A VARIATIONAL RECOVERY METHOD FOR VIRTUAL VIEW SYNTHESIS

Akira Kubota[1*] *and* *Takahiro Saito*[2]

[1] Interdisciplinary Graduate School of Science and Technology, Tokyo Institute of Technology, Japan
[2] Dept. of Electric Engineering, Kanagawa University, Japan

ABSTRACT

This paper presents a novel method based on image recovery scheme for virtual view synthesis. First, using multiple hypothetical depths, we generate multiple candidate images for the desired virtual view. The generated images suffer from blending artifacts (seen like blur) due to pixel mis-correspondence. From these blurry images, we recover an image without artifacts (i.e. an all infocus image) by minimizing an energy functional of unknown textures at all the hypothetical depths. The desired image is finally reconstructed as the sum of all the estimated textures. Simulation result shows that texture color value exist over all the hypothetical depths (i.e. depth is not uniquely identified for every pixel) nevertheless the desired image can be reconstructed with adequate quality.

Index Terms— virtual view synthesis, image based rendering, image recovery, energy minimization, total variation

1. INTRODUCTION

Virtual view synthesis problem using multi-view images has recently attracted further interests in image processing community. Two main approaches to this problem are image-based modeling and rendering (IBMR) [1] and image-based rendering (IBR) [2]. IBMR approach first reconstructs the scene geometry or estimate some information about the scene such as feature correspondences. Once the scene information is obtained, synthesizing a novel view can be easily done. It is however generally hard to obtain scene geometry in precise. In contrast to IBMR, IBR approach treats the view synthesis problem as a sampling problem [3] without estimating scene information. It samples light rays (captures multi-view images) densely enough to resample them to create novel light rays (a novel view) without aliasing artifacts. The required number of samples, given by the light-field sampling theorem [4], is quite many for the most practical applications.

In this paper, we tackle the view synthesis problem using an image recovery technique. The proposed method consists of two steps. In the first step, multiple candidate images for the desired virtual view are generated based on multiple hypothetical depths. The resultant candidate images suffer from blending artifacts (seen like blur or ghosting artifacts). These artifacts are due to pixel mis-correspondences arising from the difference between the hypothetical and the actual depths.

In the second step, we recover the virtual view image without artifacts (i.e. all infocus image) from the blurry candidate images. To this end, we formulate an energy functional of unknown textures existing at the hypothetical depths and minimize it to estimate these textures. The energy functional consists of data-fidelity and regularization terms. The former evaluates errors between the candidate images and the images that we model as a linear combination of all textures with artifacts. The latter imposes smoothness of pixel values in the final virtual image by evaluating a total variation of the image. The virtual view image is finally reconstructed as the sum of the estimated textures.

The minimizing process does not require feature matching; hence all the estimated textures have some color value at each pixel, i.e. the depth of each pixel can not be uniquely identified. However the final virtual image can be reconstructed with adequate quality as the sum of such distributed textures.

2. PROBLEM SETTING

We set the XYZ world coordinate system in a 3D space and assume that all cameras are arranged on the XY plane with regularly spaced and parallel to the Z axis. In this case, the Z axis represents depth from the cameras. Let $f_{s,t}(x,y)$ be the reference image captured with the camera $C_{s,t}$ at a grid position (X_s, Y_t) on the XY plane, where (x,y) is image coordinate and $(s,t) \in \mathbb{Z}^2$ is the index for both the reference images and the capturing cameras. The distance between cameras is Δ ($= |X_{s+1} - X_s| = |Y_{t+1} - Y_t|$).

The view synthesis problem we address in this paper is, given a virtual camera C_v at an arbitrary position (X_v, Y_v, Z_v), to reconstruct the virtual view f_v using the reference images $\{f_{s,t}(x,y)\}$. Note that the scene geometry is not known but we assume the depth range $[Z_{\min}, Z_{\max}]$ is known.

