SUPER-RESOLUTION PLANE SWEEPING FOR FREE-VIEWPOINT IMAGE SYNTHESIS

Keita TAKAHASHI † *Masato ISHII* ‡ *Takeshi NAEMURA* †

† The University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo, 113-8656, Japan
‡ NEC Corporation, Shimonumabe 1753, Nakahara-ku, Kawasaki-shi, Kanagawa, 211-8666, Japan

ABSTRACT

Free-viewpoint image synthesis (FVIS) refers to the process of generating novel viewpoint images from a set of multi-view images. Most of the conventional FVIS methods were based on image blending, so that they are subject to a fundamental limitation in resolution: the output resolution is lower than or at most equal to that of the input images. A reasonable approach to overcome this limitation is to replace image blending with reconstruction-based super-resolution. Following this idea, we propose a new FVIS method named as *super-resolution plane sweeping* by extending general plane sweeping methods. We also propose an adaptive weighting scheme to make super-resolution reconstruction operate only on the pixels where it improve the quality. Experimental results with real images are presented to show the effectiveness of our method.

Index Terms— Free-viewpoint image, Super-resolution, 3-D imaging

1. INTRODUCTION

Free-viewpoint image synthesis (FVIS) is the process of combining multiple images from different viewpoints to generate new images from arbitrary viewpoints where no camera was located actually. This technology has attracted much research interest recently [1, 2], because it has a great potential in providing realistic 3-D visual experiences, which are desired for telecommunication and broadcasting with sufficient bandwidths in the near future.

Typical FVIS methods consist of two steps: first, some shape or depth model is estimated from the input images, and then, the input images are blended together to paint the model, which can finally be seen from arbitrary viewpoints. However, due to the nature of the blending operation, this framework has a fundamental limitation in the resulting resolution; the resolution of the synthesized images is lower than or at most equal to that of the input images.

To overcome this limitation, a new framework of FVIS that does not depend only on image blending is necessary. Especially, it is reasonable to incorporate reconstruction-based super-resolution methods [3] with the FVIS framework, because we have multi-view images as the input. In this paper, we propose a new FVIS algorithm named as *super-resolution plane sweeping* that can increase the resulting resolution by extending general plane sweeping methods [4, 5, 6]. The success of our method relies on an adaptive weighting scheme that makes super-resolution reconstruction operate only on the pixels where it improve the quality.

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

1.1. Background

Reconstruction-based super-resolution is the process of estimating an underlying high-resolution image from multiple low-resolution images of the same scene [3]. The process is formulated as the inverse problem of an image formation model, which describes the relation between the high and low resolution images, with some priors for regularization. However, in general, this technology is not intended to generate images from new viewpoints. Usually, the viewpoint of the resulting high-resolution image is selected from those of low-resolution input images.

As mentioned earlier, traditional FVIS methods were based on image blending [1, 2], so that they are subject to the fundamental limitation in resolution. Recently, several researchers have applied reconstruction-based super-resolution technology to the FVIS problem to achieve high-resolution view synthesis [7, 8]. These methods are object-oriented, but not designed to achieve full-frame super-resolution, because they perform silhouette-based surface reconstruction before or during the process of super-resolution.

In contrast to the prior works [7, 8], our method aims to achieve full-frame super-resolution of free-viewpoint images. The basic idea is to apply super-resolution reconstruction to each of the depth planes in the plane sweeping framework [4, 5, 6]. We also propose an adaptive weighting scheme, which controls the strength of the image formation model for each pixel according to the relevance of the depth assumption. Thanks to this weighting scheme, our method can naturally extend the standard plane sweeping algorithm and can improve the resolution even if the depth accuracy is not improved.

2. METHOD

We first describe a general plane-sweeping algorithm as the baseline method in Section 2.1, followed by our *super-resolution plane sweeping* algorithm in Section 2.2.

In this paper, images are represented as 1-D vectors and denoted by bold-face lower-case letters, e.g. $\mathbf{x}$, where n-th element (pixel) of $\mathbf{x}$ is written as $\mathbf{x}(n)$. Image operations, such as projection, down/up-sampling, are represented as 2-D matrices and denoted by bold-face upper-case letters, e.g. $\mathbf{P}$. In all of the image operations, bicubic interpolation is adopted to interpolate pixels.

