Theory of Optimal View Interpolation with Depth Inaccuracy

Keita Takahashi*

The University of Tokyo, IRT Research Initiative, Hongo 7-3-1,
Bunkyo-ku, 113-8656 Tokyo, Japan
keita.takahashi@ieee.org

Abstract. Depth inaccuracy greatly affects the quality of free-viewpoint image synthesis. A theoretical framework for a simplified view interpolation setup to quantitatively analyze the effect of depth inaccuracy and provide a principled optimization scheme based on the mean squared error metric is proposed. The theory clarifies that if the probabilistic distribution of disparity errors is available, optimal view interpolation that outperforms conventional linear interpolation can be achieved. It is also revealed that under specific conditions, the optimal interpolation converges to linear interpolation. Furthermore, appropriate band-limitation combined with linear interpolation is also discussed, leading to an easy algorithm that achieves near-optimal quality. Experimental results using real scenes are also presented to confirm this theory.

1 Introduction

Free-viewpoint image synthesis is the process of combining a set of multi-view images to generate an image which can be seen from a new viewpoint where no camera was actually located [11,16,6]. The quality of free-viewpoint images is greatly affected by the inaccuracies of camera calibration, geometry estimation, and other precedent procedures. However, to our knowledge, there are few studies that rigorously analyze the quantitative quality of resulting images in the presence of such inaccuracies.

This paper presents a theoretical framework for the free-viewpoint image synthesis problem to quantitatively analyze the effect of depth inaccuracy and provide a principled optimization scheme based on the mean squared error (MSE) metric. For simplicity of analysis, the scope of discussion is limited to a fundamental view interpolation setup where the new viewpoint is located between two given input views. I first formulate the relation between the accuracy of depth and the resulting quality of view interpolation. Based on the formulation, I then derive an optimal view interpolation scheme that minimizes the mean squared error (MSE) of the synthesized image, and discuss an appropriate band-limitation combined with conventional linear interpolation. Furthermore, I reveal that the

* This research was supported by National Institute of Information and Communication Technology (NICT).

K. Daniilidis, P. Maragos, N. Paragios (Eds.): ECCV 2010, Part IV, LNCS 6314, pp. 340–353, 2010.
© Springer-Verlag Berlin Heidelberg 2010

optimal interpolation converges to linear interpolation under specific conditions. Experimental results using real scenes are presented to validate this theory.

A key finding of this work is that the optimal interpolation which outperforms conventional linear interpolation can be achieved if the probabilistic distribution of disparity errors is available. This work also gives a theoretical base for the use of linear interpolation as an approximation of the optimal interpolation, although in previous works linear interpolation was used heuristically without sufficient theoretical justification. In addition, the band-limitation scheme presented in this work achieves near-optimal quality just combined with linear interpolation.

2 Background

Depth inaccuracy is one of the main factors that degrades the quality of free-viewpoint image synthesis and view interpolation. It results from not only the fundamental limitation of geometry estimation methods but also practical reasons such as computational time and the rendering algorithm. For example, depth information is quantized into discrete values in layer-based rendering methods [10,14,12]. In communication systems, depth inaccuracy is induced by lossy data compression.

The tradeoff between depth inaccuracy and camera interval has been studied from the perspective of the sampling problem, resulting in the limiting condition for aliasing-free view interpolation referred to as minimum sampling curve [2,15,8]. If necessary, anti-alias filtering can be used to enforce the limiting condition. However, in practice, the quality of the interpolated view tends to degrade *continuously* as the depth inaccuracy increases, which cannot be captured using the yes-no-question of with/without aliasing artifacts. Furthermore, the anti-alias filter often worsens the result because it filters out some of the valid signal components with the aliasing component [13]. By contrast, I adopt the MSE instead of aliasing for the quality metric to overcome these limitations.

In the context of video or multi-view image compression, depth inaccuracy is quantitatively considered for analyzing coding performance [4,9]. This approach uses the Fourier domain analysis to derive the MSE of the inter-image prediction. It was applied to the view interpolation problem [13], but the discussion was limited to the case where the new viewpoint is located at the midpoint between the two given views. Inspired by those works, this research presents a more comprehensive framework for theoretical analysis of the view interpolation problem, resulting in a principled optimization scheme.

3 Theory

Figure 1 illustrates the configuration used throughout this paper. Let $v_L(x, y)$ and $v_R(x, y)$ be the input images located on the left and right in parallel, respectively. $v(x, y)$ is the target image I want to synthesize from these images by view interpolation. The target viewpoint internally divides the baseline connecting the left and right viewpoints into the ratio $r : 1 - r$, where $r \in [0, 1]$. Let

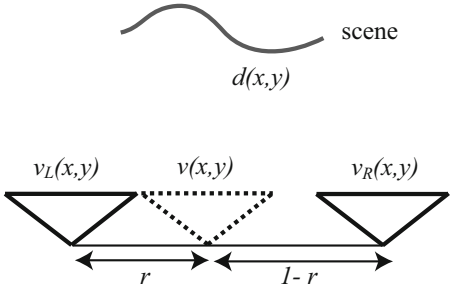


Fig. 1. Configuration

$d(x, y)$ be a pixel-wise depth map where (x, y) denotes pixel positions on the target image and the depth values are represented in terms of disparities (pixel shifts) between the left and right images. I assume that the depth information is provided with certain amount of errors, and analyze the effect of depth errors on the result of view interpolation. Depth estimation itself, i.e., how to estimate $d(x, y)$, is out of scope of this paper.

