View Interpolation using Defocused Multi-View Images

Akira Kubota* and Yoshinori Hatori†
* Chuo University, Kasuga, Bunkyo-ku, 112-8551 Tokyo, Japan
E-mail: kubota@elect.chuo-u.ac.jp
† Tokyo Institute of Technology, Yokohama 226-8502 Japan
E-mail: hatori@ip.titech.ac.jp

***Abstract*—This paper presented a view interpolation method for reconstructing intermediate all-in-focus images between defocused stereo images of a scene consisting of two depths. In the presented method, the novel view can be reconstructed as a sum of the filtered stereo images. The reconstruction filter was derived as a stable and linear space-invariant one; thus the presented method does not require region segmentation. The presented method can also reconstruct dense light field from multi-view images with different foci, so that free viewpoint images can be generated with better quality.**

I. Introduction

This paper addresses the problem of reconstructing a novel view using stereo images. In most conventional approaches, a central issue in this problem is to estimate information on 3-D scene geometry such as disparity map and feature correspondence [1], [2]. In this paper, avoiding this issue, we present a new approach that reconstructs the novel view directly from the stereo images through space-invariant filtering.

Since it is difficult to avoid this issue in general case, this paper considers a simple case where a scene has foreground and background objects at different two constant depths (Fig. 1). In our approach, unlike the conventional methods using pin-hole cameras, we use real aperture cameras to acquire stereo images with different focus settings: the left image g^L focused on the foreground and the right image g^R focused on the background. Our goal is to reconstruct an all-in-focus intermediate image f that would be captured with a pin-hole virtual camera between the stereo cameras. The presented method reconstructs the desired image by summing the filtered stereo images as follows:

$$f = k^L * g^L + k^R * g^R, \tag{1}$$

where $*$ denotes 2D convolution operation. We derive the filters, k^L and k^R, that are linear and space-invariant, independent of the scene geometry; hence this filter-based view interpolation does not require region segmentation. It takes advantage of defocus not for depth estimation (for instance [5]) but for directly reconstructing the novel view: the frequency response of the derived filter reveals that defocusing makes the filters more stable.

Our previously presented method [6] used defocused images for view interpolation in the same filtering framework. It, however, required two images at each viewpoint with different focus settings (i.e., four images in total); on the contrary the presented method requires a single image at each viewpoint (two images in total), which facilitates capturing dynamic scene.

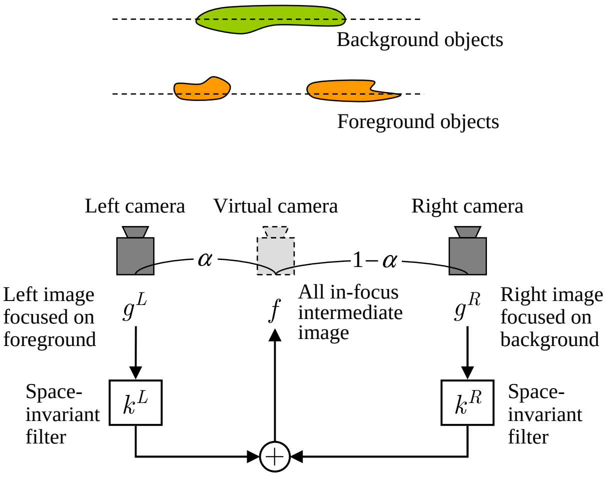


Fig. 1. View interpolation problem and our solution using space-invariant filters.

Our method can be extended to reconstructing dense light field using multi-view images that are captured with different alternate foci. We show that free viewpoint image can be generated with better quality from the reconstructed light field than from original sparse light field.

II. View Interpolation based on Space-Invariant Filtering

A. Problem description

As shown in Fig. 1, our problem setting is as follows. Stereo cameras are arranged such that their optical axes are in parallel. The scene consists of foreground and background objects that are located at two constant depths from the baseline of the stereo camera. The stereo cameras have different focus settings such that the left camera is focused on the foreground objects and the right camera is focused on the background objects. The captured left image g^L has in-focus foreground regions and defocused background regions; the right image g^R is vice versa (see Fig. 2 (a) and (b)).