*kubota@ip.titech.ac.jp

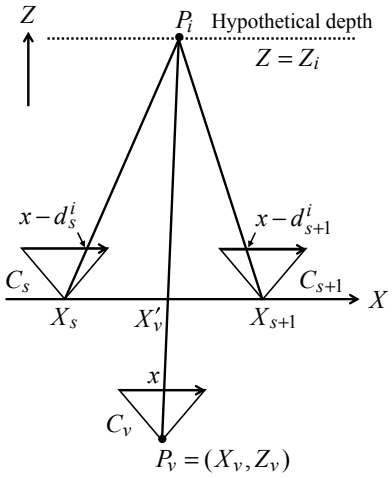


Fig. 1: Candidate image generation by light field rendering based on the hypothetical depth

3. THE PROPOSED METHOD

3.1. Step 1: Generating candidate images

We generate candidate images for the desired virtual view by using light field rendering (LFR) method [5, 6] based on multiple hypothetical depths. For simple notation, we neglect parameters Y and y, as shown in fig. 1.

Let $g_i(x)$ be the candidate image generated based on the depth Z_i, where $i = 1, ..., N$; N is the number of the candidate images (the same as that of the depths). The image $g_i(x)$ is computed as the weighted average of the shifted two reference images:

$$g_i(x) = w_s \cdot f_s(x - d_s^i) + w_{s+1} \cdot f_{s+1}(x - d_{s+1}^i). \quad (1)$$

The three image coordinates in the above equation, x, $x - d_s^i$ and $x - d_{s+1}^i$, are the corresponding pixel coordinates with respect to the point P_i at depth Z_i (see fig. 1). The displacement d_s^i is calculated as

$$d_s^i(x) = (X_s - X_v + Z_v x)/Z_i, \quad (2)$$

where focal length is normalized to be 1 for both capturing and virtual cameras.

Two images f_s and f_{s+1} used are selected such that their camera position be nearest to the position X'_v that is the intersection of the line $P_v P_i$ with the X axis. The weighting values w_s and w_{s+1} are determined to be $w_s = |X_{s+1} - X'_v|/\Delta$ and $w_{s+1} = |X_s - X'_v|/\Delta$, respectively. Note that $w_s + w_{s+1} = 1$ holds.

It is clear that none of the generated candidate images $\{g_i\}$ can be the same as the desired view f_v since we assumed the scene geometry is a plane. In the candidate images, the regions appear in focus when the hypothetical depth is at their actual depth; otherwise the regions appear blurry due to pixel mis-matching (see images in fig. 2 (a)-(d).).

3.2. Step 2: Recovering an all-focused virtual view by regularized variational method

3.2.1. Linear image formation model

We introduce image formation models for the candidate images $\{g_i\}$ and the desired all infocus virtual view f_v. These models were firstly presented in our previous paper [7]. We follow them for the most part.

We assume that the desired all focused view f_v is composed of the sum of N components $\{\varphi_j\}$ $(j = 1, ..., N)$:

$$f_v = \sum_{j=1}^{N} \varphi_j. \quad (3)$$

The component φ_j is defined as the unknown texture existing at depth Z_j. No other constraints on each texture are used in this paper, which is different from the texture model in [7].

The model of the candidate images g_i is expressed by the simultaneous equations [7]:

$$\begin{cases} g_1 = h_{11} \circ \varphi_1 + h_{12} \circ \varphi_2 + \cdots + h_{1N} \circ \varphi_N \\ g_2 = h_{21} \circ \varphi_1 + h_{22} \circ \varphi_2 + \cdots + h_{2N} \circ \varphi_N \\ \quad \vdots \\ g_L = h_{N1} \circ \varphi_1 + h_{N2} \circ \varphi_2 + \cdots + h_{NN} \circ \varphi_N, \end{cases} \quad (4)$$

where h_{ij} denotes the blurring process on the texture φ_j in g_i. For $i = j$, h_{ij} becomes an identity operation.

