Direct Filtering Method for Image based rendering

Akira Kubota
Dept. of Information Processing
Tokyo Institute of Technology
4259-G2-31 Nagatsuta, Midori-ku
Yokohama 226-8502, Japan
kubota@ip.titech.ac.jp

Kiyoharu Aizawa
Dept. of E.E., University of Tokyo
7-3-1 Hongo, Bunkyo-ku
Tokyo 113-8656, Japan
aizawa@hal.t.u-tokyo.ac.jp

Tsuhan Chen
Dept. of ECE, CMU
5000 Forbes Avenue
Pittsburgh, PA 15213-3890, USA
tsuhan@cmu.edu

***Abstract*— Image based rendering (IBR) is basically a light ray resampling method for generating a novel image without aliasing artifacts from a given set of sampled light rays. To avoid aliasing artifacts, conventional methods require an estimate of the scene geometry. In this paper, we present a novel IBR method by linear filtering without estimating the scene geometry, for the simplified case when generating a virtual image at the center of 2×2 sparse camera array for a two depth-layers scene. The reconstruction filter used in the proposed method can derived by integrating all the process in our previously proposed method into an one-shot process.**

I. Introduction

How to visualize a scene from an arbitrary viewpoint, given a set of sampled light rays from different viewpoints? Image based rendering (IBR) [1], [2] tackles this problem by resampling the sampled light rays without introducing aliasing artifacts. In other words, IBR tries to solve a high-dimensional sampling problem; therefore the sampling theorem gives an answer to how many samples are needed. For the 4-D case where light rays are sampled with cameras spaced on a plane, the plenoptic sampling theory [3] presents a minimum sampling rate for non-aliased resampling. If the sampling rate is larger than the minimum sampling rate, then novel images placed at arbitrary position on the camera plane (i.e., dense light rays) can be reconstructed by a 4-D low-pass filtering of the sampled light rays without using any depth information. Once sufficient light rays are densely reconstructed, novel images from arbitrary positions can be easily generated from them. More importantly, the plenoptic sampling theory shows that there is a trade-off between the sampling rate and the depth resolution available that are required for non-aliased resampling. When the sampling rate is inadequate (we call this case the light rays are under-sampled), the scene geometry is needed to avoid aliasing artifacts.

In this paper, we address the problem of reconstructing a novel view from an under-sampled set of light rays without estimating the scene geometry. We consider the following simplified case (Fig. 1); four images captured with a 2×2 planar camera array, and a scene consisting of only two layers at different depths. Our goal is to generate an image of a novel view at the center of the camera array by directly filtering the captured images. The reconstruction filter is derived through integration of all steps of our previously presented method [4] into a one-shot process. The previous method is composed of two steps. In the first step, multiple images for a novel view point at the center of the camera array are generated by light field rendering (LFR) [5] based on different multiple depths. Each of the resulting images suffers from aliasing artifacts due to the limited camera spacing density. In the second step, the desired image without aliasing artifacts is reconstructed by fusing the multiple images. We show that both processing steps can be replaced by linear filtering, independent on the scene geometry. This suggests that a direct relationship can be derived between the captured images and the desired resulting image. A reconstruction filter can be designed from the obtained relationship. By using the reconstruction filter, we can directly generate the desired image from captured images. The presented approach is much efficient compared with the previous method.

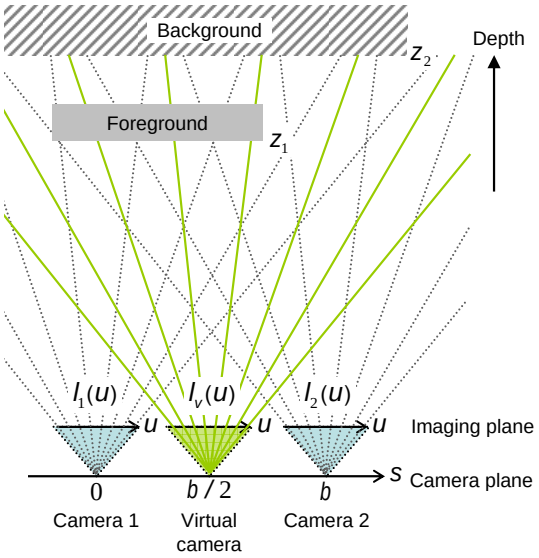
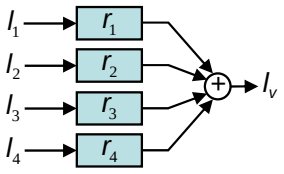
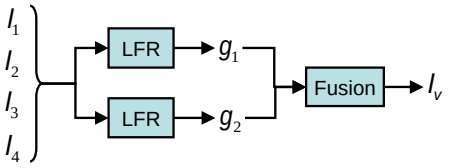