2.1. Baseline Algorithm of Plane Sweeping

Figure 1 illustrates the configuration. Let $\mathbf{y}_k$ (k=1,..,4) be the input images from different viewpoints and $\mathbf{x}$ be the image to synthesize from a new viewpoint. A set of planes, referred to as the depth planes, are located in the scene to simultaneously perform depth evaluation and image synthesis. The depth planes are usually evaluated in serial order (near-to-far order, typically), thereby, the entire process is often called as plane sweeping. Here, we describe

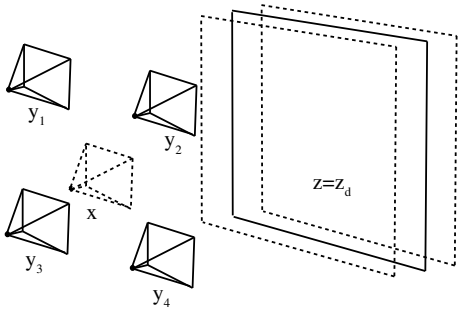

Fig. 1. Configuration.

a general baseline algorithm of plane sweeping based on [4, 5, 6], which can be easily extended to more general configurations.

Let z_d be the location of the depth plane where d is the index of depth values. At each depth, we calculate

$$\mathbf{x}_d = \sum_k \mathrm{diag}(\mathbf{a}_k)\mathbf{P}^{\mathbf{y}_k \to \mathbf{x}}_{z=z_d}\mathbf{y}_k \tag{1}$$

$$\mathbf{m}_d = \sum_{k,k'} \mathrm{match}(\mathbf{P}^{\mathbf{y}_k \to \mathbf{x}}_{z=z_d}\mathbf{y}_k, \mathbf{P}^{\mathbf{y}_{k'} \to \mathbf{x}}_{z=z_d}\mathbf{y}_{k'}). \tag{2}$$

In the above, $\mathbf{P}^{\mathbf{y}_k \to \mathbf{x}}_{z=z_d}$ represents the projection from $\mathbf{y}_k$ to $\mathbf{x}$ via the depth plane at $z = z_d$. Consequently, $\mathbf{P}^{\mathbf{y}_k \to \mathbf{x}}_{z=z_d}\mathbf{y}_k$ means that $\mathbf{y}_k$ is back-projected onto the depth plane and seen from the viewpoint of $\mathbf{x}$. We assume $\mathbf{P}^{\mathbf{y}_k \to \mathbf{x}}_{z=z_d}$ is a square matrix so that the length of a vector (the number of pixels) is not changed by the projection. Let $\mathbf{a}_k$ be a vector of weighting coefficients satisfying $\sum_k \mathbf{a}_k(n) = 1,\ \forall n$. The function "match" evaluates the consistency between the input images at each pixel position where any evaluation criterion can be used. In this work, we adopt squared differences and apply shiftable window aggregation (7×7 pixels) with normalization by the texture edge intensity.

In a physical sense, $\mathbf{x}_d$ is the image from the target viewpoint under the assumption that *the scene is exactly located on the depth plane at* $z = z_d$, and $\mathbf{m}_d$ has pixel-wise evaluation values of how much this assumption is correct (the smaller, the better). As shown in Fig. 2, only the regions of the scene that are near to the depth plane are clearly visible (as if they are *focused* [6]) in $\mathbf{x}_d$, and the corresponding regions have small values in $\mathbf{m}_d$.

Given $\mathbf{x}_d$ and $\mathbf{m}_d$ for all of the depths, the final resulting image from the target viewpoint is obtained by their integration. The simplest way is to apply the winner-takes-all rule over the depth indices d for each pixel as

$$\mathbf{x}(n) = \mathbf{x}_{\mathbf{d}(n)}(n), \quad \text{where} \quad \mathbf{d}(n) = \arg\min_d \mathbf{m}_d(n) \tag{3}$$

where $\mathbf{d}$ can be regarded as a rough depth map, and each pixel of $\mathbf{x}$ is taken from the most relevant $\mathbf{x}_d$ based on the estimated depth map $\mathbf{d}$. Examples of $\mathbf{x}$ and $\mathbf{d}$ are shown in Fig. 3.