We wish to reconstruct the novel image f that would be captured with a virtual camera between the stereo cameras.

The virtual camera has infinite depth-of-field and its optical axis is parallel to those of the stereo cameras. The novel viewpoint lies on the baseline and divides it in the ratio α and $1-\alpha$ (for $0 \leq \alpha \leq 1$).

B. Imaging model

To model the desired all-in-focus intermediate image f, we introduce two unknown textures, f_1 and f_2, that are defined as the segmented foreground and background textures, respectively, seen from the novel viewpoint. We model the desired all-in-focus intermediate image f as a sum of these textures:

$$f(x,y) = f_1(x,y) + f_2(x,y). \tag{2}$$

To model the defocused stereo images, we use a combination of these foreground and background textures that are shifted and blurred according to the viewpoints and the focus settings. Let d_1 and d_2 denote disparities of the foreground and the background objects between the stereo cameras. In modeling the left image g^L, we assume that it consists of the foreground texture that is shifted by αd_1 and the background texture that is shifted by αd_2 and blurred by the camera's point-spread function (PSF), say h. The right image g^R is model as a combination of the background texture that is shifted by $-(1-\alpha)d_2$ and the foreground that is shifted by $-(1-\alpha)d_1$ and blurred by PSF h. Thus, the model of the defocused stereo images can be expressed by

$$\left\{ \begin{array}{l} g^L(x,y) = f_1(x-\alpha d_1, y) + h(x,y) * f_2(x-\alpha d_2, y) \\ g^R(x,y) = h(x,y) * f_1(x+(1-\alpha)d_1, y) \\ \qquad\qquad\qquad + f_2(x+(1-\alpha)d_2, y) \end{array} \right. . \tag{3}$$

The PSF $h(x,y)$ is assumed to be Gaussian function: $h(x,y) = 1/(\pi\sigma^2)\exp(-(x^2+y^2)/\sigma^2)$, where σ is a blur parameter that indicates the degree of blur. The same blur parameter can be used for modeling defocused regions in both images [7]. We also assume that all the parameters σ, d_1 and d_2 are given through camera calibration. Note that these parameters are constant and independent of pixel position. Hence, unknowns in the models (2) and (3) are the texture images f_1 and f_2.

The imaging models in this paper neglect effect of unseen background (i.e., the occluded regions) unlike the layered representation considering occlusion (e.g. in [8]). Our model, however, has an advantage that it simply uses convolution operations without taking care of pixel-wise processing. Also the satisfactory result was obtained using our model as shown in the experimental results. The limitation of the model will be discussed in section III-B.

C. Filter Design in the frequency domain

We take the Fourier transform of equations (2) and (3) to obtain those imaging models in the frequency domain, as follows:

$$F(\omega_x,\omega_y) = F_1(\omega_x,\omega_y) + F_2(\omega_x,\omega_y) \tag{4}$$

and

$$\left\{ \begin{array}{l} G^L(\omega_x,\omega_y) = e^{-j\omega_x \alpha d_1} F_1(\omega_x,\omega_y) \\ \qquad\qquad + e^{-j\omega_x \alpha d_2} H(\omega_x,\omega_y) F_2(\omega_x,\omega_y) \\ G^R(\omega_x,\omega_y) = e^{j\omega_x (1-\alpha) d_1} H(\omega_x,\omega_y) F_1(\omega_x,\omega_y) \\ \qquad\qquad + e^{j\omega_x (1-\alpha) d_2} F_2(\omega_x,\omega_y) \end{array} \right. . \tag{5}$$

The capital letter function is the 2-D Fourier transform of the corresponding small letter function and (ω_x, ω_y) denotes horizontal and vertical radian frequencies. The function H is given by $\exp(-(\omega_x^2+\omega_y^2)\sigma^2/4)$.

