Super-Resolved Free-Viewpoint Image Synthesis Using Semi-global Depth Estimation and Depth-Reliability-Based Regularization

Keita Takahashi[1] and Takeshi Naemura[2]

[1] The University of Electro-Communications
1-5-1 Chofugaoka, Chofu-shi, Tokyo 182-8585, Japan
keita.takahashi@ieee.org
http://nae-lab.org/~keita/
[2] The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

Abstract. A method for synthesizing high-quality free-viewpoint images from a set of multi-view images is presented. First, an accurate depth map is estimated from a given target viewpoint using modified semi-global stereo matching. Then, a high-resolution image from that viewpoint is obtained through super-resolution reconstruction. The depth estimation results from the first step are used for the second step. First, the depth values are used to associate pixels between the input images and the latent high-resolution image. Second, the pixel-wise reliabilities of the depth information are used for regularization to adaptively control the strength of the super-resolution reconstruction. Experimental results using real images showed the effectiveness of our method.

Keywords: free-viewpoint image, semi-global stereo, super-resolution, depth reliability, regularization.

1 Introduction

Free-viewpoint image synthesis refers to the process of combining a set of multi-view images to generate an image from a new viewpoint where no camera was actually located. This technology has attracted much research interest due to its potential for representing 3-D visual information [1]; using this technology, users can fly through the 3-D space and can also display real objects on auto-stereoscopic 3-D displays with tens of parallax views [2].

In this work, we reconsider the framework of free-viewpoint image synthesis. In general, this synthesis consists of two steps: first, the depth/shape of the target scene is estimated from input images; then, using the estimated depth/shape, the input images are registered and blended together to produce a new image. The blending operation in the second step can obscure depth/shape errors by blurring the image. However, this scheme has a fundamental limitation in the resolution of the resultant image; fine textures are decayed due to the blurring nature of blending.

Y.-S. Ho (Ed.): PSIVT 2011, Part I, LNCS 7087, pp. 22–35, 2011.

A promising solution for improving resolution is to replace blending by super-resolution (SR) reconstruction [3] because multiple observations of the same scene are given as input. However, SR reconstruction is very sensitive to registration errors; it could even be destructive if applied with large registration errors. Estimating perfect depth/shape information from images alone is far beyond the capability of current computer vision technologies, so some extent of registration errors should be accepted. Consequently, we conceived of the idea to combine blend-based synthesis and SR-based synthesis adaptively according to the reliability of the estimated depth information.

On the basis of this idea, we propose a new method for super-resolved free-viewpoint image synthesis. Our method has three features. First, we adopt a view-dependent approach like that in Refs. [4,5]; we focus on image synthesis from the given target viewpoint rather than complete reconstruction of the 3-D structure. More precisely, our method works directly on the coordinate system of the target free-viewpoint image. The second feature is semi-global depth estimation based on that in Refs. [6,7], which achieves accurate depth estimation with considerably low computational costs. The final feature is depth-reliability-based regularization, which can control the strength of SR reconstruction according to the pixel-wise reliability of the depth information. This regularization is the key to achieving high-quality synthesis and also provides a new framework where blend-based synthesis and SR-based synthesis are adaptively combined. The effectiveness of our method was confirmed by experiments using real images.

1.1 Background

Super-resolution (SR) reconstruction [3] combines multiple low-resolution images to restore a latent high-resolution image. One of the input images is selected as the basis image to which other input images are registered and for which the resulting high-resolution image is synthesized. Then, an image formation model is established between the input and latent high-resolution images. Finally, by inverting the image formation model with prior knowledge, the latent high-resolution image can be restored. However, the viewpoint of the resulting image is limited to that of one of the input images because this technology is not designed for producing free-viewpoint images.

Free-viewpoint image synthesis has been studied in a different context [1]. As mentioned above, most conventional methods use blend-based synthesis, resulting in the fundamental limitation of the image resolution. To our knowledge, only a few works use SR reconstruction for free-viewpoint image synthesis. Tung et al. [8] super-resolved input multi-view images, and Goldluecke et al. [9] synthesized texture maps using SR reconstruction. Their purpose was to generate a complete 3-D model of a single object. In contrast, our method takes a view-dependent approach for synthesizing free-viewpoint images and deals with the entire scene (which includes both objects and backgrounds). The most similar work to ours is that of Mudenagudi et al. [10]. They formulated view-dependent SR reconstruction of an entire scene as a multiple-labeling problem, where a label corresponds to the color of each pixel of the resulting image. However,

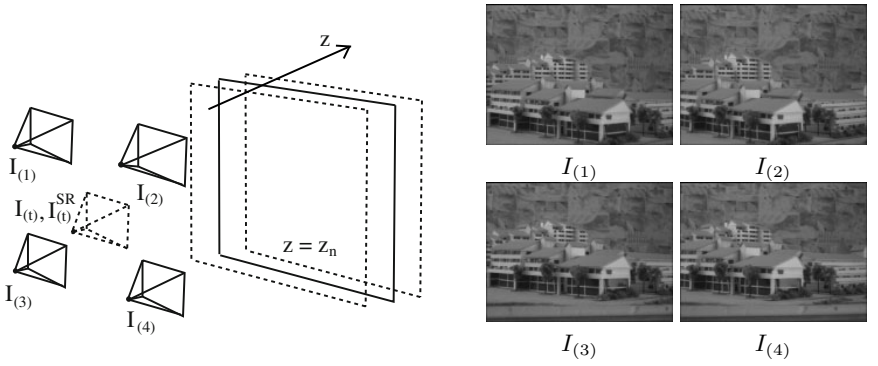


Fig. 1. Configuration used in proposed method (left) and input images (right)

their method was computationally complex and expensive due to the nature of the formulation. Our method, which consists of view-dependent depth estimation followed by SR reconstruction with depth-reliability-based regularization, is computationally more tractable and would be easier to speed-up for real-time applications in the future. Our previous work in Ref. [11] also aimed for view-dependent SR reconstruction. But due to the poor depth estimation and non-efficient algorithm design, it is incomparable to the method presented in this paper.