2.2. Super-Resolution Plane Sweeping

The baseline plane sweeping algorithm only achieves lower or at most equivalent resolution compared to the input images, because, as shown in Eq. (1), the input images are blended together to generate the target image. In this subsection, we modify the baseline algorithm to achieve higher resolution. Specifically, a reconstruction-based super-resolution scheme is incorporated into the image synthesis process at each depth, by replacing Eq. (1) with Eq. (4). Our algorithm is named as *super-resolution plane sweeping*.

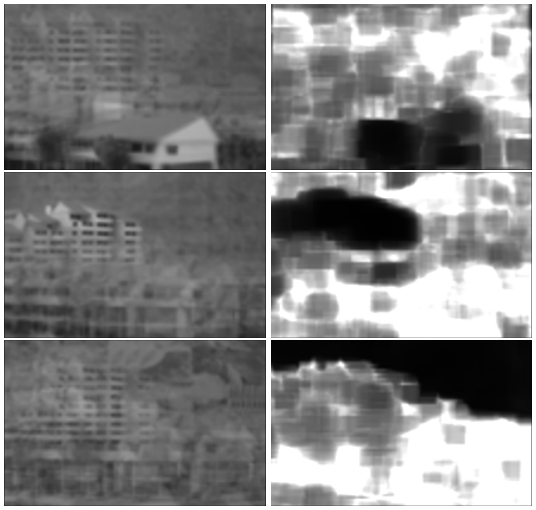

Fig. 2. Examples of $\mathbf{x}_d$ (left column) and $\mathbf{m}_d$ (right colomn) with different depths. Best viewed on the screen.

Fig. 3. Examples of $\mathbf{d}$ (left) and $\mathbf{x}$ (right). Best viewed on the screen.

Let r denote the magnification factor. At each depth, the latent high-resolution image, $\mathbf{x}_d$, whose size is $r \times r$ times of the input images y_k, is determined by the following energy minimization:

$$\mathbf{x}_d = \arg\min_{\mathbf{x}_d} \left\{ \sum_k \mathrm{Fidelity}(\mathbf{x}_d, \mathbf{y}_k) + \lambda \cdot \mathrm{Prior}(\mathbf{x}_d) \right\} \tag{4}$$

where λ is a positive constant to coordinate the relative strengths of the two terms, and was set to 0.001 in this paper.

The fidelity term evaluates how much the solution is suited to the image formation model:

$$\mathrm{Fidelity}(\mathbf{x}_d, \mathbf{y}_k) = \mathbf{e}^T_{k,d}\ \mathrm{diag}(\mathbf{P}^{\mathbf{x} \to \mathbf{y}_k}_{z=z_d}\mathbf{w}_{k,d})\ \mathbf{e}_{k,d} \tag{5}$$

$$\text{where} \quad \mathbf{e}_{k,d} = \mathbf{y}_k - \mathbf{D}_{1/r^2}\mathbf{P}^{\mathbf{x} \to \mathbf{y}_k}_{z=z_d}\mathbf{x}_d. \tag{6}$$

Here, $\mathbf{e}_{k,d}$ is the difference between the observation ($\mathbf{y}_k$) and the image formation model ($\mathbf{D}_{1/r^2}\mathbf{P}^{\mathbf{x} \to \mathbf{y}_k}_{z=z_d}\mathbf{x}_d$), where $\mathbf{x}_d$ is the latent high-resolution image from the target viewpoint, $\mathbf{P}^{\mathbf{x} \to \mathbf{y}_k}_{z=z_d}$ is the projection from $\mathbf{x}$ onto $\mathbf{y}_k$ via the depth plane at $z = z_d$, and $\mathbf{D}_{1/r^2}$ is a down-sampling matrix with the factor r. The point spreading function is not explicitly described, but is absorbed in $\mathbf{D}_{1/r^2}\mathbf{P}^{\mathbf{x} \to \mathbf{y}_k}_{z=z_d}$. $\mathbf{w}_{k,d}$ represents pixel-wise weighting coefficients defined for each $\mathbf{y}_k$, whose definition and meaning are discussed below.