Fig. 1. This figure shows the reconstruction problem dealt with in this paper, illustrated on a 2-D subspace of the original 4-D light field. Given the set of light rays $l_1(u)$ and $l_2(u)$ (dotted gray lines), originating from camera 1 and camera 2, respectively, for the scene consisting of two layers (foreground at depth z_1 and background at z_2), we wish to reconstruct the set of light rays $l_v(u)$ (solid green lines), originating from a virtual camera centered between camera 1 and 2, without estimating the scene geometry.



(a) The proposed approach: Direct filtering method.



(b) Our previous approach: the multi-layered rendering and fusing method [4]

Fig. 2. A comparison between the proposed filtering approach and the multi-layered rendering and fusion method [4]. In this paper, we show that similar results as the multi-layered rendering and fusion approach can be obtained by directly filtering the captured images.

II. Direct filtering method

A. Problem formulation

An example of the reconstruction problem of light rays discussed in this paper is shown in Fig. 1. Assuming a scene containing only two depths, given four images captured with a 2×2 camera array, we generate a novel image positioned at center of the camera array. In this section, we formulate the problem and objective using the light field representation in [6]. We denote the four images by $l(u, v, 0, 0)$, $l(u, v, b, 0)$, $l(u, v, 0, b)$, and $l(u, v, b, b)$, where (u, v) is the pixel position on each imaging plane. The latter coordinate, $(0, 0)$, $(b, 0)$, $(0, b)$ and (b, b), denotes the camera positions on the camera plane. We use $l_i(u, v)$, $(i = 1, 2, 3, 4)$, as shorthand notation for the captured images and $\{l_i\}$ for the entire set of captured images. The desired image $l(u, v, b/2, b/2)$ is denoted as $l_v(u, v)$. In this paper, we assume that there are only two depths in the scene, denoted by z_1 and z_2, but with unknown depth map or the scene geometry. Our objective is to generate the virtual image l_v directly from a the set of images l_i without estimating the depth map. This is illustrated in figure 1, where the solid green rays are generated from the given dotted gray rays.

In this paper, we propose a direct filtering method based on the following reconstruction model:

$$l_v(u, v) = \sum_{i=1}^{4} \{r_i(u, v) * l_i(u, v)\}, \tag{1}$$

where r_i is a reconstruction filter which is applied to the captured images l_i. The operation "$*$" indicates a 2-D convolution. The proposed reconstruction model is illustrated in Fig. 2 (a). The reconstruction filter r_i is derived by integrating all the processing steps in the multi-layered rendering and fusing method (Fig. 2 (b)) our previously proposed method [4].

B. Image formation model in the multi-layered rendering and fusing method

As shown in Fig. 2 (b), the multi-layered rendering and fusing method [4] consists of two steps. In the first step, two virtual images, g_1 and g_2, are generated at the center point of the camera array by using the conventional LFR method based on two depths, z_1 and z_2, respectively. In the second step, final image l_v is the fusion of the two computed images g_1 and g_2 obtained in the step one.