2 Overview of Proposed Method

The configuration used by our method is shown in Fig. 1. The input images, denoted by $I_{(m)}$ $(m = 1, \ldots M)$, are captured from viewpoints that are arranged roughly on the same plane. The camera parameters are estimated beforehand. The distance from the input camera plane is denoted by z. The goal of our method is to synthesize an image viewed from a new viewpoint, referred to as the target viewpoint, which is denoted by t. We define two synthesized images, $I_{(t)}$ and $I^{SR}_{(t)}$. $I_{(t)}$ is produced by blend-based synthesis and has the same resolution as the input images; $I^{SR}_{(t)}$ is produced by our new SR-based scheme. We assume that four images are given as the input and that the resolution of $I^{SR}_{(t)}$ is twice that of $I_{(t)}$ both in the horizontal and vertical directions. However, our framework can easily be extended to more general setups.

In general, our method first registers the input images to the coordinate system of the target viewpoint t and then applies a SR scheme to obtain a high-resolution resulting image. Registration of multi-view images is equivalent to depth estimation. In particular, if pixel-wise depth information from the target viewpoint is available, all pixels of the target image can be associated with the pixels of the input images, which is sufficient for constructing an image formation model for SR reconstruction. Thus, the first step of our method is to estimate a depth map viewed from the target viewpoint (described in Section

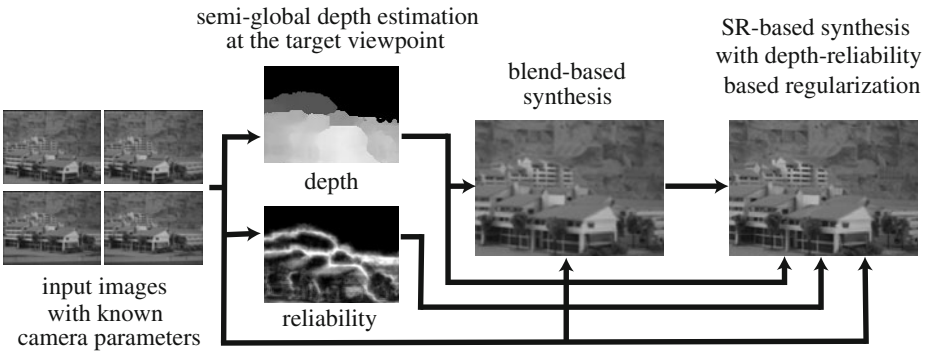


Fig. 2. Flowchart of our method

3). To estimate accurate depth with reasonable speed, we use the semi-global stereo method [6,7], modified for our problem. The depth map is estimated in the same resolution as the input images. It is then upsampled to the resolution of $I_{(t)}^{SR}$ and used for the next SR reconstruction step (described in Section 4). In SR reconstruction, the per-pixel reliability of the depth information, which is obtained through the depth-estimation step, is also used to control the strength of the regularization. This scheme, referred to as depth-reliability-based regularization, is the key to achieving high-quality synthesis, since it can adaptively combine blend-based synthesis and SR-based synthesis. Our entire method is summarized in Fig. 2.

3 Semi-global Depth Estimation

The first step of our method is to estimate a depth map from the target viewpoint. In Section 3.1, we briefly describe the semi-global stereo method [6,7], which is extended to our free-viewpoint setup in Section 3.2. The obtained depth map is further refined in Section 3.3.

3.1 Semi-global Stereo Matching

The purpose of stereo matching, given two or more input images, is to find pixel correspondences between the images. This is equivalent to depth estimation if the camera parameters are known. Typically, one of the input images is selected as the basis image for which the depth value of each pixel is estimated.

Modern stereo methods consider not only the photometric consistency between the input images (i.e., corresponding pixels should exhibit similar intensities/colors) but also the inter-pixel relations in the estimated depth map (i.e., depth values should not vary drastically except around the object boundaries). These conditions are represented as an energy minimization problem, whose optimal solution can be found using sophisticated numerical methods. The most common choices for optimization are belief-propagation and graph-cut, but they

are computationally prohibitively expensive since many iterations are required for convergence [12]. In contrast, semi-global stereo matching [6,7] can find a near-optimal solution with much lower computational cost because no iterative calculations are needed.

The energy function for semi-global stereo matching is described as

$$E_{sm}(D) = \sum_{\boldsymbol{p}} \left\{ C(\boldsymbol{p}, D(\boldsymbol{p})) + \sum_{\boldsymbol{q}\in N_{\boldsymbol{p}},\ |D(\boldsymbol{p})-D(\boldsymbol{q})|=1} \lambda_1 + \sum_{\boldsymbol{q}\in N_{\boldsymbol{p}},\ |D(\boldsymbol{p})-D(\boldsymbol{q})|>1} \lambda_2 \right\}, \quad (1)$$

where D is the resulting depth map and $D(\boldsymbol{p})$ is the depth of a pixel $\boldsymbol{p}$ that is represented as an integer disparity value. The first term evaluates the photometric consistency between the input images for each pixel $\boldsymbol{p}$ with the assumed depth $D(\boldsymbol{p})$. The second and third terms penalize depth discontinuities; $N_{\boldsymbol{p}}$ is the neighbor of $\boldsymbol{p}$, and λ_1 and λ_2 are non-negative weights, where $\lambda_1 \leq \lambda_2$.