3.1 View Interpolation with Disparity Error

First, I define a model that describes the relation between the input and target images. Assume that the target scene is a diffusive surface without occlusions. Basically, the left and right images, $v_L(x, y)$ and $v_R(x, y)$, are associated with the target image $v(x, y)$ using the depth map $d(x, y)$. In addition, noise terms, denoted as $n_L(x, y)$ and $n_R(x, y)$, are introduced to compensate the incompleteness of the above assumption. Consequently, the model is described as:

$$\begin{aligned} v_L(x, y) &= v\left(x - rd(x, y), y\right) + n_L(x, y), \\ v_R(x, y) &= v\left(x + (1-r)d(x, y), y\right) + n_R(x, y). \end{aligned} \tag{1}$$

The noise terms include occlusions, non-diffusive reflections, and other components that are present in $v_L(x, y)$ or $v_R(x, y)$ but unpredictable from $v(x, y)$. They are referred to as *unpredictable noise components* in this paper.

The view interpolation process consists of two steps; disparity shifting is applied to the input images, and these images are combined linearly to produce the target image. These steps should be formulated by taking the presence of the depth inaccuracy into account. Let ξ be the disparity error considered as a probabilistic variable, and apply disparity shifting of $(d(x, y)+\xi)$ pixels to the input images to obtain $v'_L(x, y, \xi)$ and $v'_R(x, y, \xi)$ as follows:

$$\begin{aligned} v'_L(x, y, \xi) &= v_L\left(x + r\left(d(x, y)+\xi\right), y\right) &&= v\left(x + r\xi, y\right) + n'_L(x, y) \\ v'_R(x, y, \xi) &= v_R\left(x - (1-r)\left(d(x, y)+\xi\right), y\right) &&= v\left(x - (1-r)\xi, y\right) + n'_R(x, y) \end{aligned} \tag{2}$$

where $n'_L(x, y)$ and $n'_R(x, y)$ are disparity-shifted versions of $n_L(x, y)$ and $n_R(x, y)$, respectively. If $\xi = 0$, $v'_L(x, y, \xi)$ and $v'_R(x, y, \xi)$ are equal to $v(x, y)$, except the

unpredictable noise components. By combining these disparity-shifted images, the result of view interpolation is obtained:

$$\hat{v}(x,y,\xi) = f_L(x,y) \circ v'_L(x,y,\xi) + f_R(x,y) \circ v'_R(x,y,\xi), \tag{3}$$

where $\circ$ denotes convolution, and $f_L(x,y)$ and $f_R(x,y)$ are spatially invariant linear filters, which are referred to as *combining filters* in this paper. It should be noted that *the combining filters integrate the functions of weighting coefficients and prefilters*. This unified formalization leads to a comprehensive optimization scheme. The synthesized image is written as $\hat{v}(x,y,\xi)$ because it is an estimate of the true target image $v(x,y)$ with the disparity error ξ.

In many previous works [3,7,5,8,12][1], linear interpolation with respect to the distance from the target viewpoint

$$\hat{v}(x,y,\xi) = (1-r) \cdot v'_L(x,y,\xi) + r \cdot v'_R(x,y,\xi) \tag{4}$$

was adopted without sufficient theoretical justification. This interpolation is straightforward and often produces reasonable results, however, it is just a special case of (3) with

$$f_L(x,y) = (1-r) \cdot \delta(x,y),\ f_R(x,y) = r \cdot \delta(x,y) \tag{5}$$

where $\delta(x,y)$ is Kronecker's delta function. Here, $f_L(x,y)$ and $f_R(x,y)$ degenerate into weighting coefficients without the function of prefilters. By contrast, I seek truly optimal forms for $f_L(x,y)$ and $f_R(x,y)$ based on the minimization of the MSE of the resulting image. It is clarified that under specific conditions, the optimal interpolation converges to linear interpolation as a limit.

3.2 MSE of View Interpolation

In this subsection, I derive the MSE of view interpolation by taking the expectation of the estimation error over the probabilistic disparity error ξ.

The theory is constructed in the frequency domain to easily deal with pixel shifts and filtering convolutions. Substitution of (2) into (3) and Fourier transform over (x,y) yields

$$\begin{aligned}\hat{V}(\omega_x,\omega_y,\xi) = \{&F_L(\omega_x,\omega_y)e^{jr\xi\omega_x} + F_R(\omega_x,\omega_y)e^{j(r-1)\xi\omega_x}\}V(\omega_x,\omega_y)\\ &+F_L(\omega_x,\omega_y)N'_L(\omega_x,\omega_y) + F_R(\omega_x,\omega_y)N'_R(\omega_x,\omega_y),\end{aligned} \tag{6}$$

where ω_x and ω_y ($\omega_x, \omega_y \in [-\pi,\pi]$) are the angular frequencies of x and y, and j denotes the imaginary unit. $\hat{V}$, V, F_L, F_R, N'_L, and N'_R are the Fourier transforms of $\hat{v}$, v, f_L, f_R, n'_L, and n'_R, respectively. The estimation error from the ground truth $V(\omega_x,\omega_y)$ is

[1] Although those works consider more general configurations, their interpolation schemes are equivalent with (5) when the configuration is simplified into that in Fig. 1. Another choice is angular penalty [1], which becomes equivalent with (5) when the object is located far from the cameras.

$$E(\omega_x, \omega_y, \xi) = V(\omega_x, \omega_y) - \hat{V}(\omega_x, \omega_y, \xi). \tag{7}$$