The LFR process for generating g_j $(j = 1, 2)$ in the first step can be modeled as an average of shifted versions of the captured images based on assumed depth z_j. Therefore, the generated images g_j can be modeled as

$$\begin{aligned} g_j(u, v) = {} & (1/4) \cdot \delta(u - d_j, v - d_j) * l_1(u, v) \\ & + (1/4) \cdot \delta(u + d_j, v - d_j) * l_2(u, v) \\ & + (1/4) \cdot \delta(u - d_j, v + d_j) * l_3(u, v) \\ & + (1/4) \cdot \delta(u + d_j, v + d_j) * l_4(u, v), \end{aligned} \tag{2}$$

where d_j is the disparity of layer at depth z_j between the virtual image l_v and any of captured images: $d_j = bf/(2z_j)$. In the second step, l_v, g_1 and g_2 are modeled by linear combinations as follows:

$$l_v(u, v) = \phi_1(u, v) + \phi_2(u, v) \tag{3}$$

$$\begin{cases} g_1(u, v) = \phi_1(u, v) + h(u, v) * \phi_2(u, v) \\ g_2(u, v) = h(u, v) * \phi_1(u, v) + \phi_2(u, v) \end{cases} \tag{4}$$

where ϕ_1 and ϕ_2 is respectively foreground and background texture visible from the virtual point. Both ϕ_1 and ϕ_2 are unknown since the depth map is not also unknown. h represents a filtering model of ghosting artifacts. From geometrical relationship between the virtual camera and the 2x2 camera array, it can be shown that the same artifact is caused on both textures and the filter $h(u, v)$ is represented by the following linear spatially invariant filter:

$$\begin{aligned} h(u, v) = {} & (1/4) \cdot \delta(u - \Delta, v - \Delta) + (1/4) \cdot \delta(u - \Delta, v + \Delta) \\ & + (1/4) \cdot \delta(u + \Delta, v - \Delta) + (1/4) \cdot \delta(u + \Delta, v + \Delta), \end{aligned} \tag{5}$$

where $\delta(\cdot, \cdot)$ is the 2-D Dirac delta function and $\Delta = (bf/2)(1/z_1 - 1/z_2)$. f is the distance between the imaging plane and the camera plane, which is equal to the focal length of cameras.

C. Derivation of the reconstruction filter

We derive a relationship between the captured images $\{l_i\}$ and the desired image l_v by eliminating the four intermediate functions g_1, g_2, ϕ_1 and ϕ_2 from (2), (3), and (4).

First, we eliminate ϕ_1 and ϕ_2 from (3) and (4) by solving (4) for ϕ_1 and ϕ_2, and by substituting them into (3). We use the iterative Gauss-Seidel method to solve (4) in the spatial domain. The iteration rules from $\tau - 1$ to τ are:

$$\begin{cases} \phi_1^{(\tau)} = g_1 - h * \phi_2^{(\tau-1)} \\ \phi_2^{(\tau)} = g_2 - h * \phi_1^{(\tau)} \end{cases}, \tag{6}$$

which can be written as

$$\begin{cases} \phi_1^{(\tau)} = g_1 - h * g_2 + h^{\langle 2 \rangle} * \phi_1^{(\tau-1)} \\ \phi_2^{(\tau)} = g_2 - h * g_1 + h^{\langle 2 \rangle} * \phi_2^{(\tau-1)} \end{cases} \quad (7)$$

where $h^{\langle n \rangle} = h * h^{\langle n-1 \rangle}$. At the τ-th iteration, the solution of l_v is given by their sum:

$$\begin{aligned} l_v^{(\tau)} &= \phi_1^{(\tau)} + \phi_2^{(\tau)} \\ &= (g_1 + g_2) - h * (g_1 + g_2) + h^{\langle 2 \rangle} * l_v^{(\tau-1)} \end{aligned} \quad (8)$$

By setting $l_v^{(0)} = (g_1 + g_2)/2$ and expanding equation (8), we can derive the following simple relation between $l_v^{(\tau)}$ and $(g_1 + g_2)$:

$$l_v^{(\tau)} = k^{(\tau)} * (g_1 + g_2). \quad (9)$$

where $k^{(\tau)}$ is a linear filter defined by:

$$k^{(\tau)} = (\delta + h^{\langle 2 \rangle} + \cdots + h^{\langle 2\tau-2 \rangle}) * (\delta - h) + h^{\langle 2\tau \rangle}/2. \quad (10)$$

The certainty of the solution is increased by not estimating ϕ_1 and ϕ_2 directly For instance, this avoids estimating DC component of ϕ_1 and ϕ_2, which cannot be identified uniquely. A detailed analysis on this effect is our future work.