The optimization procedure is very similar to dynamic programming. First, the photometric consistency cost $C(\boldsymbol{p}, n)$ is obtained for all pixels and all depth levels. Then, it is accumulated along the 1-D path with a direction r as

$$\begin{aligned} L_{\boldsymbol{r}}(\boldsymbol{p}, n) = C(\boldsymbol{p}, n) &- \min_k L_{\boldsymbol{r}}(\boldsymbol{p}-\boldsymbol{r},\ k) \\ &+ \min\{L_{\boldsymbol{r}}(\boldsymbol{p}-\boldsymbol{r}, n),\ L_{\boldsymbol{r}}(\boldsymbol{p}-\boldsymbol{r},\ n-1)+\lambda_1, L_{\boldsymbol{r}}(\boldsymbol{p}-\boldsymbol{r},\ n+1)+\lambda_1, \\ &\quad \min_k L_{\boldsymbol{r}}(\boldsymbol{p}-\boldsymbol{r},\ k)+\lambda_2\}. \end{aligned} \quad (2)$$

The accumulated costs for 8 or 16 directions (8 are used in this work) are added to yield $S(\boldsymbol{p}, n)$. Finally, a semi-optimal depth map $D(\boldsymbol{p})$ is obtained through a minimum search over the depth levels for each pixel $\boldsymbol{p}$.

$$D(\boldsymbol{p}) = \arg\min_n S(\boldsymbol{p}, n), \text{ where } S(\boldsymbol{p}, n) = \sum_{\boldsymbol{r}} L_{\boldsymbol{r}}(\boldsymbol{p}, n). \quad (3)$$

In the post-processing step, isolated noises are removed from the resulting depth map, but this step is beyond the scope of this paper.

3.2 Extension to Arbitrary Viewpoint Setups

The semi-global stereo method [6,7] was designed to work on the coordinate system of the basis image that is selected from the input images, similar to most stereo methods. However, our purpose is to estimate a depth map directly from the arbitrary target viewpoint where free-viewpoint image synthesis is performed. In this subsection, the coordinate system of the resulting depth map D is set to the target viewpoint, and we introduce three modifications to the original semi-global stereo method [6,7].

First, disparities cannot uniquely be defined to represent depth in our problem because the target viewpoint is set to an arbitrary position. Instead of using disparities, we quantize the depth space into N levels as

$$\frac{1}{z_n} = \frac{1}{z_{\max}} + \frac{n-1/2}{N}\left(\frac{1}{z_{\min}} - \frac{1}{z_{\max}}\right) \ (n = 1, \ldots, N), \quad (4)$$

where $z_{\min}$ and $z_{\max}$ are the minimum and maximum of the object depths. This quantization is natural because the disparity space (which is proportional to the inverse of the depth) is evenly divided, similar to most stereo methods. In our method, each pixel of the depth map $D(\boldsymbol{p})$ takes an integer that represents the depth index. The physical depth value for $D(\boldsymbol{p})$ can be written as $z_{D(\boldsymbol{p})}$.

Second, the photometric consistency in Eq. (1) should be given over the coordinate system of the target viewpoint. Consequently, we have to map the pixels from the target viewpoint to the input viewpoints in evaluating the consistencies. Specifically, we define $C(\boldsymbol{p},\ D(\boldsymbol{p}))$ as

$$C(\boldsymbol{p},\ D(\boldsymbol{p})) = \frac{1}{\mathcal{Z}} \sum_{\boldsymbol{q} \in B_{\boldsymbol{p}}} \left\{ \sum_{m \neq m'} C_{m,m'}(\boldsymbol{q},\ D(\boldsymbol{q})) \right\}, \tag{5}$$

where $B_{\boldsymbol{p}}$ is a window centered at $\boldsymbol{p}$, m and m' are the indices of input images, and $\mathcal{Z}$ is a constant for normalization. $C_{m,m'}(\boldsymbol{p},\ D(\boldsymbol{p}))$ evaluates the consistency for a pixel $\boldsymbol{p}$ of the target image as

$$C_{m,m'}(\boldsymbol{p},\ D(\boldsymbol{p})) = \mathrm{diff}\left(I_{(m)}(\mathrm{P}_{t \to m}(\boldsymbol{p}, z_{D(\boldsymbol{p})}),\ I_{(m')}(\mathrm{P}_{t \to m'}(\boldsymbol{p}, z_{D(\boldsymbol{p})}))\right), \tag{6}$$

where $P_{\alpha \to \beta}(\boldsymbol{p}, z)$ is a function that maps a point $\boldsymbol{p}$ on the camera α onto the camera β with a known depth z. The derivation details are described in the appendix A.1. The function diff is defined as

$$\mathrm{diff}(a, b) = \min\left\{ \|a - b\|^2, \mathrm{diff}_{\max} \right\}, \tag{7}$$

where $\mathrm{diff}_{\max}$ is an upper limit for the difference values. Giving an upper limit is useful for handling occlusions because we have multiple pairs of input images.

Third, λ_2 is set to a constant in our method, while it was set to be proportional to the inverse of the image gradient in the original [6,7]. In our problem, the image gradients are unavailable directly because the coordinate system is set to a new viewpoint from which no image was captured.