Each blurring process is modeled as a spatially varying filtering as follows. Consider the case when the scene contains one plane object at depth Z_j. In this case, the model (4) is represented by

$$g_i(x) = h_{ij} \circ \varphi_j(x), \quad i = 1, ..., N. \quad (5)$$

Assuming surface property of the object plane is lambertian, we have the relationship

$$f_v(x) = \varphi_j(x) = f_s(x - d_s^j) = f_{s+1}(x - d_{s+1}^j) \quad (6)$$

and substitute this into eq. (1) to obtain

$$g_i(x) = w_s \varphi_j(x - d_s^i + d_s^j) + w_{s+1} \varphi_j(x - d_{s+1}^i + d_{s+1}^j). \quad (7)$$

Comparing the above equation with eq. (5) tells us that the operation h_{ij} can be modeled as a filter whose coefficients are the weighting values w_s and w_{s+1} and that it is linear but shift varying since the displacements (e.g. d_s^i) varie with x (shown in eq. (2)).

3.2.2. Energy functional

We define an energy functional of textures $\{\varphi_j\}$ as follows:

$$D[\varphi_1, ..., \varphi_N] = \int_\Omega \Big(\|\nabla f_v\| + \frac{\lambda}{2} \textstyle\sum_{i=1}^{N} e_i^2 \Big) dx, \quad (8)$$

where Ω denotes the domain of the image space and λ is a positive parameter. The first term in the energy functional is the regularization term that evaluates the total variation of the desired view f_v, imposing smoothness constraint on it. The second term is the data-fidelity term that evaluates the square of error e_i defined as

$$e_i = (h_{i1} \circ \varphi_1 + \cdots + h_{ii} \circ \varphi_i + \cdots + h_{iN} \circ \varphi_N) - g_i,$$

which is the difference between g_i and its formation model in eq. (4).

3.2.3. Energy minimization

Euler-Lagrange equation minimizing the energy functional $D[\varphi_1, ..., \varphi_N]$ with respect to φ_j is given as the following partial differential equation (PDE):

$$\mathrm{div}\left[\frac{\nabla\varphi_j}{\|\nabla f_v\|}\right] - \lambda \sum_{i=1}^{N} \left(h_{ij}^* \circ e_i\right) = 0, \tag{9}$$

where operator h_{ij}^* denotes the adjoint operator of h_{ij}.

We solve the solution of the PDE as the steady-state solution of the following time-evolution PDE:

$$\begin{aligned} \frac{\partial}{\partial\tau}\varphi_j &= \mathrm{div}\left[c(x;\tau)\nabla\varphi_j\right] - \lambda \sum_{i=1}^{N} \left(h_{ij}^* \circ e_i\right), \qquad (10) \\ c(x;\tau) &= \min(1, 1/\|\nabla f_v\|), \\ \varphi_j(x;0) &= g_j/N, \end{aligned}$$

where τ is an artificial time-variable and $\varphi_j(x;0)$ is the initial estimate for the texture $\varphi_j(x)$. The final solution of the desired view f_v is given as the sum of the obtained solutions of $\varphi_1, ..., \varphi_N$.

The time-evolution PDE in eq. (10) acts as a nonlinear diffusion process [8] with the conduction coefficient of c when λ equals to zero. In addition, if c is a constant, it reduces to the isotropic heat diffusion, which is identical to a blurring process by Gaussian kernel. To prevent large diffusion that causes blur in the solution, we use the conduction coefficient c as $\min(1, 1/\|\nabla f_v\|)$ instead of use of $1/\|\nabla f_v\|$. This is a similar idea used in robust anisotropic diffusion [9]. The second term in eq. (10) acts as a pseudo-inverse process, i.e. a back projection image recovery.

Notably, the presented recovering process does not perform feature matching but find the combination of textures φ_j that gives the minimum of the energy functional we define.

4. SIMULATION

We tested the performance of the presented method for the synthetic scene that consists of a glossy sphere and a pole in front of reflective two walls, involving reflections, shadows and occlusions (fig. 2). The scene exists from 65 to 100 in depth. For this scene, we created a set of 9x9 reference images $\{f_{s,t}\}$ (which are 24 bits color images of 320x240 pixels). The distance between cameras was set to $\Delta = 1$.