Secondary, substituting (2) into (9) yields the relationship between l_i and l_v, from which the reconstruction filter r_i is designed. We only show r_1, the other filters are similar:

$$\begin{aligned} r_1(u,v) &= k^{(\tau)} * \frac{1}{4}\big\{\delta(u-d_1, v-d_1) + \delta(u-d_2, v-d_2)\big\} \\ &= \frac{1}{4}\big\{k^{(\tau)}(u-d_1, v-d_1) + k^{(\tau)}(u-d_2, v-d_2)\big\} \end{aligned} \quad (11)$$

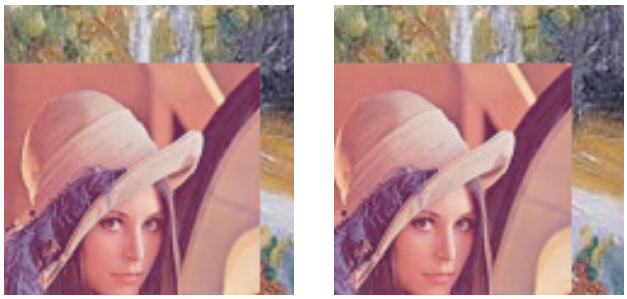
The resulting filters are spatial invariant; hence, the reconstruction of a virtual image is possible, using (1) and filters r_i, directly from the captured images without local processing (e.g., depth map estimation). The proposed filtering method does not either need the multiple LFR processes nor the estimation of texture.

III. Simulation results

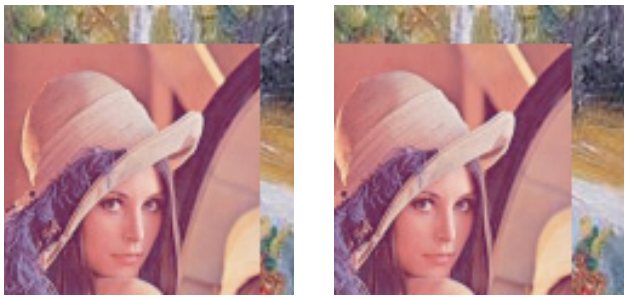
The performance of our reconstruction method is tested by using four synthetic images l_i of a scene containing only two layers. The image of Lena is used as the foreground texture, and a painting is used as background texture (Fig. 3). In Table I, the camera parameters, disparities d_1 and d_2, Δ, and the image resolution are shown for the test scene.

The simulation results are shown in Fig. 4 with the ground truth image (Fig. 4 (a)) at the center of the camera array. In Fig. 4 (b), an image rendered using LFR based on the optimal depth [5] is shown. The optimal depth is selected such that the aliasing artifacts are minimized for a given depth range. In our example the optimal depth is given by $2(1/z_1+1/z_2)^{-1}$ = 666. The resulting image by LFR is blurry and contains ghosting artifacts, because the sampling density in the camera plane is not sufficient (i.e., distance between cameras b is too large for the depth range.). This shows that a depth map is required for generating a non-aliased image.

Two images generated using the proposed method are shown in Fig. 4 (c) and (e), with τ set to 1 and 5, respectively.



(a) l_1 (b) l_2



(c) l_3 (d) l_4

Fig. 3. The four test input images

The results are sharper and closer to the ground truth than the image generated using LFR. Comparison between the two results obtained using the proposed method shows that the result created with τ = 5 is better in visual quality. These observations are quantitatively proved by the mean squared errors shown in table 2. Ghosting artifacts are visible in occluded boundaries in both the error image (Fig. 4 (d) and (f)) and the results generated with the proposed method (Fig. 4 4(c) and 4(e)). These ghosting artifacts are due to the fact that the linear model in (4) for g_1 and g_2 is incorrect for occluded boundaries.