3.3 Depth Refinement

The depth map D obtained through the previous step takes discrete integer values. These values can be refined by fitting parabolic curves to the energy values $S(\boldsymbol{p}, n)$ around the minimums. For each pixel $\boldsymbol{p}$, the refined depth value $\hat{D}(\boldsymbol{p})$ and corresponding energy value $\hat{S}_{\min}(\boldsymbol{p})$ are given by

$$\hat{D}(\boldsymbol{p}) = D(\boldsymbol{p}) + \frac{S^{pre}_{\min}(\boldsymbol{p}) - S^{next}_{\min}(\boldsymbol{p})}{2(S^{pre}_{\min}(\boldsymbol{p}) - 2S_{\min}(\boldsymbol{p}) + S^{next}_{\min}(\boldsymbol{p}))} \tag{8}$$

$$\hat{S}_{\min}(\boldsymbol{p}) = S_{\min}(\boldsymbol{p}) - \frac{(S^{pre}_{\min}(\boldsymbol{p}) - S^{next}_{\min}(\boldsymbol{p}))^2}{8(S^{pre}_{\min}(\boldsymbol{p}) - 2S_{\min}(\boldsymbol{p}) + S^{next}_{\min}(\boldsymbol{p}))}, \tag{9}$$

where $S_{\min}(p) = S(\boldsymbol{p}, D(\boldsymbol{p}))$, $S^{next}_{\min}(p) = S(\boldsymbol{p}, D(\boldsymbol{p}) + 1)$, and $S^{pre}_{\min}(\boldsymbol{p}) = S(\boldsymbol{p}, D(\boldsymbol{p}) - 1)$ (see appendix A.2 for the derivation). This refinement is equivalent to

interpolation in the disparity space because the value of $D(\boldsymbol{p})$ is proportional to the inverse of the depth. The resulting depth map $\hat{D}$ takes continuous values, but the corresponding physical depths can also be obtained from Eq. (4) by simply treating n as a continuous value. Thus, without any inconsistency, the physical depth for $\hat{D}(\boldsymbol{p})$ can be described as $z_{\hat{D}(\boldsymbol{p})}$.

Using the refined depth map $\hat{D}(\boldsymbol{p})$, we can obtain the image from the target viewpoint, $I_{(t)}$, whose resolution is the same as those of the input images.

$$I_{(t)}(\boldsymbol{p}) = \frac{1}{M} \sum_{m} I_{(m)}(P_{t \to m}(\boldsymbol{p}, z_{\hat{D}(\boldsymbol{p})})) \tag{10}$$

This image is referred to as a *blend-based* image because the input images are blended together to produce it.

4 Super-Resolved Free-Viewpoint Image Synthesis

After the depth estimation from the target viewpoint, which was described in Section 3, we have a depth map $\hat{D}$, cost map $\hat{S}_{\min}$, and synthesized image $I(t)$, with the same resolutions as those of the input images. As the pre-process of super-resolution, the images are upsampled to the target resolution by using a standard interpolation method (in this work, bicubic interpolation), to obtain $\hat{D}_{\uparrow}$, $\hat{S}_{\min\uparrow}$, and $I(t)_{\uparrow}$. The super-resolved image from the target viewpoint is denoted as $I^{SR}_{(t)}$; the inference process is described in this section.

4.1 Formulation with Depth-Reliability-Based Regularization

Following the standard reconstruction-based SR scheme, the problem can be described by minimization of a energy function E_{sr} as

$$E_{sr}(I^{SR}_{(t)}) = E^{(1)}_{sr}(I_{(1)}, ..., I_{(M)} | I^{SR}_{(t)}) + \lambda\, E^{(2)}_{sr}(I^{SR}_{(t)}), \tag{11}$$

where $E^{(1)}_{sr}$ is a fidelity term, $E^{(2)}_{sr}$ is a regularizer, and λ is a positive weight.

The fidelity term evaluates the relation between the input images $I_{(m)}$ and the desired super-resolved image $I^{SR}_{(t)}$. We formulated it as

$$\mathrm{E}^{(1)}_{sr} = \sum_{m} \sum_{\boldsymbol{p} \in I_{(m)}} \| I_{(m)}(\boldsymbol{p}) - \hat{I}_{(m)}(\boldsymbol{p}) \|^2, \text{ where } \hat{I}_{(m)} = f_{t_{\uparrow} \to m}(I^{SR}_{(t)},\ \hat{D}_{\uparrow}). \tag{12}$$

In brief, the function $f_{t_{\uparrow} \to m}$ represents an image formation model. Using the given depth map $\hat{D}_{\uparrow}$, $f_{t_{\uparrow} \to m}$ transforms the latent image $I^{SR}_{(t)}$ into the m-th input image. The pixel correspondences between the two cameras, $t_{\uparrow}$ and m, are captured by $P_{t_{\uparrow} \to m}$, where $t_{\uparrow}$ means the target image has double the resolution. Occlusions and the point-spreading function are also considered in this transform (see appendix A.3 for more details).