Eliminating unknowns F_1 and F_2 from equations (4) and (5) yields the following simple sum-of-products formula:

$$F = K^L G^L + K^R G^R. \tag{6}$$

The functions K^L and K^R are the frequency response of the filters k^L and k^R in eq. (1) that we wish to derive. They are expressed by

$$\left\{ \begin{array}{l} K^L(\omega_x,\omega_y) = \dfrac{e^{j\omega_x \alpha d_1} - H e^{j\omega_x (d_1-(1-\alpha)d_2)}}{1 - H^2 e^{j\omega_x (d_1-d_2)}} \\ K^R(\omega_x,\omega_y) = \dfrac{e^{-j\omega_x (1-\alpha) d_2} - H e^{j\omega_x (\alpha d_1 - d_2)}}{1 - H^2 e^{j\omega_x (d_1-d_2)}} \end{array} \right. . \tag{7}$$

The derived filters are space-invariant because they consist of PSF and phase shift function. The required parameters for designing the filters are disparities d_1 and d_2 and blur parameter σ that can be obtained through camera calibration beforehand.

The denominator of the filters in equation (7), $1 - H^2 e^{j\omega_x(d_1-d_2)}$, determines the stability of the filters. Except for the DC (i.e., $(\omega_x,\omega_y) = (0,0)$), the denominator has non-zero value because of $H < 1$. This means defocus makes the filters stable; if we capture the stereo images with in-focus, then $H = 1$ and the denominator has zero values at the radian frequencies ($\omega_x = n\pi/(d_1 – d_2); n :$ integer) that satisfy $e^{j\omega_x(d_1-d_2)} = 1$. At the DC, it is found that both numerator and denominator of the filters become zero. Taking the limit of the filter to the DC, we obtain finite values as follows:

$$\lim_{\omega_r \to 0} K^L = \left\{ \begin{array}{ll} 1-\alpha & (\theta \neq \pm\pi/2) \\ 1/2 & (\theta = \pm\pi/2) \end{array} \right. \tag{8}$$

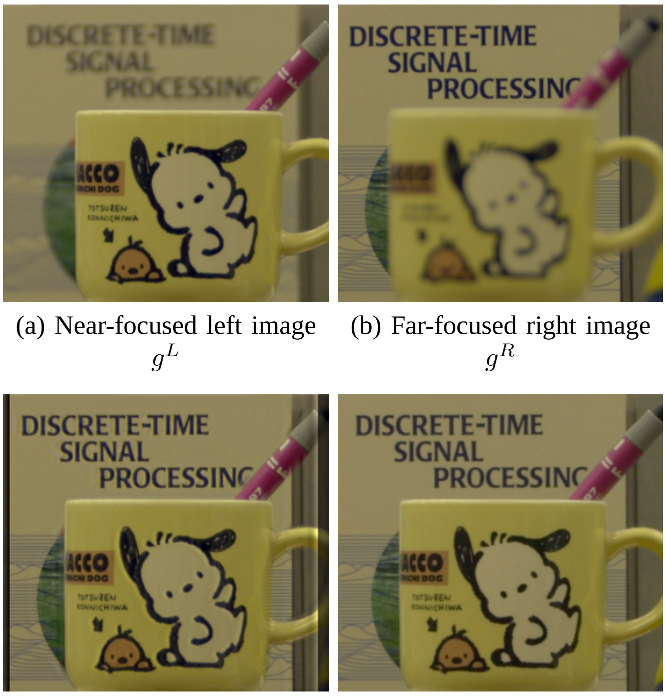
$$\lim_{\omega_r \to 0} K^R = \left\{ \begin{array}{ll} \alpha & (\theta \neq \pm\pi/2) \\ 1/2 & (\theta = \pm\pi/2) \end{array} \right. , \tag{9}$$

where $\omega_r = \sqrt{\omega_x^2 + \omega_y^2}$ and $\theta = \arctan(\omega_y/\omega_x)$. This indicates that the derived filters are stable. Although the DC components do not have unique limit value (except for the case of $\alpha = 0.5$), we use $1-\alpha$ and α as the DC component of K^L and K^R, respectively. Note that we experimentally confirmed that this did not affect image quality.

