IMAGE BASED RENDERING WITH DEPTH CAMERAS: HOW MANY ARE NEEDED?

Christopher Gilliam, James Pearson, Mike Brookes and Pier Luigi Dragotti

Electrical and Electronic Engineering Department, Imperial College London
Exhibition Road, SW7 2AZ, London, United Kingdom.

ABSTRACT

Image based rendering is a technique for producing arbitrary viewpoints of a scene using multiple images instead of exact object models. The recent emergence of low-price, fast, and reliable cameras for measuring depth makes possible the augmentation of traditional color images with depth images. This combination promises to improve the rendering quality of an arbitrary viewpoint and thus have a great impact on IBR. A key issue is to understand, for any particular scene of interest, how many depth images and how many color images are necessary in order to obtain good rendering results. In this paper, using a framework akin to the plenoptic function, we perform a spectral analysis of multi-view depth images in order to determine the relationship between the number of depth and color images required. Our analysis is then validated using both synthetic and real images.

***Index Terms*—** Depth cameras, Image-Based Rendering, plenoptic function, sampling, spectral analysis

1. INTRODUCTION

Image Based Rendering (IBR) is an effective technique for rendering novel views from a set of available multi-view images. Instead of rendering views of 3-D scenes by projecting objects and their textures, novel views are rendered by interpolating available nearby images. The advantage of such a method is that it produces convincing photorealistic results since the interpolated viewpoints are obtained through combinations of real images. The main drawback is the fact that a huge amount of data needs to be captured.

Clearly, knowledge of the scene geometry reduces the number of images required. The interplay between geometry and sampling rate (number and spacing of cameras) has been extensively studied in the recent past (e.g., [1, 2, 3, 4, 5]). Unfortunately, 3-D reconstruction techniques from passive cameras, are still not reliable and do not work well in many cases. This fact has profoundly limited the use of IBR ideas. Recent advances in sensing technologies may, however, soon allow large-scale deployment of 3-D cameras using active depth sensing systems. These cameras are able to estimate depth and geometry with good accuracy and reliability, and for this reason can be very useful in IBR [6]. A natural question then is to understand the interplay between the number of depth and color cameras. Specifically, given a scene of interest with a certain geometry, how many depth cameras are necessary to infer the geometry and how many color cameras are then needed, given the inferred geometry, to render novel photorealistic views?

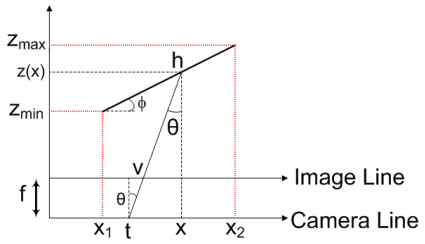


Fig. 1. Scene model of a slanted plane where $z(x)$ is the depth at x, f is the focal length and h is the curvilinear coordinate. Note that θ is the viewing angle and ϕ is the slant of the plane.

To answer this question we put ourselves in the typical Shannon sampling framework and perform a spectral analysis of both multi-view depth images and multi-view color images. In that respect we continue the work of several researchers [1, 2, 3, 4]. In particular we use the formalism developed in a previous paper [7] and expand it to include the case of depth cameras. We show that the interplay between the required number of depth and color cameras mostly depends on the resolution of the color cameras and the bandwidth of the texture of the scene. Our analysis is then validated using both synthetic and real images.

The paper is organized as follows: In the next section, we review the spectral analysis of multi-view color images, in particular the concept of the plenoptic function [8]. In Sec. 3, we present a spectral analysis of multi-view depth images of a slanted plane and determine a maximum acceptable spacing between the depth cameras. We present results validating this analysis for both synthetic and real images in Sec. 4. We finally conclude in Sec. 5.

2. THE PLENOPTIC FUNCTION

At the heart of IBR is the idea that a scene can be represented as a collection of light rays emanating from the scene. The light rays in question are described using the 7-D plenoptic function [8]. The number of dimensions, however, can

be reduced by constraining the sensing setup. For example, the case when cameras lie on a plane leads to the 4-D lumigraph [9] or lightfield [10] parametrization. This parametrization is obtained by using two parallel planes: the camera plane (s,t) and the image plane (u,v). The distance between the two planes is the focal length, f. In this parametrization, the function $p(s,t,u,v)$ represents the intensity of the light ray at camera location (s,t) and pixel location (u,v).