The regularizer should reflect the prior knowledge about the resulting image $I^{SR}_{(t)}$, where we introduce two assumptions. First, $I^{SR}_{(t)}$ resembles the upsampled

version of the blend-based image $I_{(t)\uparrow}$. Second, the image formation model is less reliable where depth estimation is less accurate. On the basis of these assumptions, we define the regularization term as

$$E_{sr}^{(2)} = \sum_{\boldsymbol{p} \in I_{(t)}^{SR}} w(\boldsymbol{p}) \| I_{(t)}^{SR}(\boldsymbol{p}) - I_{(t)\uparrow}(\boldsymbol{p}) \|^2 \tag{13}$$

$$\text{where } w(\boldsymbol{p}) = \max\{ \| \hat{S}_{\min\uparrow}(\boldsymbol{p}) \|^4, w_{\min} \}. \tag{14}$$

Note that the second assumption is reflected in the pixel-wise weighting factor $w(\boldsymbol{p})$, which introduces adaptivity to the regularization. We observe that $\hat{S}_{\min}(\boldsymbol{p})$ takes large values around occlusion boundaries, for example, where the estimated depths are likely to be erroneous (see Fig. 4 (b)). Thus, for such regions, we increase the weight for the regularization term to stabilize the result. When the weight is ultimately large for a pixel $\boldsymbol{p}$, the result for $\boldsymbol{p}$ converges to the blend-based synthesis, i.e., $I_{(t)}^{SR}(\boldsymbol{p}) \sim I_{(t)\uparrow}(\boldsymbol{p})$. Meanwhile, for the regions where the depth estimation is sufficiently reliable, we decrease the weight for regularization to encourage the resolution enhancement that is enabled by the image formation model. This scheme, referred to as depth-reliability-based regularization, is very important in practice because depth information cannot be perfect. Moreover, this regularization is a natural extension of the conventional blend-based approach since it can adaptively combine blend-based synthesis with SR-based synthesis.

4.2 Implementation

Let X, $\bar{X}$, and Y_m be 1-D vector representations of $I_{(t)}^{SR}$, $I_{(t)\uparrow}$, and $I_{(m)}$, respectively. Let A_m be a matrix that represents the relation between the inputs and outputs of the function $f_{t_\uparrow \to m}$ in Eq. (12). Let W denote a diagonal matrix given by $\mathrm{diag}(w)$. Equation (11) can be rewritten as

$$E_{sr}(X) = \sum_m \| Y_m - A_m X \|^2 + \lambda \, W \| (X - \bar{X}) \|^2. \tag{15}$$

We set the initial value of X as $X_0 = \bar{X}$ and iterate

$$X_{j+1} = X_j - \alpha_j \nabla E_{sr}(X_j), \quad \alpha_j = \frac{\| \nabla E_{sr}(X_j) \|^2}{\nabla E_{sr}(X_j)^T (\nabla^2 E_{sr}) \nabla E_{sr}(X_j)} \tag{16}$$

until it converges, where $\nabla E_{sr}(X_j)$ denotes the gradient of E_{sr} at $X = X_j$. In our test, this solution is faster and more stable than solving a linear equation $\nabla E_{sr}(X_j) = 0$ using MATLAB's numerical solver.

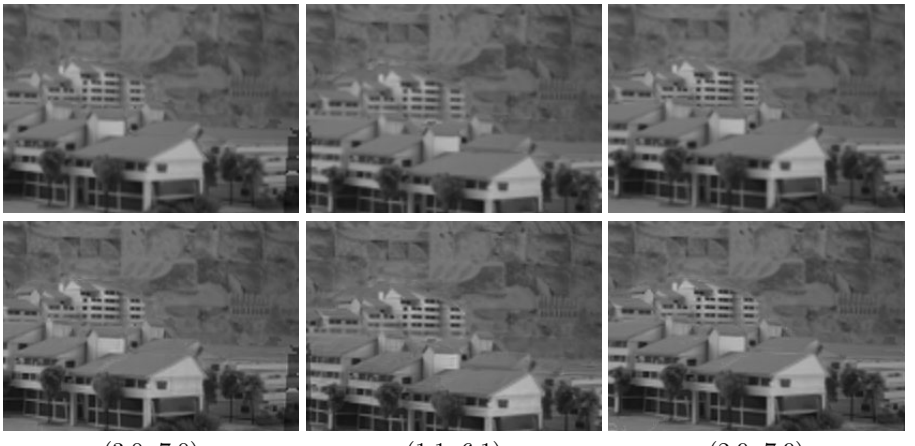
5 Experiments

The four images of a city diorama shown in Fig. 1, which were taken from the *Multi-view Image Database of University of Tsukuba, Japan*, were used as input

Table 1. Default values for parameters

Eq. (1)	$\lambda_1 = 100$, $\lambda_2 = 400$
Eq. (4)	$z_{\min}$=250 mm (21.00*), $z_{\max}$=1900 mm (2.76*), $N = 40$
Eq. (5)	size of $B_{\boldsymbol{p}}$: 3×3 pixels
Eq. (7)	$\mathrm{diff}_{\max} = 150$
Eq. (11)	$\lambda = 5.0 \times 10^{-13}$
Eq. (14)	$w_{\min} = 10$

∗corresponding disparities (in pixels) between input images



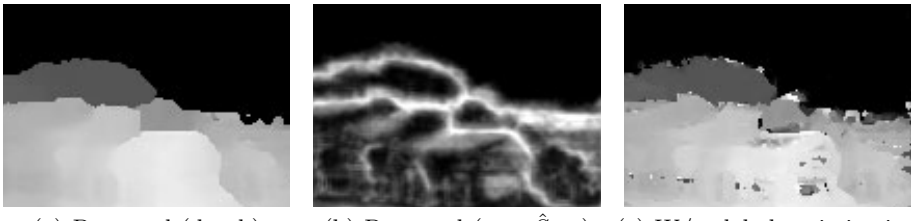
(3.0, 7.0) (1.1, 6.1) (2.9, 7.9)

Fig. 3. Resulting images from various viewpoints by (top) blend-based synthesis and (bottom) SR-based synthesis. A demo video is available from our website.