In this simulation, $k^{(1)}$ and $k^{(5)}$ is a 9x9 and a 41x41 filter, respectively. $k^{(1)}$ is given by

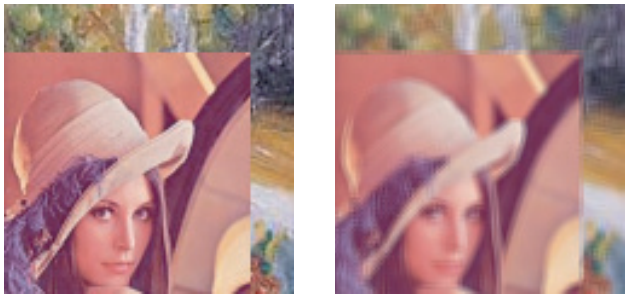
$$k^{(1)} = \frac{1}{32}\begin{pmatrix} 1 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -8 & 0 & 0 & 0 & -8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 36 & 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -8 & 0 & 0 & 0 & -8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 1 \end{pmatrix}. \quad (12)$$

$k^{(1)}$ has many zero coefficients, because d_1, d_2 and Δ are integer values in this simulation. The bigger value of τ is set, the larger the size of the reconstruction filter is. We could perform a filtering in the frequency domain using discrete Fourier transform for much faster reconstruction.

TABLE I

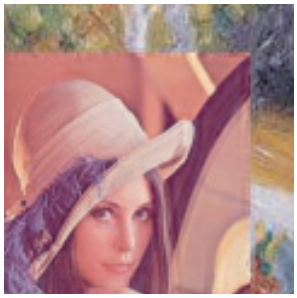
SIMULATION CONDITIONS AND PARAMETERS USED FOR CREATING THE INPUT IMAGES SHOWN IN FIG. 3

b	z_1	z_2	d_1	d_2	Δ	resolution
			[pixels]			
10	500	1000	4	2	2	256x256



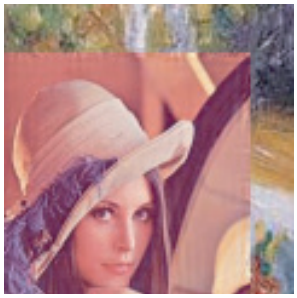
(a) Ground truth, l_v

(b) Optimal depth LFR



(c) The proposed method ($\tau = 1$)

(d) Error on (c): $2 \times \{(c) - (a)\} + 128$



(e) The proposed method ($\tau = 5$)

(f) Error on (e): $2 \times \{(e) - (a)\} + 128$

Fig. 4. Simulation results. (a) the ground truth synthesized at the center of the camera array. (b) the desired image generated by LFR method using an the optimal depth. (c) and (e) the desired image generated by the proposed method at τ = 1 and 5, respectively. (d) and (f) the corresponding error images of (c) and (e).

TABLE II

A COMPARISON OF THE MEAN SQUARE ERROR (MSE) OF EACH COLOR CHANNEL BETWEEN THE PROPOSED METHOD AND CONVENTIONAL LIGHT FIELD RENDERING (LFR).

method	red	green	blue
LFR (optimal) in fig. 4 (b)	262.2	250.4	225.2
Proposed ($\tau = 1$) in fig. 4 (c)	39.5	36.7	33.6
Proposed ($\tau = 5$) in fig. 4 (e)	22.2	20.7	22.9

IV. CONCLUSION AND FUTURE WORK

In this paper, we have proposed a direct filtering method that can generate a virtual image from a set of captured images of a scene containing two depth layers. The direct linear filter requires no depth map to be estimated. For future work, we would like to extend this approach to a general case where a scene can contain multiple layers and/or continuous depth objects. Such an extension would require a non-linear filter to be designed that can suppress errors due to occlusion boundaries. Finally, we would like to derive a generalized filter for reconstructing at arbitrary sample positions in the camera plane.

REFERENCES

[1] 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, Jan. 2004.

[2] H-Y. Shum, S. B. He, and S-C. Chan, "Survey of Image-Based Representations and Compression Techniques", *IEEE Trans. on CSVT*, Vol. 13, No. 11, pp. 1020 - 1037, 2003

[3] J-X. Chai, X. Tong, S.-C. Chany, H.-Y. Shum, "Plenoptic Sampling," *SIGGRAPH2000*, pp. 307-318, 2000

[4] A. Kubota, K. Aizawa, T. Chen, "Virtual View Synthesis through Linear Processing without Geometry," *ICIP2004*, pp. 3009-3012, 2004

[5] M. Levoy, P. Hanrahan, "Light field rendering," *SIGGRAPH96*, pp.31-42, 1996