The expectation of the squared error is obtained by

$$\Phi(\omega_x, \omega_y) = \int_{-\infty}^{\infty} p(\xi) \, \|E(\omega_x, \omega_y, \xi)\|^2 \, d\xi, \tag{8}$$

where $p(\xi)$ denotes the probability distribution of ξ. I assume $p(\xi)$ is constant over the image to simplify the discussion. Integration of $\Phi(\omega_x, \omega_y)$ over the entire spectra results in the MSE of view interpolation.[2]

$$\text{MSE} = \frac{1}{4\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \Phi(\omega_x, \omega_y) \, d\omega_x d\omega_y. \tag{9}$$

To further investigate the characteristics of $\Phi(\omega_x, \omega_y)$, (6) and (7) are substituted into (8), resulting in

$$\Phi(\omega_x, \omega_y) = G(\omega_x, \omega_y)\|V(\omega_x, \omega_y)\|^2 + \|\bar{N}(\omega_x, \omega_y)\|^2, \tag{10}$$

$$G() = 1 + \|F_L()\|^2 + \|F_R()\|^2 - F_L()\psi(r\omega_x) - F_L^*()\psi(-r\omega_x) - F_R()\psi((r-1)\omega_x)$$
$$-F_R^*()\psi((1-r)\omega_x) + F_L()F_R^*()\psi(\omega_x) + F_L^*()F_R()\psi(-\omega_x), \tag{11}$$

$$\|\bar{N}()\|^2 = \|F_L()N_L'()\|^2 + \|F_R()N_R'()\|^2. \tag{12}$$

In (11) and (12), (ω_x, ω_y) is abbreviated as (), and * denotes a complex conjugate. ψ is defined as

$$\psi(z) = \int_{-\infty}^{\infty} p(\xi) e^{j\xi z} d\xi. \tag{13}$$

It is assumed that $n_L'(x, y)$ and $n_R'(x, y)$ have no correlation with $v(x, y)$ and with each other, thereby cross spectra between them are ignored.

As shown by (10), $\Phi(\omega_x, \omega_y)$ consists of three terms. $\|V(\omega_x, \omega_y)\|^2$ denotes the power spectral density of the target image, $G(\omega_x, \omega_y)$ serves as a *gain* that characterizes the magnification factor of $\|V(\omega_x, \omega_y)\|^2$ to produce $\Phi(\omega_x, \omega_y)$, and $\|\bar{N}(\omega_x, \omega_y)\|^2$ summarizes the unpredictable noise components as (12). The disparity error, characterized by $p(\xi)$, affects the gain $G(\omega_x, \omega_y)$ via ψ, as can be seen from (11) and (13). The combining filters, $F_L(\omega_x, \omega_y)$ and $F_R(\omega_x, \omega_y)$, affect both $G(\omega_x, \omega_y)$ and $\|\bar{N}(\omega_x, \omega_y)\|^2$ and are the subjects of optimization, which is described next.

3.3 Derivation of Optimal View Interpolation

The essential idea of optimization is to determine such combining filters, $F_L(\omega_x, \omega_y)$ and $F_R(\omega_x, \omega_y)$, that minimize the MSE defined by (9). This is

[2] Parseval's formula proves that (9) is equivalent with the mean of squared errors calculated in the pixel (spatial) domain.

equivalent to minimizing $\Phi(\omega_x, \omega_y)$ for all frequencies. The optima, denoted as $\hat{F}_L(\omega_x, \omega_y)$ and $\hat{F}_R(\omega_x, \omega_y)$, should satisfy

$$\frac{\partial \Phi(\omega_x, \omega_y)}{\partial \hat{F}_L^*(\omega_x, \omega_y)} = 0, \ \frac{\partial \Phi(\omega_x, \omega_y)}{\partial \hat{F}_R^*(\omega_x, \omega_y)} = 0. \tag{14}$$

Substitution of (10)–(12) into (14) leads to simultaneous equations

$$\begin{pmatrix} 1+\theta_L & \psi(-\omega_x) \\ \psi(\omega_x) & 1+\theta_R \end{pmatrix} \begin{pmatrix} \hat{F}_L(\omega_x, \omega_y) \\ \hat{F}_R(\omega_x, \omega_y) \end{pmatrix} = \begin{pmatrix} \psi(-r\omega_x) \\ \psi((1-r)\omega_x) \end{pmatrix}, \tag{15}$$

$$\text{where} \quad \theta_L = \frac{\|N_L(\omega_x, \omega_y)\|^2}{\|V(\omega_x, \omega_y)\|^2}, \ \theta_R = \frac{\|N_R(\omega_x, \omega_y)\|^2}{\|V(\omega_x, \omega_y)\|^2} \tag{16}$$

whose unique solution is given as

$$\begin{aligned} \hat{F}_L(\omega_x, \omega_y) &= \frac{(1+\theta_R)\psi(-r\omega_x) - \psi(-\omega_x)\psi((1-r)\omega_x)}{(1+\theta_L)(1+\theta_R) - \psi(\omega_x)\psi(-\omega_x)} \\ \hat{F}_R(\omega_x, \omega_y) &= \frac{(1+\theta_L)\psi((1-r)\omega_x) - \psi(\omega_x)\psi(-r\omega_x)}{(1+\theta_L)(1+\theta_R) - \psi(\omega_x)\psi(-\omega_x)} \end{aligned} \tag{17}$$

I refer to $\hat{F}_L(\omega_x, \omega_y)$ and $\hat{F}_R(\omega_x, \omega_y)$ as *the optimal combining filters*, and view interpolation with those filters as *the optimal view interpolation*.