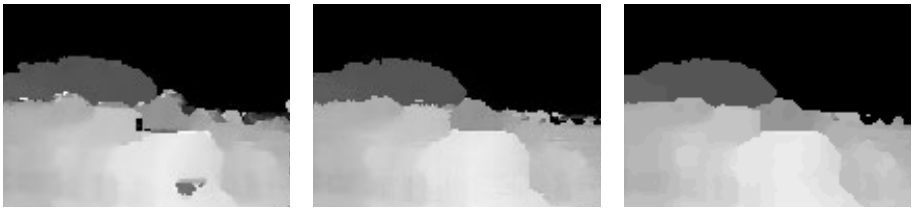
for our method. The input viewpoints were located at the corners of a square of 16 × 16 mm. The original images had 640 × 480 pixels in RGB color. We converted them to grayscale and reduced them to 160 × 120 pixels to use them as the input.

The target viewpoints were located inside the square formed by the input viewpoints. Our method first estimated a depth map with 160 × 120 pixels from a given target viewpoint, then generated a resulting image with 320 × 240 pixels from that viewpoint. The parameter settings, which were empirically determined based on several tests, are in Table 1.

Images from different viewpoints were generated by blend-based synthesis ($I_{(t)\uparrow}$) and SR-based synthesis ($I^{SR}_{(t)}$), shown in the top and bottom rows respectively of Fig. 3. The tuples of numbers below the images indicate the coordinates of the viewpoints according to the database notation, where the input viewpoints were described as (1, 6), (3, 6), (1, 8), and (3, 8). The figure shows that free-viewpoint images were successfully synthesized and that SR-based synthesis achieves better quality with finer texture details. A video for further demonstration is available from our website http://nae-lab.org/˜keita/.



(a) Proposed (depth) (b) Proposed (cost $\hat{S}_{\min}$) (c) W/o global optimization



(d) W/o occlusion handling (e) W/o block matching (f) W/o depth refinement

Fig. 4. Comparison of depth estimation results

5.1 Detailed Evaluation

To evaluate our method more closely, we fixed the target viewpoint to the center of the square, i.e., $(2, 7)$ according to the database notation, where the ground truth image was available from the database.

First, we evaluated the depth estimation part of our method. We disabled each element of our method one by one and estimated the depth. The results are shown in Fig. 4. As shown in (a), the proposed method produced a good result. The cost map $\hat{S}_{\min}$, shown in $1/10$ scale in (b), was used for the depth-reliability-based regularization mentioned later. When the global optimization was turned off by setting $\lambda_1, \lambda_2 = 0$, the resulting depth map was very noisy, as shown in (c). Unless the occlusions were handled properly, depth estimation was erroneous around the occlusion boundaries, as shown in (d). When the block matching was disabled by setting the block size to 1×1 pixels, the depth map became granular, as shown in (e). When depth refinement was skipped, the depth map took only the quantized values, as shown in (f).

Next, we evaluated the adaptive regularization scheme in SR-based synthesis, which is represented by Eq. (14). Images synthesized with different regularization factors (λ in Eq. (11)) are shown in Fig. 5. The top row shows the results with adaptive regularization, and the bottom shows those without it, where $w(\boldsymbol{p})$ was fixed to 2000 for all pixels. When λ became larger (meaning stronger regularization), the resulting images by SR-based synthesis converged to $I_{(t)\uparrow}$ in both cases. Meanwhile, when λ became smaller (meaning weaker regularization), the resulting images were sharper, but some regions, such as occlusion boundaries, became noisy due to mis-registrations. By our regularization scheme, the resulting quality was successfully optimized around $\lambda = 5.0 \times 10^{-13}$ because the regions with less reliable depth are more strongly regularized. Without this

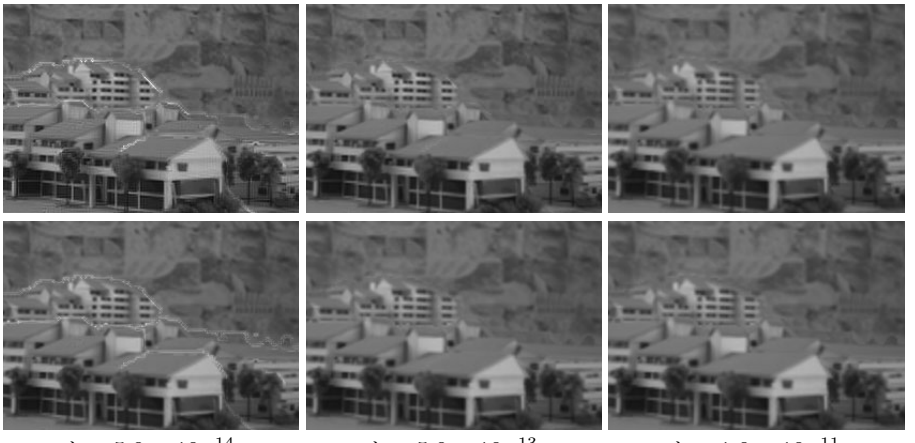


Fig. 5. Resulting images with (top) and without (bottom) adaptive regularization based on pixel-wise depth reliabilities