A further simplification made in [1] is to fix s and u, corresponding to the situation where the camera positions are constrained to a 1-D camera line and only one scan-line is considered in each image, see Fig. 1. In this case the lightfield is reduced to two dimensions: $p(t,v)$. This representation is also known as the Epipolar Plane Image (EPI).

2.1. Spectral Analysis of the Plenoptic Function

IBR can be seen as the problem of sampling and interpolating the plenoptic function. Therefore by examining its spectral properties we can determine the maximum acceptable spacing between the color cameras. Using the EPI parametrization, the plenoptic spectrum is defined as $P(\omega_t,\omega_v)=\mathcal{F}\{p(t,v)\}$, where $\mathcal{F}$ is the Fourier transform operator. The properties of the plenoptic spectrum were studied for the first time in [1]. By assuming a Lambertian scene with no occlusion, the authors showed that the spectrum is approximately bounded by lines related to the maximum and minimum depths of the scene and that finite camera resolution bandlimits the spectrum. This spectral analysis was shown to be exact in [2] and extended to more general cases, in particular non-Lambertian and occluded scenes, in [3].

This spectral analysis was re-examined in [7] for the simple case of a slanted plane, see Fig. 1, but with two additional constraints: finite scene width (FSW) and cameras with finite field of view (FFoV). Using these constraints the authors presented a closed-form expression for the plenoptic spectrum of a slanted plane with bandlimited texture. The resulting spectrum was band-unlimited in both ω_t and ω_v. However, by assuming the function is bandlimited to an essential bandwidth, [7] determined a maximum acceptable camera spacing for the reconstruction of the plenoptic function. The essential bandwidth was defined such that it contained 90% of the signal's energy. Based on this spectral result, [5] formulated an algorithm to determine the optimum positioning for a finite number of cameras to sample a scene with a smoothly varying surface. However, in order for the algorithm to operate it requires prior knowledge of the scene geometry.

3. THE PANTELIC FUNCTION

We propose treating multi-view depth images as samples of a function, $q(t,v)$, which we term the Pantelic[1] function. It describes the inverse depth of the scene captured at camera location t and pixel location v. Therefore, by performing spectral analysis on the function, we can determine the minimum number of depth cameras required to reconstruct the scene geometry. The reconstructed geometry can then be used in the adaptive sampling algorithm proposed in [5].

We focus on the spectral analysis of the pantelic function for a slanted plane, see Fig. 1. The scene geometry equations for this scene are

$$\mathcal{G}_s=\begin{cases} x=h\cos(\phi)+x_1 \\ z(x)=(x-x_1)\tan(\phi)+z_{min}\end{cases} \tag{1}$$

where $x\in[x_1,x_2]$ is the spatial position, $z\in[z_{min},z_{max}]$ is the depth and ϕ is the angle between the plane and the line $z=z_{min}$. The finite width of the plane is T, hence $h\in[0,T]$ is the curvilinear coordinate.

Having defined the scene geometry, we use the functional framework outlined in [2] in order to relate a point on the scene at $(x,z(x))$ to a camera location t and pixel location v, hence

$$t=x-z(x)\frac{v}{f}. \tag{2}$$

Note that the FFoV constraint restricts v to $v\in[-v_m,v_m]$. The relationship in (2) is restricted to a one to one mapping using a no-occlusion constraint

$$\frac{f}{v_m}>|z'(x)|=|\tan(\phi)|\,, \tag{3}$$

where $z'(x)$ is the first differential of $z(x)$ with respect to x. On a last note, the absolute depth of a point on the scene is independent of the camera location, thus using (2) the following is true $q(t,v)=\hat{q}(x)=1/z(x)$.

3.1. Derivation of Pantelic Spectrum

Starting with the Fourier transform, $Q(\omega_t,\omega_v)$, of the pantelic function, $q(t,v)$, we apply the FFoV and FSW constraints to obtain

$$\begin{aligned} Q(\omega_t,\omega_v)&=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}q(t,v)\,e^{-j(\omega_t t+\omega_v v)}\,dtdv,\\ &\overset{(i)}{=}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(1-z'(x)\frac{v}{f}\right)\hat{q}(x)e^{-j(\omega_t(x-z(x)v/f)+\omega_v v)}dxdv,\\ &\overset{(ii)}{=}\int_{x_1}^{x_2}\hat{q}(x)e^{-j\omega_t x}\int_{-v_m}^{v_m}\left(1-z'(x)\frac{v}{f}\right)e^{-j(\omega_v-z(x)\omega_t/f)v}dvdx,\\ &\overset{(iii)}{=}2v_m\int_{x_1}^{x_2}\hat{q}(x)\left[\text{sinc}\,(\Omega)-j\frac{\tan(\phi)v_m}{f}\text{sinc}'(\Omega)\right]e^{-j\omega_t x}dx, \end{aligned} \tag{4}$$