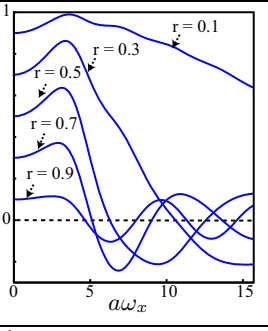
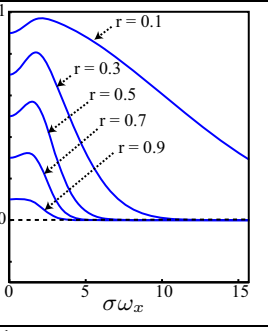
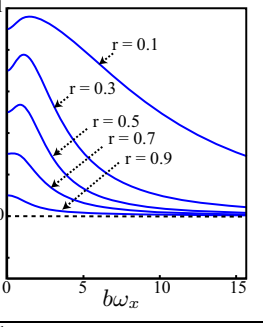
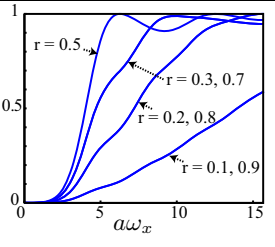
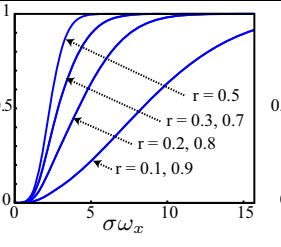
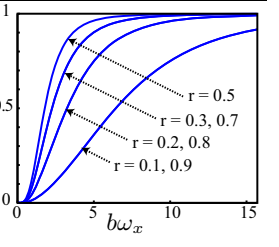
In the remainder, I assume that the unpredictable noise components, $N_L(\omega_x, \omega_y)$ and $N_R(\omega_x, \omega_y)$, are small relative to $V(\omega_x, \omega_y)$, and set $\theta_L = \theta_R = 0$. Because it is difficult to accurately estimate $N_L(\omega_x, \omega_y)$ and $N_R(\omega_x, \omega_y)$, this assumption greatly simplifies further analysis. Note that under this assumption, $\hat{F}_L$, $\hat{F}_R$, and G can be written as functions of ω_x without relation to ω_y.

Several interesting observations on (17) are obtained. First, the optimal combining filters are determined by the disparity errors characterized by $p(\xi)$ and the viewpoint of the target image denoted by r, *without* relation to the spectral shape of the target image $V(\omega_x, \omega_y)$. Second, several boundary conditions that are naturally required by definition are satisfied; $(\hat{F}_L, \hat{F}_R) = (1, 0)$ for $r = 0$, $(\hat{F}_L, \hat{F}_R) = (0, 1)$ for $r = 1$, and $\hat{F}_L = \hat{F}_R$ for $r = 1/2$ can be easily confirmed. Finally, when $\psi(z)$ is even, $\hat{F}_L$ and $\hat{F}_R$ are symmetric with respect to $r = 1/2$, which means that $\hat{F}_L$ for r and $\hat{F}_R$ for $1 - r$ are the same.

3.4 Examples

Let three probability distributions, uniform, Gaussian, and Laplacian, be assumed for $p(\xi)$ as typical models of the disparity error. See Table 1 for a summary. The forms of $p(\xi)$ and $\psi(z)$ are listed in the second and third rows, respectively. In the fourth row, each graph illustrates the frequency response of $\hat{F}_L$ with different values of r. Note that the horizontal axes are scaled by a, σ, or b for the uniform, Gaussian, or Laplacian distributions, respectively. The responses of $\hat{F}_R$ can be obtained by just replacing r with $1 - r$ because F_L and F_R are symmetric. As can be seen from these graphs, the optimal filter emphasizes middle

Table 1. Optimal combining filter ($\hat{F}_L(\omega_x)$) and gain term ($G(\omega_x)$) for disparity errors ($p(\xi)$) following uniform, Gaussian, and Laplacian distributions (a, σ, $b \geq 0$)

	uniform	Gaussian	Laplacian
$p(\xi)$	$\begin{cases} 1/(2a) & (-a \leq \xi \leq a) \\ 0 & (\text{otherwise}) \end{cases}$	$\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\xi^2}{2\sigma^2}\right)$	$\frac{1}{2b} \exp\left(-\frac{\|\xi\|}{b}\right)$
$\psi(z)$	$\frac{\sin(az)}{az}$	$\exp\left(-\frac{(\sigma z)^2}{2}\right)$	$\frac{1}{1+(bz)^2}$
$\hat{F}_L(\omega_x)$			
$G(\omega_x)$			

frequencies and reduces high frequencies. The bottom row shows the gain term $G(\omega_x)$ for each distribution, whose form is obtained by using (11).

As mentioned before, $G(\omega_x)$ is the gain between the original image and the view interpolation error in the frequency domain; the smaller $G(\omega_x)$ means less error. *The optimal combing filters are to minimize $G(\omega_x)$ for all frequencies because they are designed to minimize the MSE.* At least, they should be better than linear interpolation, whose combining filters in the frequency domain are given by $F_L(\omega_x) = 1 - r$ and $F_R(\omega_x) = r$. The comparison between the optimal and linear interpolations for the three distribution models are shown in Fig. 2. It is clear that $G(\omega_x)$ of the optimal interpolation is always smaller than that of linear interpolation.