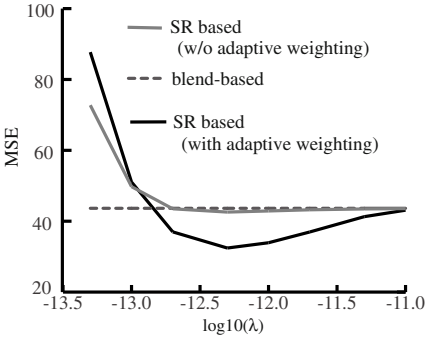


Fig. 6. Regularization factor vs. quality

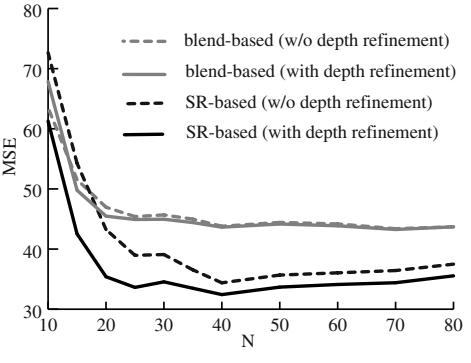


Fig. 7. Number of depths vs. quality

adaptive regularization we cannot obtain good results with any value of λ. The same results are shown quantitatively in Fig. 6. The horizontal axis denotes the value of λ in log scale, and the vertical axis is the mean squared error (MSE) against the ground truth image. The dashed line represents the quality of blend-based synthesis. The SR-based synthesis successfully improved the quality (reduced the MSE) if and only if the adaptive regularization was enabled.

Finally, we evaluated the performance change with regard to the number of candidate depths (N in Eq. (4)) and depth refinements (Eqs. (8) and (9)). The graph in Fig. 7 shows the relation between the number of candidate depths and the resulting image quality in MSE. As an overall trend, the quality improved as the number of depths increased, but using more than 40 depths had no benefit in our environment. As clearly seen in the graph, SR-based synthesis performed better than blend-based synthesis with a sufficient number of depths. Moreover,

depth refinement was effective for improving the quality, especially when it was combined with SR-based synthesis.

6 Conclusion

We proposed a method for free-viewpoint image synthesis with resolution improvement. The main features of our method are its view-dependent approach focused on a given target viewpoint, fast and accurate semi-global depth estimation, and super-resolution-based synthesis with depth-reliability-based regularization. Experimental results validated the effectiveness of our method. Future work will be focused on its real-time implementation. Our current implementation with unoptimized MATLAB codes performs at an unsatisfactory speed. We plan to transplant it to C++ and CUDA codes to improve the processing rate.

Acknowledgments. This research is supported by the Strategic Information and Communication R&D Promotion Programs (SCOPE) of the Ministry of Internal Affairs and Communications, Japan.

A Appendix

A.1 Derivation of Mapping Function

Here we show how to derive the point correspondence between two cameras α and β with a known depth z. Let $P_{(\alpha)}$ be the 3×4 projection matrix of the camera α. An object point $\boldsymbol{X}$ is projected onto an image point $\boldsymbol{u}_\alpha$ as

$$\boldsymbol{p}_{(\alpha)} = P_{(\alpha)}\boldsymbol{X}, \text{ where } \boldsymbol{p}_{(\alpha)} = (u_\alpha, v_\alpha, 1)^t, \boldsymbol{X} = (X, Y, Z, 1)^t. \tag{17}$$

A plane located at $Z = z$ can be written as

$$[0, 0, 1, -z] \cdot \boldsymbol{X} = 0. \tag{18}$$

By combining Eqs. (17) and (18), we obtain

$$\left(\begin{array}{c} \boldsymbol{p}_{(\alpha)} \\ \hline 0 \end{array}\right) = \hat{P}_{(\alpha)}\boldsymbol{X}, \text{ where } \hat{P}_{(\alpha)} = \left(\begin{array}{c} P_{(\alpha)} \\ \hline 0\ 0\ 1\ -z \end{array}\right). \tag{19}$$

Similarly, we can also derive $\hat{P}_{(\beta)}$ for the camera β. By using them, we obtain the point correspondence between the two cameras as

$$\left(\begin{array}{c} \boldsymbol{p}_{(\beta)} \\ \hline 0 \end{array}\right) = \hat{P}_{(\beta)}\hat{P}^{-1}_{(\alpha)} \left(\begin{array}{c} \boldsymbol{p}_{(\alpha)} \\ \hline 0 \end{array}\right). \tag{20}$$

This is equivalent to the mapping function $P_{\alpha\to\beta}(\boldsymbol{u}_{\boldsymbol{\alpha}}, z)$ in Eq. (6).

A.2 Derivation of Depth Refinement Procedure

Assume a parabolic function, $y = ax^2 + bx + c$, to locally approximate the energy function $S(\boldsymbol{p}, n)$ around $n = D(\boldsymbol{p})$. We substitute three points, $(x, y) = (D(\boldsymbol{p}),\ S_{\min}(\boldsymbol{p}))$, $(D(\boldsymbol{p}) - 1,\ S^{pre}_{\min}(\boldsymbol{p}))$, and $(D(\boldsymbol{p}) + 1,\ S^{next}_{\min}(\boldsymbol{p}))$, to obtain the coefficients a, b, and c. This function clearly takes the minimum $c - b^2/4a$ at $x = -b/2a$, which are equivalent to Eqs. (8) and (9), respectively.