III. Experimental results

A. Experimental setup

In our experiment, we captured defocused stereo images by moving a single digital camera (Nikon D1) in horizontal direction. The test scene consists of a cup containing a pencil



(a) Near-focused left image g^L

(b) Far-focused right image g^R

(c) Reconstructed all-in-focus center image (the proposed method)

(d) Captured all-in-focus center image (used as the ground truth)

Fig. 2. Experimental results. (a) and (b) are the stereo images captured with different focus settings when the baseline was 8 mm. (c) is the all-in-focus center image reconstructed by the presented method. (d) is the ground truth image that was captured at the center with small aperture setting.

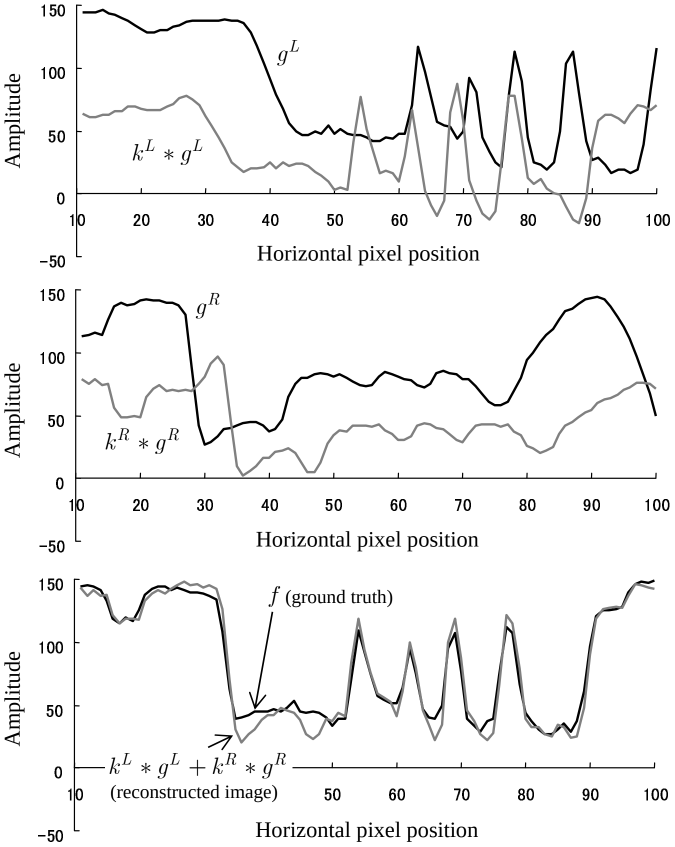


Fig. 3. Comparison among scanline signals of the original stereo images, their filtered images, the reconstructed image and the ground truth.

as the foreground object and text books as the background objects. These objects are located roughly at 1000 and 1600 [mm] from the baseline.

The captured stereo images (280×256 [pixel]) are shown in Fig. 2 (a) and (b) when the baseline length was set to 8 [mm]. Difference of focus setting is clearly shown between the images: the foreground is in-focus and the background is out-of-focus in the left image g^L; while vice versa in the right image g^R. Different disparities can be seen between the foreground and background regions (see the change of the horizontal position of the foreground cup and the background letters). The disparities and blur parameter obtained by our previously presented method [6] were $d_1 = 18.5$, $d_2 = 11.4$ and $\sigma = 3.3$ [pixel]. In addition, image size and brightness correction between captured images were conducted.

B. Results

Figure 2 (c) shows the novel image reconstructed by the presented method at the center of the baseline (i.e., $\alpha = 0.5$). This image was reconstructed as a sum of the filtered versions of the input stereo images using the derived filters in eq. (7). It can be seen that the in-focus regions in the stereo images are successfully fused into the reconstructed image. The image quality is similar to that of the ground truth image (Fig. 2 (d)) that was the real image captured at the same center position with small aperture. The measured PSNR of the reconstructed image was 32.3 [dB] in green channel.