Another important observation is that the gain of the optimal interpolation never exceeds 1.0, which means that *the optimal interpolation keeps the error below the original signal for all frequencies.* By contrast, linear interpolation amplifies the error in high frequencies with the gain that exceeds 1.0. However, if linear interpolation is combined with appropriate band-limitation, near-optimal

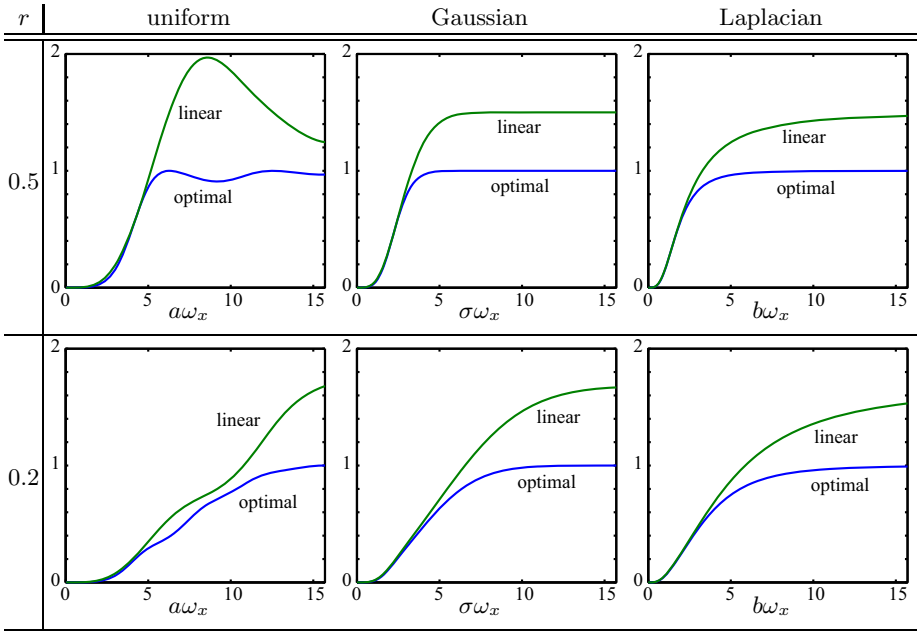


Fig. 2. Comparison of gain term $G(\omega_x)$ between optimal and linear interpolations

interpolation can be achieved. Let $G_{lin}(\omega_x)$ be the gain term of linear interpolation. If the input images are band-limited to the spectral range where $G_{lin}(\omega_x) \leq 1.0$, the resulting gain becomes $\min\{G_{lin}(\omega_x), 1.0\}$, avoiding the overshoots in high frequencies. This method, referred to as *band-limited linear interpolation*, has a practical advantage in that the implementation is much easier than the naive implementation of the optimal interpolation. The spectral components where $G_{lin}(\omega_x) \leq 1.0$ are kept without any change and other components are just reduced to 0. The cut-off frequency of this method is derived from the shape of the gain term $G(\omega_x)$, while the band-limitation in [2,8] was based on the anti-alias condition.

3.5 Relation with Linear Interpolation

Finally, I investigate the theoretical relation between the optimal and linear interpolations. I give a theoretical basis to linear interpolation used just *heuristically* without sufficient justification.

I first discuss the three distribution models mentioned above. As seen from the third row of Table 1, $\psi(z)$ can be written as a function of kz, where k takes a, σ, or b, respectively. k represent the magnitude of the disparity error; a larger k corresponds to a larger error. For all of those distributions, $\psi(z)$ can be expanded around $kz = 0$ as:

$$\psi(z) = 1 - A \cdot (kz)^2 + O\left((kz)^4\right), \tag{18}$$

where A takes 1/6, 1/2, and 1 for the uniform, Gaussian, and Laplacian distributions, respectively. Substituting (18) into (17),

$$\lim_{k\omega_x \to 0} \hat{F}_L(\omega_x, \omega_y) = 1 - r, \ \lim_{k\omega_x \to 0} \hat{F}_R(\omega_x, \omega_y) = r, \tag{19}$$

can be proven. This result can be confirmed from the fourth row of Table 1; $\hat{F}_L$ reaches $1-r$ as $k\omega_x$ reaches 0 for all distribution models. The physical meaning is straightforward. *For small disparity errors ($k \sim 0$) or low frequencies ($\omega_x \sim 0$), the optimal interpolation converges to conventional linear interpolation.* Equation (19) also holds true when $p(\xi)$ is represented as a weighted sum of those three distributions.

As shown by (8)–(10), the final MSE depends on the spectral shape of the target image $V(\omega_x, \omega_y)$. It should be noted that $V(\omega_x, \omega_y)$ of a natural image tends to be low-frequency oriented (most of the image energy resides in low frequencies). In addition, as seen in Fig. 2, the difference in $G(\omega_x)$ between the optimal and linear interpolations is marginal for small $k\omega_x$. Consequently, the difference in MSE between the optimal and linear interpolations would be quite small for practically small values of k, which would justify the use of linear interpolation as an approximation of the optimal interpolation.

The discussion above can be partly extended to general distributions. Equation (13) can be rewritten as

$$\psi(z) = \int_{-\infty}^{\infty} p(\xi) \sum_{n=0}^{\infty} \frac{(j\xi z)^n}{n!} d\xi = \sum_{n=0}^{\infty} \left\{ \int_{-\infty}^{\infty} p(\xi) \xi^n d\xi \right\} \frac{(jz)^n}{n!} = \sum_{n=0}^{\infty} \mu_{\xi,n} \frac{(jz)^n}{n!}, \tag{20}$$

where $\mu_{\xi,n}$ is the n-th order moment of ξ. $\mu_{\xi,0} = 1$ by definition of the probability distribution $p(\xi)$. $\mu_{\xi,1}$ is the mean of ξ, and let it be 0. $\mu_{\xi,2}$ is the variance and described as σ_ξ^2. If all of the higher order moments are convergent, Equation (20) can be approximated around $z = 0$ as

$$\psi(z) \sim 1 - \sigma_\xi^2 z^2/2 \tag{21}$$

Substitution of (21) into (17) leads to

$$\lim_{\omega_x \to 0} \hat{F}_L(\omega_x, \omega_y) = 1 - r, \ \lim_{\omega_x \to 0} \hat{F}_R(\omega_x, \omega_y) = r. \tag{22}$$

Unlike (19), (22) gives no information about k. This is reasonable because no explicit form is assumed for $p(\xi)$, thereby there is no parameter such as k that characterizes $p(\xi)$. However, also for this case, the optimal interpolation converges to linear interpolation as ω_x reaches 0.