A.3 Derivation of warping function

A pseudo-code of the function $f_{t_\uparrow \to m}$ in Eq. (12) is given as follows.

```
00:  function I'_(m) = f_{t↑→m}(I^SR_(t), D̂↑)
01:
02:  for each m
03:      D_(m) = depth_warping(D̂↑, t↑ → m)
04:  end
05:
06:  I'_(m)(p) = 0 for all p ∈ I'_(m)
07:  for each p ∈ I^SR_t
08:      p_(m) = P_{t↑→m}(p, z_{D̂↑(p)})
09:      get integer pixel positions p_(m),i (i = 1, 2, ...) around p_(m)
10:      for each p_(m),i
11:          if ||D_(m)(p_(m),i) − D̂↑(p)| ≤ 1
12:              get r_i based on |p_(m) − p_(m),i| and PSF
13:              I'_(m)(p_(m),i) = I'_(m)(p_(m),i) + r_i I^SR_(t)(p)
14:          end
15:      end
16:  end
```

In lines 02–04, depth maps viewed from input viewpoints $D_{(m)}$ are obtained by warping $\hat{D}_\uparrow$ to the input viewpoints, whose details are given later. In line 06, all pixels of $I'_{(m)}$ are initialized with zero. In line 08, a pixel on $I^{SR}_{(t)}$, $\boldsymbol{p}$, is warped onto the m-th input camera, resulting in $\boldsymbol{p}_{(m)}$. Since $\boldsymbol{p}_{(m)}$ is not an integer pixel position in general, neighboring integer pixels $\boldsymbol{p}_{(m),i}$ are selected. After the occlusion test in line 11, we determine the contribution weight r_i in line 12. This weight is calculated from the distance between $\boldsymbol{p}_{(m)}$ and $\boldsymbol{p}_{(m),i}$, and the shape of the point spreading function (PSF). We use a box-shaped PSF that is equal to the pixel in size. In line 13, $I^{SR}_{(t)}(\boldsymbol{p})$ is weighted by r_i and added to $I'_{(m)}(\boldsymbol{p}_{(m),i})$. These procedures are iterated for every pixel $\boldsymbol{p} \in I^{SR}_t$.

The function depth_warping() is given as follows. In line 02, each pixel is initialized with 0, which corresponds to the infinite distance. In line 04, each pixel on $\hat{D}_\uparrow$ is warped to the m-th input viewpoint. Depth values of $D_{(m)}$ are updated with the occlusion test as shown by lines 05–07.

00: function $D_{(m)} = \text{depth\_warping}(\hat{D}_{\uparrow},\ t_{\uparrow} \rightarrow m)$
01:
02: $D_{(m)}(\boldsymbol{p}) = 0$ for all $\boldsymbol{p} \in D_{(m)}$
03: for each $\boldsymbol{p} \in I_t^{SR}$
04: $\boldsymbol{p}_{(m)} = \text{round}(P_{t_{\uparrow} \rightarrow m}(\boldsymbol{p},\ z_{\hat{D}_{\uparrow}(\boldsymbol{p})}))$
05: if $D_{(m)}(\boldsymbol{p}_{(m)}) \leq \hat{D}_{\uparrow}(\boldsymbol{p})$
06: $D_{(m)}(\boldsymbol{p}_{(m)}) = \hat{D}_{\uparrow}(\boldsymbol{p})$
07: end
08: end

References

1. Kubota, A., et al.: Multiview Imaging and 3DTV. IEEE Signal Processing Magazine 24(6), 10–111 (2007)
2. Taguchi, Y., Koike, T., Takahashi, K., Naemura, T.: TransCAIP: A Live 3D TV System Using a Camera Array and an Integral Photography Display with Interactive Control of Viewing Parameters. IEEE Trans. Visualization and Computer Graphics 15(5), 841–852 (2009)
3. Park, S.-C., et al.: Super-resolution Image Reconstruction: A Technical Overview. IEEE Signal Processing Magazine 20(3), 21–36 (2003)
4. Matusik, W., Buehler, C., Raskar, R., Gortler, S.-J., McMillan, L.: Image-Based Visual Hulls. In: Proc. ACM SIGGRAPH, pp. 369–374 (2000)
5. Yang, R., Welch, G., Bishop, G.: Real-Time Consensus-Based Scene Reconstruction Using Commodity Graphics Hardware. In: Proceedings of Pacific Graphics, pp. 225–235 (2002)
6. Hirschmueller, H.: Accurate and Efficient Stereo Processing by Semi-Global Matching and Mutual Information. In: IEEE CVPR, pp. 807–814 (2005)
7. Hirschmuller, H.: Stereo Processing by Semiglobal Matching and Mutual Information. In: IEEE TPAMI, vol. 30(2), pp. 328–341 (2008)
8. Tung, T., Nobuhara, S., Matsuyama, T.: Simultaneous Super-Resolution and 3D Video Using Graph-Cuts. In: IEEE CVPR, pp. 1–8 (2008)
9. Goldluecke, B., Cremers, D.: Superresolution Texture Maps for Multiview Reconstruction. In: IEEE ICCV, pp. 1677–1684 (2009)
10. Mudenagudi, U., Gupta, A., Goel, L., Kushal, A., Kalra, P., Banerjee, S.: Super Resolution of Images of 3D Scenecs. In: Yagi, Y., Kang, S.B., Kweon, I.S., Zha, H. (eds.) ACCV 2007, Part II. LNCS, vol. 4844, pp. 85–95. Springer, Heidelberg (2007)
11. Takahashi, K., Ishii, M., Naemura, T.: Super-Resolution Plane Sweeping for Free-Viewpoint Image Synthesis. In: IEEE ICIP, pp. 2013–2016 (2011)
12. `http://vision.middlebury.edu/stereo/`