where $\text{sinc}'(\Omega)$ is the first derivative of the sinc function with respect to the argument, and $\Omega=\omega_v v_m-(z(x)v_m\omega_t)/f$. Step (i) follows from applying (2) and $q(t,v)=\hat{q}(x)$. Step (ii) follows from applying the FFoV and FSW constraints, and the last step from solving the integral in v.

[1] With a slight abuse, we derived the word pantelic from the Greek $\pi\alpha\nu$ meaning all, and $\tau\eta\lambda\epsilon$ meaning at distance.

By changing the variable of integration from x to h, using (1), the equation in (4) becomes

$$Q(\omega_t,\omega_v) = 2v_m \left[\int_0^T \frac{\cos(\phi)\text{sinc}(\hat{\Omega})}{h\sin(\phi)+z_{min}} e^{-j\omega_t \cos(\phi)h} dh \right.$$
$$\left. -j\frac{v_m \sin(\phi)}{f} \int_0^T \frac{\text{sinc}'(\hat{\Omega})}{h\sin(\phi)+z_{min}} e^{-j\omega_t \cos(\phi)h} dh \right], \quad (5)$$

where

$$\hat{\Omega} = \omega_v v_m - (h\sin(\phi)+z_{min})\frac{v_m}{f}\omega_t.$$

Note that for clarity in this derivation, and without loss of generality, we set $x_1 = 0$. At this point, we perform another change of variable, coupled with integration by parts, to obtain the following expression for the pantelic spectrum

$$Q(\omega_t,\omega_v) = j\,2v_m \left(\frac{\text{sinc}(a)}{z_{max}\omega_t} e^{-j(a-b)c} - \frac{\text{sinc}(b)}{z_{min}\omega_t} \right)$$
$$- j\frac{2v_m^2}{f} e^{jbc} \int_b^a \frac{\text{sinc}(\hat{\Omega})}{(\omega_v v_m - \hat{\Omega})^2} e^{-j\hat{\Omega}c}\, d\hat{\Omega} \quad (6)$$

where

$$a = \omega_v v_m - \omega_t \frac{z_{max} v_m}{f}, \qquad b = \omega_v v_m - \omega_t \frac{z_{min} v_m}{f},$$
$$\text{and} \qquad c = \frac{-f}{\tan(\phi) v_m}.$$

An exact closed-form expression for the pantelic spectrum is obtained from (6) by solving the remaining integral.[2] Note that (6) is only valid for $\omega_t \neq 0$, if $\omega_t = 0$ then

$$Q(0,\omega_v) = \frac{2v_m}{\tan(\phi)} \ln\left(\frac{z_{max}}{z_{min}}\right) \left(\text{sinc}(\omega_v v_m) \right.$$
$$\left. -j\frac{v_m \tan(\phi)}{f} \text{sinc}'(\omega_v v_m) \right). \quad (7)$$

A comparison between the pantelic spectrum for a slanted plane and its corresponding plenoptic spectrum is shown in Fig. 2. The pantelic spectrum is illustrated in Fig. 2(a) and the plenoptic spectrum, assuming sinusoidal texture, is in Fig. 2(b). Note that the spectrum in Fig. 2(a) is computed using our expression however if we compare it to that calculated numerically the PSNR between the two is 72.8dB.

3.2. Essential Bandwidth of the Pantelic Spectrum

Similar to the plenoptic spectrum, the pantelic spectrum of a slanted plane under FSW and FFoV is band-unlimited in both ω_t and ω_v. Therefore the maximum acceptable spacing between the depth cameras is determined using the essential bandwidth of the pantelic spectrum. One approach to this problem would be to use the essential bandwidth for the

[2] We omit the detail of this derivation due to the lack of space.