The image formation model, $\mathbf{D}_{1/r^2}\mathbf{P}^{\mathbf{x} \to \mathbf{y}_k}_{z=z_d}\mathbf{x}_d$, holds true *if and only if the scene is located on the depth plane*. Thereby, the strength of the fidelity term should be coordinated for each pixel according to the relevance of the depth assumption; otherwise, physically-incorrect models are applied to $\mathbf{x}_d$, which produces undesired re-

sults. For this purpose, we introduce weighting coefficients, whose vector representation $\mathbf{w}_{k,d}$ is given by:

$$\mathbf{w}_{k,d} = \sum_{k' \neq k} \frac{1}{1 + \alpha \cdot \mathrm{match}(\mathbf{P}_{z=z_d}^{\mathbf{y}_k \to \mathbf{x}} \mathbf{y}_k, \mathbf{P}_{z=z_d}^{\mathbf{y}_{k'} \to \mathbf{x}} \mathbf{y}_{k'})} \tag{7}$$

where α is a positive constant. The function "match" is the same with that of Eq. (2), except that normalization is not performed here. For the regions that are located on the depth plane at z_d, the matching function returns small values, resulting in large weighting coefficients. Meanwhile, for the regions that are apart from the depth plane, the weighting coefficients would be smaller. In this way, $\mathbf{w}_{k,d}$ changes the strength of the fidelity term for each pixel according to the relevance of the depth assumption. This scheme is referred to as adaptive weighting.

The remaining problem is how to define the prior terms. Our purpose is natural extension of the plane sweeping algorithm. The design scheme here is that only the regions that are located on the depth plane at z_d should be super-resolved, but other regions, i.e. the regions that are apart from the depth plane, should remain unchanged from the result of the baseline plane sweeping algorithm. Consequently, we define the prior term as

$$\mathrm{Prior}(\mathbf{x}_d) = (\mathbf{x}_d - \mathbf{U}_{r^2} \tilde{\mathbf{x}}_d)^2 \tag{8}$$

where $\tilde{\mathbf{x}}_d$ is the result of Eq. (1) in the baseline plane sweeping algorithm, and $\mathbf{U}_{r^2}$ is an up-sampling matrix. This prior means that $\mathbf{x}_d$ should not be too far from $\mathbf{U}_{r^2} \tilde{\mathbf{x}}_d$. It can be easily checked that for the regions apart from the depth plane whose weighting coefficients are near to zero, minimization of Eq. (4) results in $\mathbf{x}_d \approx \mathbf{U}_{r^2} \tilde{\mathbf{x}}_d$.

Equations (2) and (3) in the baseline algorithm are common to our algorithm except that $\mathbf{m}_d$ is also up-sampled.

3. EXPERIMENTS

The test images were taken from miniature city dataset of "multi-view image database of University of Tsukuba, Japan" . The dataset consists of 81 images whose viewpoints are arranged in a 9×9 square grid[1]. As the input ($\mathbf{y}_k$), we used four images from the grid-points at $(0, 5)$, $(4, 5)$, $(4, 9)$, and $(0, 9)$. These images were down-sampled to 160×120 pixels beforehand. The output resolution was set to 320×240 pixels ($r = 2.0$), but 50 pixels from the boundaries were excluded from the evaluation. The viewpoint of the output image $\mathbf{x}$ was set to $(2, 7)$, where the ground truth image is available. We set $\mathbf{a}_k(n) = 0.25, \forall n, \forall k$ in Eq. (1), because the target viewpoint is located at the center of the four input viewpoints. The energy of Eq. (4) was minimized by gradient descent method. The software was implemented with MATLAB.

3.1. Synthesis with a single depth plane

We first demonstrate intermediate results of our algorithm.

The upper row (a–c) of Fig. 4 shows images synthesized with a single depth plane located at the farthest building. These images correspond to $\mathbf{x}_d$ in Eqs. (1) or (4). We compared (a) the baseline algorithm with bicubic up-sampling, (b) super-resolution with adaptive weighting ($\alpha = 2.0$ in Eq. (7)), and (c) super-resolution without adaptive weighting ($\alpha = 0$ in Eq. (7)). The lower row (d–g) of Fig. 4 shows weighting coefficients $\mathbf{w}_{k,d}$ for $k = 1, ..., 4$, respectively, which were used to synthesize (b).

