P. Hanrahan, "Light field rendering", in *proc. of ACM SIGGRAPH'96*, pp. 31–42, 1996.

[7] H.-Y. Shum and L.-W. He, "Rendering with concentric mosaics," in *proc. of ACM SIGGRAPH'99*, pp. 299–306, 1999.

[8] J-X. Chai, X. Tong, S.-C. Chany, H.-Y. Shum, "Plenoptic Sampling," in *proc. ACM SIGGRAPH'00*, pp. 307–318, 2000.

[9] R. Ng, M. Levoy, M. Bredif, G. Duval, M. Horowitz, and P. Hanrahan, "Light Field Photography with a Hand-Held Plenoptic Camera," in *Technical Report CSTR 2005-02, Computer Science Department, Stanford University*, 2005

[10] D.G. Aliaga, T. Funkhouser, D. Yanovsky, I. Carlbom, "Sea of images," *IEEE Computer Graphics and Applications*, vol. 23, no.6, pp. 22 – 30, 2003.

[11] A. M. K. Sju, E. W. H. Lau, "Image Registration for Image-Based Rendering," *IEEE Transactions on Image Processing*, vol. 14, no. 1, pp. 241–252, 2005.

[12] R. T. Collins, "Space-Sweep Approach to True Multi-Image Matching," in *proc. of CVPR'96*, pp. 358–363, 1996.

[13] S. M. Seitz and C. R. Dyer, "Photorealistic Scene Reconstruction by Voxel Coloring," in *proc. of CVPR'97*, pp. 1067–1073, 1997.

[14] K. Kutulakos and S. M. Seitz, "A theory of shape by space carving," *Int. Journal of Computer Vision*, vol. 38, no. 3, pp. 199–218, July 2000.

[15] R. Yang, G. Welch, G. Bishop, "Real-time consensus-based scene reconstruction using commodity graphics hardware," in *Proc. of Pacific Conference on Computer Graphics and Applications'02*, pp. 225–235, 2002.

[16] I. Geys, T. P. Koninckx, and L. V. Gool, "Fast interpolated cameras by combining a GPU based plane sweep with a max-flow regularisation algorithm," in *Proc. of Intl Symp. 3D Data Processing, Visualization and Transmission*, pp. 534–541, 2004.

[17] K. Takahashi, A. Kubota, T. Naemura, "A Focus Measure for Light Field Rendering," in *proc. of ICIP'04*, pp. 2475–2478, 2004.

[18] C. Zhang and T. Chen, "A Self-Reconfigurable Camera Array," in *proc. of Eurographics Symposium on Rendering'04*, pp. 243–254, 2004

[19] J. Ens and P. Lawrence, "An investigation of methods for determining depth from focus," *IEEE Trans. on PAMI* vol. 15, no. 5, pp. 97–107, 1993.

[20] S. K. Nayar and Y. Nakagawa, "Shape from focus: An effective approach for rough surfaces," *IEEE trans. PAMI*, vol. 16, no. 8, pp. 824–831, 1994.

[21] N. Rajagopalan, S. Chaudhuri, and U. Mudenagudi, "Depth Estimation and Image Restoration Using Defocused Stereo Pairs," *IEEE trans. on PAMI*, vol. 26, no. 11, pp. 1521–1525, 2004.

[22] K. Aizawa, K. Kodama and A. Kubota, "Producing Object Based Special Effects by Fusing Multiple Differently Focused Images", *IEEE trans. on Circuits and System for Video Techinology*, vol. 10, no. 2, pp. 323–330, 2000.

[23] M. Subbarao, T-C. Wei, and G. Surya, "Focused image recovery from two defocused images recorded with different camera settings," *IEEE trans. on Image Processing*, vol. 4, no. 12, pp. 1613–1627, 1995

[24] A. Kubota and K. Aizawa, "Reconstructing arbitrarily focused images from two differently focused images using linear filters," *IEEE trans. on Image Processing* (to appear on No. 10)

[25] N. Asada, H. Fujiwara, and T. Matsuyama, "Seeing behind the scene: analysis of photometric properties of occluding edges by the reversed projection blurring model", *IEEE tanns. on PAMI*, vol. 20, no. 2, pp. 155–167, 1998.

[26] R. Hsu, K. Kodama, and K. Harashima, "View interpolation using epipolar plane images," in *Proc. IEEE ICIP'94*, pp. 745–749, 1994.

[27] A. Katayama, K. Tanaka, T. Oshino, and H. Tamura, "A viewpoint dependent stereoscopic display using interpolation of multi-viewpoint images," in *Proc. SPIE Int. Conf. on Stereoscopic Displays and Virtual Reality Systems II*, vol. 2409, pp. 11–20, 1995.

[28] M. Droese, T. Fujii and M. Tanimoto, "Ray-Space Interpolation constraining Smooth Disparities based on Loopy Belief Propagation," in *Proc. Int. Workshop on Systems, Signals and Image Processing*, pp. 247–250, 2004.

[29] A. Isaksen, M. Leonard, S. J. Gortler "Dynamically Reparameterized Light Fields," *MIT-LCS-TR-778*, 1999.