4 Experiment

Figure 3 illustrates the flow of the experiment (See Fig. 1 for the configuration). A complete MATLAB implementation is included in the supplementary material.

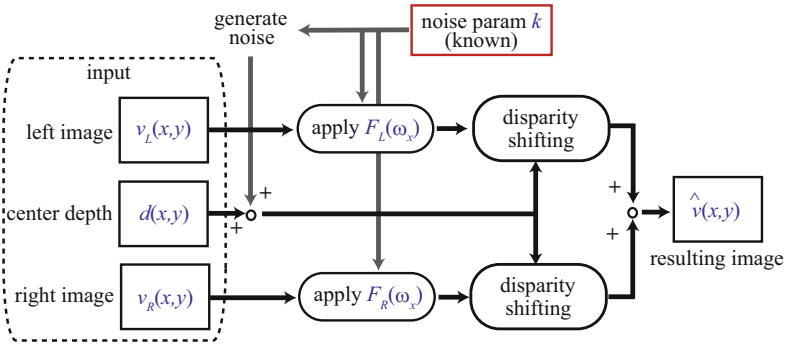


Fig. 3. Flow of experiment. Artificial noise is added to disparity map to simulate depth inaccuracy. Its parameter is used to design combining filters, F_L and F_R.

The input is a pair of stereo images, denoted as $v_L(x, y)$ and $v_R(x, y)$, respectively, and a pixel-wise depth map from the target viewpoint, denoted as $d(x, y)$. The pixel value of $d(x, y)$ represents the disparity (pixel shift) that is measured between the viewpoints of $v_L(x, y)$ and $v_R(x, y)$.

First, artificial noise is added to $d(x, y)$ to simulate the disparity error, yielding $d'(x, y)$. The disparity error is random, but the distribution $p(\xi)$ is assumed to be known. Next, combining filters are applied to the input images. The optimal filters are derived from (17) using the known distribution of the disparity error, $p(\xi)$. Although this filtering operation is applicable as convolution in the spatial domain, it is implemented in the frequency domain because the purpose of the experiments is to confirm the theory. To be precise, the left image after filtering operation, $\dot{v}_L(x, y)$, is obtained as

$$\dot{v}_L(x,y) = f_L(x,y) \circ v_L(x,y) = \mathcal{F}^{-1}\left[F_L(\omega_x,\omega_y) \cdot \mathcal{F}\left[v_L(x,y)\right]\right], \tag{23}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ represent Fourier transform and its inverse. For linear interpolation, the filtering operation is just multiplying the input images by a weighting coefficient, $1 - r$ or r, and is implemented in the spatial domain. The band-limitation introduced in Section 3.4 is also implemented as an option.

Then, disparity shifting is performed over the filtered left image $\dot{v}_L(x, y)$ using the noisy disparity map $d'(x, y)$, yielding

$$\ddot{v}_L(x,y) = \dot{v}_L\left(x + r \cdot d'(x,y),\ y\right). \tag{24}$$

In this operation, cubic spline interpolation is adopted to read fractional pixel positions. $\ddot{v}_R(x, y)$ is also obtained in the same way. Finally, the resulting image from the target viewpoint is obtained as

$$\hat{v}(x,y) = \ddot{v}_L(x,y) + \ddot{v}_R(x,y). \tag{25}$$

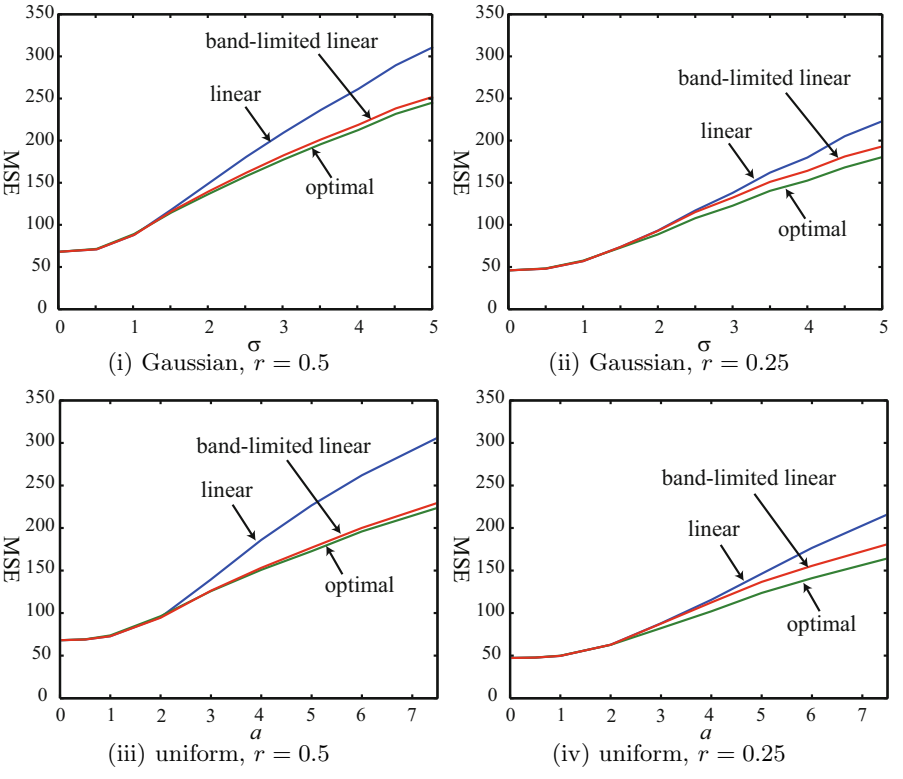


Fig. 4. Profiles of MSE (y axes) against magnitude of disparity error (x axes) obtained with *cones* dataset. Optimal interpolation outperforms linear interpolation as a whole. Band-limited linear interpolation achieves near-optimal quality.