Table 1: PSNR improvement in each color channel of the recovered final view after 300 terations for three different virtual view positions. (unit: [dB])

view point: (X_v, Y_v, Z_v)	Red	Green	Blue
P_1: (0.5, 0.5, -1.0)	2.36	2.07	1.86
P_2: (-0.5, -0.5, -1.0)	2.24	1.82	1.44
P_3: (1.5, 1.5, 2.0)	2.48	2.15	1.77

Taking into account the reflected textures on the back walls, we used four hypothetical depths at Z_1=65, Z_2=78, Z_3=98, and Z_4=130; this range is larger than the scene depth. Z_2 and Z_3 were calculated as $1/Z_2 = (3/Z_1 + 1/Z_3)/4$ and $1/Z_3 = (1/Z_1 + 3/Z_3)/4$, respectively, based on the efficient arrangement [4]. Four candidate images generated in the first step from a novel viewpoint (0.5, 0.5,-1) are shown in fig. 2 (a)-(d). In each image, the regions near the hypothetical depth appear infocus, while other regions appear blurry and contain ghosting artifacts.

In the second step, the discrete version of the PDE in eq. (10) was solved iteratively. The initial solution, which is the average of the candidate images, and the finally recovered view that is the solution after 300 iterations are shown in fig. 2 (e) and (f), respectively. Parameter λ was set to 2. Comparison of them with the ground truth in fig. 2 (g) shows that the presented method effectively recovers the all-focused view with adequate quality.

As the performance measure, PSNR improvement that is an increment of PSNR of the reconstructed image after 300 iterations are shown in table 1 for three cases of different view points. The result shows quality is improved by around 2 dB for all the cases.

The estimated textures φ_j after 300 interations are shown in fig. 3. This result somewhat impressively shows that at every pixel color value exists over all four depths not at one depth. Nevertheless, all focused image is correctly reconstructed as the sum of these textures. This follows the fact that depth of uniform regions do not need to be uniquely identified but depth of non-uniform (texture) regions should be.

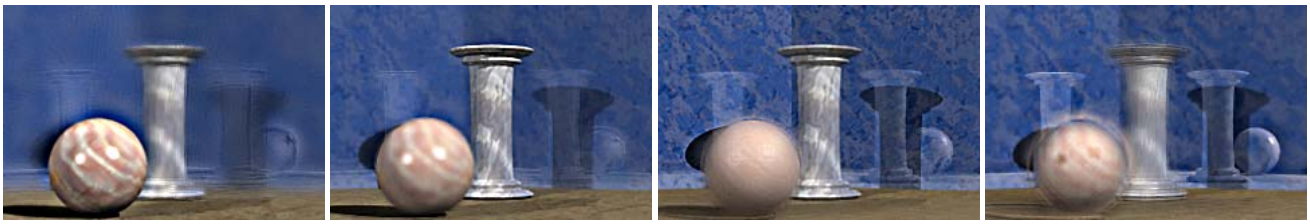
5. SUMMARY

In this paper, we have presented a novel view synthesis method based on image recovery scheme. Generating multiple candidate images and using them as initial estimates, we recover the desired image without artifacts by the regularized variational method, not requiring feature matching.

(a) g_1 (b) g_2 (c) g_3 (d) g_4

(e) The initial estimate (f) The final reconstructed image (g) The ground truth

Fig. 2: Simulation result. (a)-(d): generated candidate images; (e)-(g): comparison between the initial estimate, the finally reconstructed image after 300 iterations, and the ground truth.



(a) φ_1 (b) φ_2 (c) φ_3 (d) φ_4

Fig. 3: Obtained texture components φ_j after 300 iterations. Intensity values are multiplied by 4 for visibility.

6. REFERENCES

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

[2] 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.

[3] E. H. Adelson, J. R. Bergen, "The plenoptic function and the elements of early vision," *Computational Models of Visual Processing The MIT Press, Cambridge, Mass.* 1991.

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

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

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

[7] A. Kubota, K. Takahashi, K. Aizawa, T. Chen, "All-Focused Light Field Rendering," Eurographics Symposium on Rendering (EGSR2004), pp. 235-242, June 21-23, 2004

[8] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," IEEE Trans. Pattern. Anal. & Mach. Intell., 12, 7, pp.629-639, 1990.

[9] M. J. Black, G. Sapiro, D. H. Marimont, and D. Heeger, "Robust anisotropic diffusion," IEEE Trans. on Image Processing, 7, 3, pp.421-432, 1998.