Figure. 3 compares scanline signals (at 150th line) among the original stereo images, their filtered images, the reconstructed image and the captured ground truth image. The black line in Fig. 3 (top) is the scanline signal of the original left image g^L. The foreground region corresponds to the signal after 61st pixel position, in which range the signal has four peaks that are foreground textures in-focus. The black line in Fig. 3 (middle) is the scanline signal of the right image g^R. The signal after 43rd pixel position corresponds to the defocused foreground texture, where the four peaks almost disappear. The filtered signals $k^L * g^L$ and $k^R * g^R$ are plotted as the gray lines in Fig. 3 (top) and (middle), respectively. They contain in-focus regions that are appropriately shifted. Their sum signal is plotted as the gray line in Fig. 3 (bottom) compared with the ground truth signal of the black line. This result indicates that the reconstructed and ground truth signals are in close agreement.

To see how the reconstruction filters work, in Fig. 4, we plotted the frequency response of the reconstruction filters at $\omega_y = 0$: the magnitude response (i.e., $|K^L(\omega_x)|$ and $|K^R(\omega_x)|$) and the group shift (delay), which can be given by the gradient of phase shift. From this result, we found that the filters not just extract and shift the high-frequency

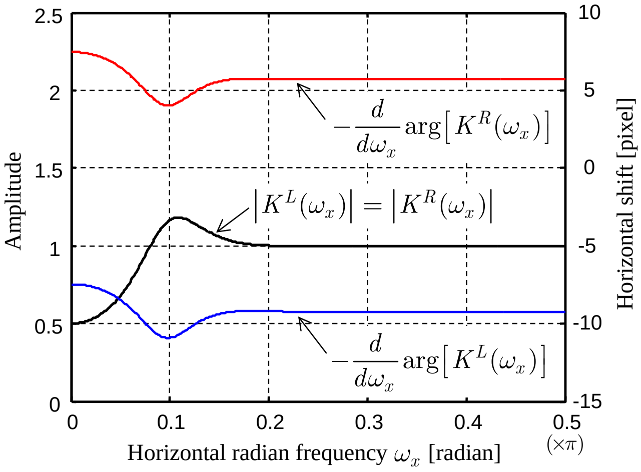


Fig. 4. Frequency magnitude responses and group shift for the reconstruction filters

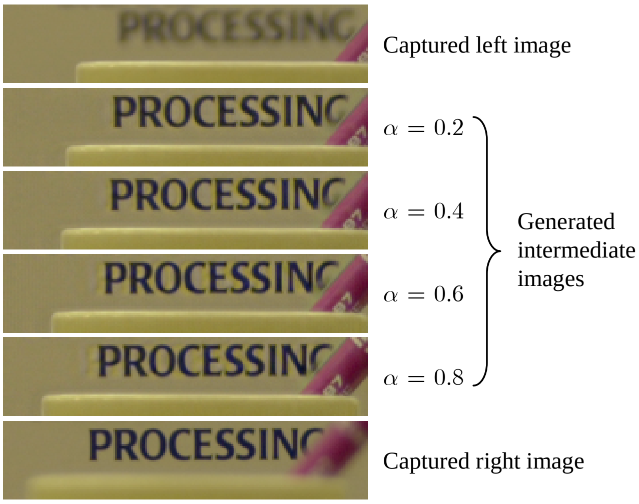


Fig. 5. Reconstructed intermediate images at different view positions.

components but appropriately amplify and shift the lower-frequency components to produce the desired image. This means that the novel image is reconstructed not only from the in-focus regions but also from the defocused regions of the stereo images.

Figure 5 shows the in-between images (close up) reconstructed by the presented method for α = 0.2, 0.4, 0.6 and 0.8 when the baseline length was 12 [mm]. It can be seen that the horizontal distance between foreground and background regions is changed according to the novel viewpoint. Although the occluded regions are not correctly reconstructed, visible artifacts do not appear. This is because the view reconstruction is based on stable filters.

A limitation of the presented method is that the image quality is degraded with an increase of the baseline length. One example of the reconstructed image is shown in Fig. 6 when the baseline length was 16 [mm]. The result shows that the image contains unclear occluded boundaries and suffers from ghosting artifacts (look like ringing artifacts) compared with the result of 8 [mm] baseline length in Fig. 2 (c). Since our method does not identify these regions, the quality degradation in occluded regions cannot be avoided. The occluded regions are filled with either of the stereo images in the presented method without visible error. The ghosting artifacts are may due to instability of the reconstruction filters. When the baseline length is increased, the difference of disparities $d_1 - d_2$ becomes large. In this case, the denominator of the filters, $1 - H^2 e^{j\omega_x(d_1-d_2)}$, is periodically close to zero in the lower frequency range where H can be approximated by 1; consequently the reconstruction filters have larger values and amplify artifacts arising from PSF calibration errors.