$\hat{v}(x, y)$ is compared with the ground truth $v(x, y)$ to evaluate the accuracy. No occlusion handling was performed because it is out of scope of this paper.[3]

Two real image datasets, *cones* and *teddy*, from Middlebury stereo website[4] are used for experiments. All images are 450 × 375 pixels and converted into 8-bit grayscale. The target viewpoint is set to the position of *im2*, whose corresponding disparity map *disp*2 is also available. When *im0* and *im4* are used for $v_L(x, y)$ and $v_R(x, y)$, respectively, r equals 0.5. Instead, when *im1* and *im5* are used, r becomes 0.25. In both cases, the accuracy of the original depth information is 1/4 pixel length in terms of disparities between the left and right images.[5] The

[3] The procedure described here seems a little different from that of (2) and (3); the disparity shifting and combing filtering are performed in the opposite order. The order is changed to avoid hole-filling problems. These two procedures are completely equivalent in the theoretical model.

[4] http://vision.middlebury.edu/stereo/

[5] Consequently, inherent quantization errors are included in the input depth data. Those relatively small errors are ignored in the experiment.

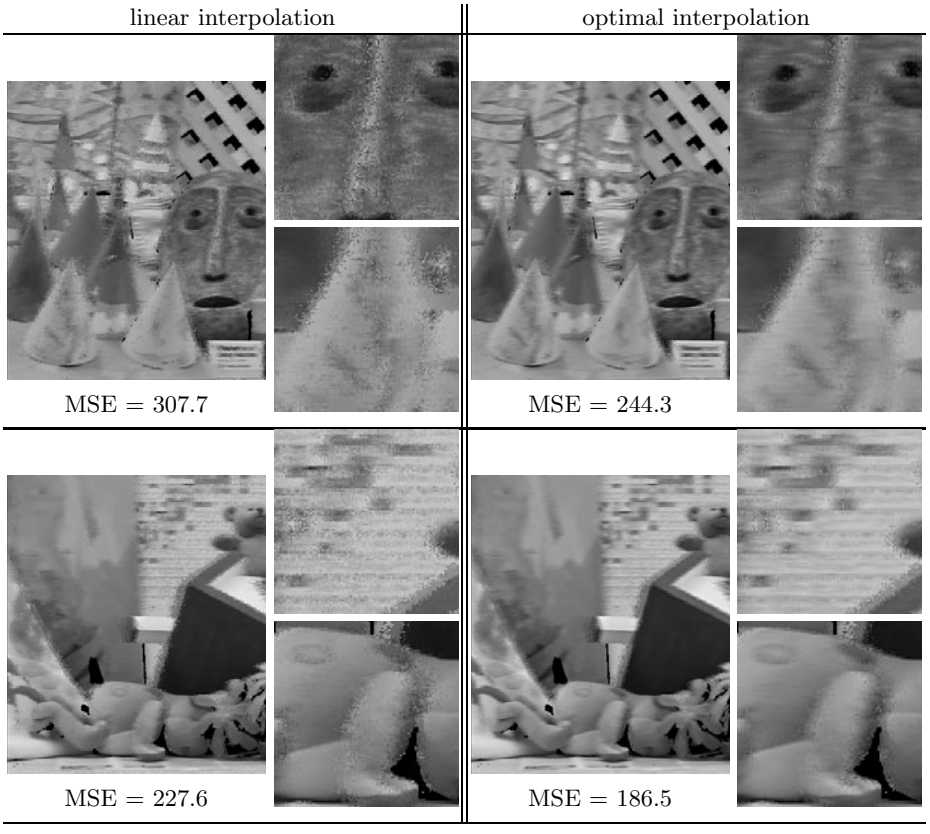


Fig. 5. Resulting images and close-ups by linear (left) and optimal (right) interpolations. Disparity errors are Gaussian of $\sigma = 5$ and viewpoint is set to $r = 0.5$. Optimal interpolation produces more plausible results.

resulting image is compared with the ground truth *im2* for evaluation of MSE, which are calculated in floating point precision without considering the pixels with void disparities and 50 pixels from both sides.

Figure 4 shows the MSE profiles against the magnitude of the disparity error obtained with *cones* dataset. The disparity error is Gaussian for the upper row and uniform for the lower row. It is obvious that the optimal interpolation outperforms linear interpolation as a whole. As the theory states, the difference between the optimal and linear interpolations decreases as the disparity error decreases. Sometimes, linear interpolation outperforms the optimal interpolation by a very narrow margin. However, as the disparity error increases, the advantage of the optimal interpolation becomes significant. Another important observation is that band-limited linear interpolation achieves comparable quality with the optimal interpolation.

Figure 5 shows several resulting images by the linear and optimal interpolations with large disparity errors (Gaussian with $\sigma = 5$). These images are low

quality because of the very noisy disparities used, but the optimal interpolation produces more plausible results than linear interpolation in visual quality. More results are included in the supplementary material.