In Fig. 4(a)–(c), only the farthest building is clearly synthesized, but other regions are blurred. This is what is expected because we place a single depth plane at the farthest building. However, as shown in (a), even the farthest building is still blurry when we use the baseline plane sweeping, due to the nature of image blending. The effectiveness of our method is obvious from Fig. 4(b); the farthest building, where the depth plane is located, is clearly super-resolved, but other parts remain unchanged from those of the baseline algorithm. In other words, super-resolution reconstruction operates only on the pixels where it improves the quality. However, if the adaptive weighting is turned off, as shown in Fig. 4(c), the fidelity term constraint is enforced even to the regions that are far from the depth plane, resulting in undesired ringing-like artifacts.

Furthermore, occlusions are properly handled by the adaptive weighting, because weighting coefficients are defined for each of the input images as Eq. (7). For example, the lower part of the farthest building is invisible from two camera due to occlusion, but visible from other two cameras. The weighting coefficients of this part for the former two cameras (Figs. 4(f) and 4(g)) are small, but those for the remaining two cameras (Figs. 4(d) and 4(e)) are still large. As a result, the lower part of the farthest building is blurry in Figs. 4(a) and 4(c), but it is clearly visible in Fig. 4(b) thanks to the proper handling of occlusions.

3.2. Result of plane sweeping

We set 20 depth planes in the scene according to the rule:

$$\frac{1}{z_d} = \frac{d - 1/2}{N_d} \left(\frac{1}{Z_{\min}} - \frac{1}{Z_{\max}} \right) \ (d = 1, 2, ..., N_d) \tag{9}$$

where $Z_{\max}$ and $Z_{\min}$ are the maximum and minimum depths of the scene, and integrated them using Eq. (3). The resulting images by the three methods described above are shown in Fig. 5. The PSNRs are 29.0 dB, 29.4 dB, and 27.5 dB for (a) the baseline algorithm with up-sampling, (b) super-resolution with adaptive weighting, and (c) super-resolution without adaptive weighting, respectively. (d) is the ground truth image for reference. Although the differences in PSNR values are relatively small, the differences in visual quality are obvious. The proposed method (Fig. 5(b)) achieves the best quality among the three, with improved resolution throughout the scene. Note that the underlying depth map, $\mathbf{d}$ in Eq. (3), is common to all methods, because we did not modify the plane integration process.

If the adaptive weighting is turned-off, severe ringing artifacts arise as in Fig. 5(c). This is mainly due to the imperfection of depth estimation. In the synthesis process for each depth plane, most ringing artifacts appear on the regions that are apart from the depth plane, as shown in Fig. 4(c). If the estimated depth were definitely correct, such regions would completely be discarded in the layer integration process of Eq. (3). However, in many applications, depth accuracy cannot absolutely be guaranteed, and some regions with ringing artifacts would remain after the layer integration process, as in Fig. 5(c). Meanwhile, thanks to the adaptive weighting scheme, our method does not produce ringing artifacts. Thereby, our method has less damage from the imperfection of depth estimation.

The main drawback of our method is insufficiency of occlusion handling. As mentioned in Section 3.1, our adaptive weighting scheme (Eq. (7)) is occlusion-sensitive. However, the depth evaluation (Eq. (2)) and layer-integration (Eq. (3)) procedures does not properly handle occlusions, causing visible artifacts around the depth discontinuities in Fig. 5 (b).

[1]The original images are in 24-bit RGB color with 640 × 480 pixels. They were converted into grayscale and down-sampled with bicubic kernel for the use in our experiments.