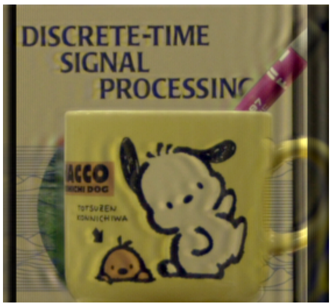


Fig. 6. Reconstructed images from the center of the stereo cameras with 16 [mm] baseline (the difference of disparities become $d1-d2$=14.2 [pixel])

IV. Generating Free Viewpoint Images

Our method can be applied to reconstructing dense light field from multi-view images with different foci. Once light field data were obtained, free viewpoint images can be generated by re-sampling the light field data. This method is called light field rendering (LFR) [3], [4].

Our experimental setup is shown in Fig. 7. The target scene was the same in the experiment in the previous section. We assume X and Z coordinates as a 2D world coordinate system. In our experiment, we captured 10 images with alternate focus settings at 10 different positions on the horizontal X axis with equal intervals of b=8 mm. Let us refer the camera at each position as C_1, C_2,..., and C_{10}, 10 captured images are 5 near focused images with cameras C_1, C_3, C_5, C_7 and C_9 and 5 far focused images with C_2, C_4, C_6, C_8 and C_{10}.

We constructed the reconstruction filters using the estimated parameters of blur amount and disparities described in the previous section. The reconstruction filters for stereo pair of far-focused left image and near-focused right image are derived as

$$\begin{cases} K^L(\omega_x,\omega_y) = \dfrac{e^{j\omega_x\alpha d_2} - He^{j\omega_x(d_2-(1-\alpha)d_1)}}{1-H^2e^{j\omega_x(d_2-d_1)}} \\ K^R(\omega_x,\omega_y) = \dfrac{e^{-j\omega_x(1-\alpha)d_1} - He^{j\omega_x(\alpha d_2-d_1)}}{1-H^2e^{j\omega_x(d_2-d_1)}} \end{cases}, \qquad (10)$$

which are given by respectively exchanging d_1 and d_2 for d_2 and d_1 in the filters of eq. (7). Based on the presented filtering method, we interpolated nine all-in-focus intermediate

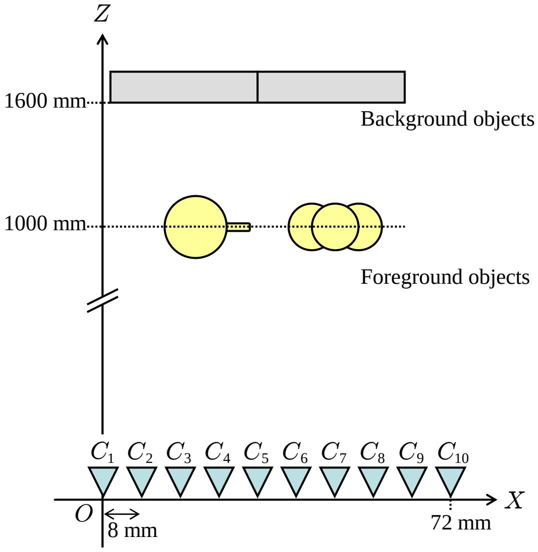

Z
1600 mm
Background objects
1000 mm
Foreground objects
C_1 C_2 C_3 C_4 C_5 C_6 C_7 C_8 C_9 C_{10}
O
X
8 mm
72 mm

Fig. 7. Configuration of scene and camera array used in the experiment.