It should be noted that the proposed theory starts with an occlusion-free diffusive surface model (see (1)), and the noise terms are actually ignored (θ_R and θ_L in (16) are assumed to be 0) in the optimization. However, the optimal interpolation works well for real scenes that contain non-diffusive reflections and occlusions, indicating that the proposed theory successfully captures the essential property of real image data despite its simplicity.

5 Conclusions

A theoretical framework for view interpolation problem to analyze the quantitative quality of the resulting image in the presence of depth inaccuracy and provide a principled optimization scheme based on the MSE metric was proposed. The theory clarified that if the probabilistic distribution of the disparity error is available, the optimal view interpolation that outperforms conventional linear interpolation can be achieved. It was also revealed that the optimal interpolation converges to linear interpolation as the disparity inaccuracy decreases. The theory was confirmed by experiment using real image data.

The main drawback of the optimal interpolation scheme is that it requires the exact shape of the disparity error distribution $p(\xi)$, which may be infeasible in practice. Furthermore, the advantage of optimal interpolation over linear interpolation is marginal unless the disparity inaccuracy is considerably large. Consequently, when the depth information is accurate to some extent, linear interpolation is a realistic choice. If the depth information is very noisy, band-limited linear-interpolation can be adopted, because it can achieve near-optimal quality only with simple band limitation. Although the cut-off frequency also depends on the shape of $p(\xi)$, this parameter might be tuned interactively.

Future work will include several directions. First, the theory will be extended to deal with more general configurations, e.g., 2-D configurations of input cameras and the target viewpoint not located on the baseline. Second, the probabilistic distribution of the disparity error should be studied from real data. For example, most stereo matching algorithms produce depth maps with spatially varying errors, which conflicts with the assumption that $p(\xi)$ is space invariant. In addition, efficient implementations or approximations for the optimal combining filters should be considered.

Finally, it should be noted that view interpolation is essentially a very complex problem. As seen from [17], many technical elements including accurate camera calibration, stable depth/correspondence estimation, and appropriate handling of occlusion boundaries, contribute to the final rendering quality. Meanwhile, the theory presented in this paper is focused only on a single aspect of the problem: *how to blend input image pixels to synthesize new output pixels when the correspondences are established with some amount of errors*, which is a common issue for any view interpolation algorithm. I believe the theory can be extended

to deal with other aspects in the future and it leads to a solid mathematical framework for general view interpolation problems.

References

1. Buehler, C., Bosse, M., McMillan, L., Gortler, S., Cohen, M.: Unstructured lumigraph rendering. In: ACM SIGGRAPH Papers pp. 425–432 (2001)
2. Chai, J., Tong, X., Chany, S.C., Shum, H.Y.: Plenoptic sampling. In: ACM Trans. Graphics (Proc. ACM SIGGRAPH), pp. 307–318 (2000)
3. Chen, S.E., Williams, L.: View interpolation for image synthesis. In: Proc. ACM SIGGRAPH, pp. 279–288 (1993)
4. Girod, B.: The efficiency of motion-compensating prediction for hybrid coding of video sequences. IEEE Journal SAC SAC 5(7), 1140–1154 (1987)
5. Gortler, S.-J., Crzeszczuk, R., Szeliski, R., Cohen, M.-F.: The lumigraph. In: Proc. ACM SIGGRAPH, pp. 43–54 (1996)
6. Kubota, A., Smolic, A., Magnor, M., Tanimoto, M., Chen, T., Zhang, C.: Special issue on multi-view imaging and 3dtv. IEEE Signal Processing Magazine 24(6), 10–111 (2007)
7. Levoy, M., Hanrahan, P.: Light field rendering. In: Proc. ACM SIGGRAPH, pp. 31–42 (1996)
8. Lin, Z., Shum, H.Y.: A geometric analysis of light field rendering. Intl. Journal of Computer Vision 58(2), 121–138 (2004)
9. Ramanathan, P., Girod, B.: Rate-distortion analysis for light field coding and streaming. EURASIP SP:IC 21(6), 462–475 (2006)
10. Shade, J.W., Gortler, S.J., He, L.W., Szeliski, R.: Layered depth images. In: Proc. ACM SIGGRAPH, pp. 231–242 (1998)
11. Shum, H.Y., Kang, S.B., Chan, S.C.: Survey of Image-Based Representations and Compression Techniques. IEEE Trans. CSVT 13(11), 1020–1037 (2003)
12. Taguchi, Y., Takahashi, K., Naemura, T.: Real-time all-in-focus video-based rendering using a network camera array. In: Proc. 3DTV-Conference, pp. 241–244 (2008)
13. Takahashi, K., Naemura, T.: Theoretical model and optimal prefilter for view interpolation. In: Proc. IEEE ICIP, pp. 1528–1531 (2008)
14. Tong, X., Chai, J., Shum, H.Y.: Layered lumigraph with lod control. The Journal of Visualization and Computer Animation 13(4), 249–261 (2002)
15. Zhang, C., Chen, T.: Spectral analysis for sampling image-based rendering data. IEEE Trans. CSVT 13(11), 1038–1050 (2003)
16. Zhang, C., Chen, T.: A survey on image-based rendering - representation, sampling and compression. EURASIP SP:IC 19(1), 1–28 (2004)
17. Zitnick, C., Kang, S.B., Uyttendaele, M., Winder, S., Szeliski, R.: High quality video interpolation using a layered representation. In: ACM SIGGRAPH Papers, pp. 600–608 (2004)