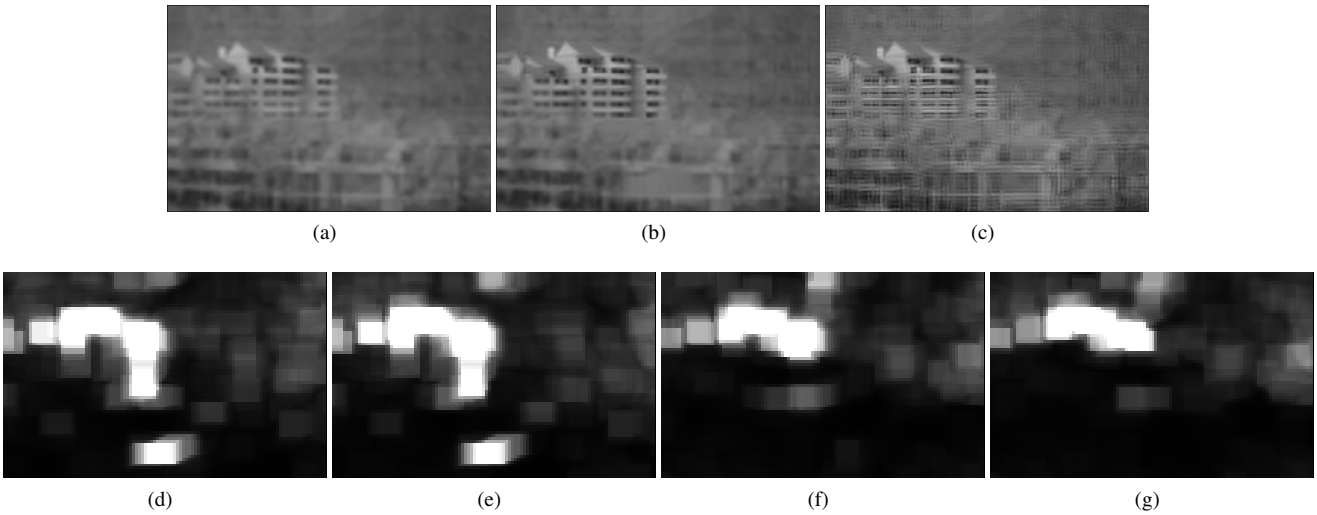

Fig. 4. Synthesis with a single depth plane: (a) the baseline algorithm with up-sampling, (b) super-resolution with adaptive weighting, and (c) super-resolution without adaptive weighting. (d)–(g) weighting coefficients for $\mathbf{y}_1$–$\mathbf{y}_4$. Best viewed on the screen.

Fig. 5. Results of plane sweeping with 20 depths: (a) the baseline algorithm with up-sampling (29.0 dB), (b) super-resolution with adaptive weighting (29.4 dB), (c) super-resolution without adaptive weighting (27.5 dB), and (d) the ground truth. Best viewed on the screen.

4. CONCLUSION

We proposed a new free-viewpoint image synthesis (FVIS) method named as *super-resolution plane sweeping*. The basic idea is to apply reconstruction-based super-resolution to each of the depth planes in the plane sweeping algorithm. The key contribution is the adaptive weighting that can control the strength of super-resolution process according to the relevance of the depth assumption for each pixel. The experimental results demonstrated that our method can successfully improve the resolution compared to the conventional approach based on image blending.

Our future work includes several directions. Our algorithm will be extended to more general configurations, where more than four input images can be used, and new viewpoints can be located anywhere, not limited on the same plane with the input viewpoints. The plane integration process should also be improved to handle occlusions properly. Another interesting topic is real-time implementation of our algorithm using GPGPU technology.

5. REFERENCES

[1] H.-Y. Shum, S.-B. Kang, and S.-C. Chan, "Survey of image-based representation and compression techniques," *IEEE Trans. on CSVT*, vol. 13, no. 11, pp. 1020–1037, 2003.

[2] A. Kubuta, A. Smolic, M. Magnor, M. Tanimoto, T. Chen, and C. Zhang, "Multiview imaging and 3dtv: Special issue overview and introduction," *IEEE Signal Processing Magazine*, vol. 24, no. 6, pp. 10–21, Nov 2007.

[3] S.-C. Park, M.-K. Park, and M.-G. Kang, "Super-resolution image reconstruction: a technical overview," *IEEE Signal Processing Magazine*, vol. 20, no. 3, pp. 21–36, May 2003.

[4] R.-T. Collins, "A space-sweep approach to true multi-image matching," *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 358–363, 1996.

[5] R. Yang, G. Welch, and G. Bishop, "Real-time consensus-based scene reconstruction using commodity graphics hardware," *10th Pacific Conference on Computer Graphics and Applications*, pp. 225–234, 2002.

[6] K. Takahashi and T. Naemura, "Layered light-field rendering with focus measurement," *EURASIP Signal Processing: Image Communication*, vol. 21, no. 6, pp. 519–530, 2006.

[7] T. Tung, S. Nobuhara, and T. Matsuyama, "Simultaneous super-resolution and 3d videousing graph-cuts," *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.

[8] B. Goldluecke and D. Cremers, "Superresolution texture maps for multiview reconstruction," *IEEE Intl. Conf. on Computer Vision*, pp. 1677–1684, 2009.