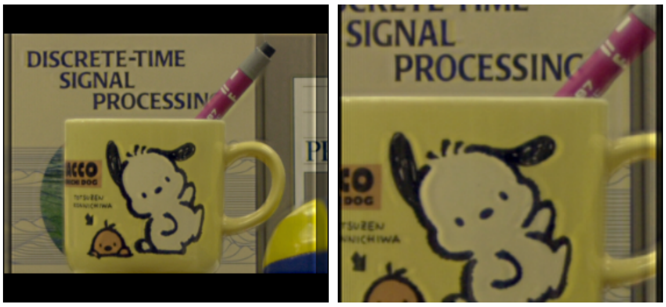
images between every pair of adjacent cameras, for a total of 91 images, including the all-in-focus images at the same positions as the cameras (that is, α=0 and 1). The parameter α was varied from 0 to 1 in equal increments of 0.1 for each interpolation.

Figure 8 shows examples of synthesized free viewpoint images when we changed the depth position of the viewpoint. We can see that the image in (b) is not merely a magnified version of the image in (a); the foreground objects (the cup and pencil) were magnified more than the background object. Figure 9 shows the novel view images from the same view positions by LFR using the all-in-focus images at the original 10 positions. The images suffer from blur or ghosting artifacts. The artifacts arise from pixel mis-correspondence due to the sparse light field.

V. Conclusion

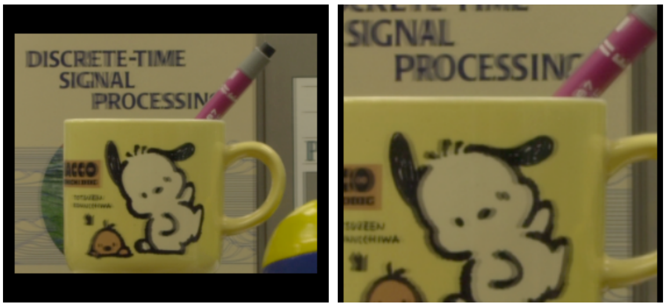
In this paper, we have presented a novel view synthesis method based on space-invariant filtering of defocused stereo image. Modeling the stereo images as a combination of layered textures with shifted and blurred, we derived the filters in the frequency domain. The experimental results on real images showed that view interpolation was possible using space-invariant filtering without requiring feature matching.

In future, we wish to extend our method to a scene with three and more depth layers. In this case, we need to capture multiple images with different focus settings; the number of images is the same of that of the layers. We also study camera calibration to accurately obtain the camera's PSF and the disparities using the acquired images or some test patterns (calibration board).



(a) (X_v, Z_v)=(20,-300) (b) (X_v, Z_v)=(20,300)

Fig. 8. Example of the novel views synthesized by LFR from the densely interpolated images. (X_v, Z_v) are the coordinates of the novel viewpoint along the horizontal and depth axes.



(a) (X_v, Z_v)=(20,-300) (b) (X_v, Z_v)=(20,300)

Fig. 9. Example of the novel views synthesized by LFR from the all-in-focus images at the original 10 positions.

References

[1] M. Oliveira, "Image-based modeling and rendering techniques: a survey," RITA–Rev. Inform. Teórica Aplicada IX(2), pp. 37–66, 2002.

[2] S.C. Chan, H.Y. Shum, and K.T. Ng, "Image-based rendering and synthesis," IEEE Signal Processing Mag. vol. 24, no. 7, pp. 22–33, 2007.

[3] T. Fujii, T. Kimoto and M. Tanimoto, "Ray Space Coding for 3D Visual Communication," proc. of Picture Coding Symposium 1996, pp. 447–451, 1996.

[4] M. Levoy, and P. Hanrahan, " Light Field Rendering, " proc. of ACM SIGGRAPH, pp. 31–42, 1996.

[5] 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.

[6] A. Kubota, K. Aizawa, T. Chen, "Reconstructing Dense Light Field From Array of Multifocus Images for Novel View Synthesis," IEEE Trans. on Image Processing, vol. 16, no. 1, pp. 269–279, 2007

[7] A. Kubota and K. Aizawa, "Reconstructing arbitrarily focused images from two differently focused images using linear filters," IEEE trans. on Image Processing, vol. 14, no. 11, pp. 1848–1859, 2005.

[8] S. W. Hainoff and K. N. Kutulakos, "A layer-based restoration framework for variable-aperture photography," proc. of ICCV2007, pp. 1–8, 2007.