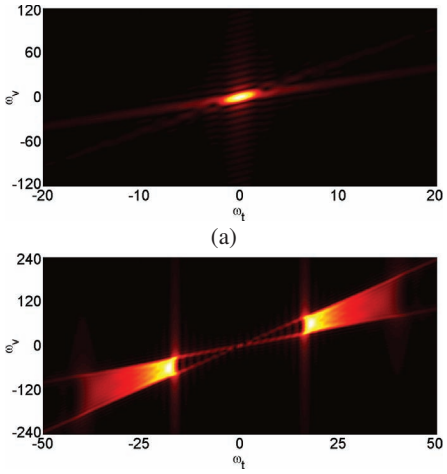


Fig. 2. Comparison of the pantelic spectrum, (a), and the plenoptic spectrum, (b), for a slanted plane with sinusoidal texture pasted to the surface.

plenoptic spectrum from [5]. However a simple visual comparison of the spectra in Fig. 2 highlights the compactness of the pantelic spectrum around the origin compared to the plenoptic spectrum. This suggest that fewer depth cameras are required than image cameras. As a result we opt for a rectangular essential bandwidth centered around the origin.

This region is determined by approximating the bandwidth along $Q(0,\omega_v)$ as the bandwidth of $\text{sinc}(\omega_v v_m)$, which is π/v_m, to give a maximum value in ω_v. This value is then projected onto the ω_t-axis using $\omega_v = z_{min}\omega_t/f$. As a result the essential bandwidth is defined by the region

$$\mathcal{B} = \left\{ \omega_t, \omega_v : |\omega_t| \in \left[0, \frac{f\pi}{v_m z_{min}}\right], |\omega_v| \in \left[0, \frac{\pi}{v_m}\right] \right\}. \quad (8)$$

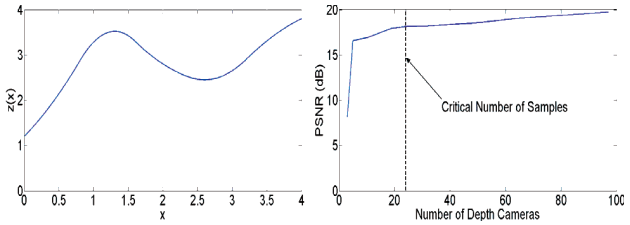
By using the above expression, the maximum acceptable depth camera spacing is given by

$$\Delta t = 2\pi \left(2\frac{f\pi}{v_m z_{min}}\right)^{-1} = \frac{z_{min} v_m}{f}. \quad (9)$$

Therefore the depth camera spacing for a slanted plane is only dependent on the minimum depth of the scene and camera characteristics. In contrast the equivalent color camera spacing is also dependent upon the scene geometry and inversely proportional to the maximum frequency of the texture. Consequently fewer depth cameras than color cameras are required for a slanted plane.

4. RESULTS

In this section we extend the analysis, empirically, to more complex synthetic and real scenes. The synthetic scene is a piecewise quadratic surface, see Fig. 3(a), with sinusoidal texture. The scene is sampled using depth cameras and an estimate of the surface obtained from $1/q(t,0)$. This surface estimate is then used as a prior to sample and reconstruct the



(a) (b)

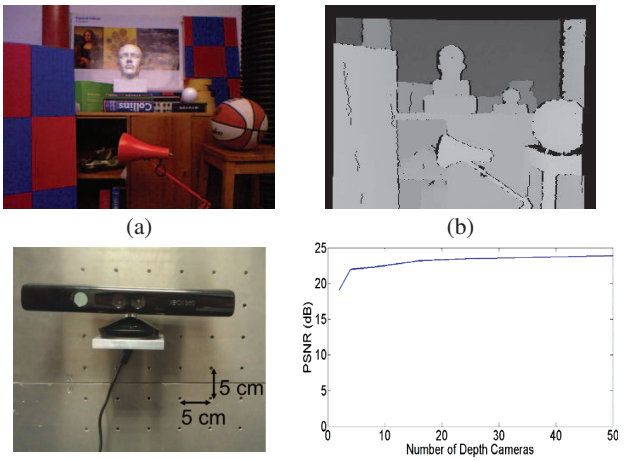
Fig. 3. The synthetic data. (a) Diagram of the piecewise quadratic surface. (b) Graph of the PSNR of the reconstructed EPI as the number of depth cameras increases. Note that the number of color cameras used in the reconstruction is fixed at 250.

plenoptic function, assuming a fixed number of color images, using the algorithm in [5]. The PSNR of the reconstructed plenoptic function is shown in Fig. 3(b) as the number of depth cameras varies. The critical sample point marked on Fig. 3(b) is obtained by using (9) as an approximation. Notice that the PSNR of the reconstruction begins to saturate after this point, which suggests that (9) is a good approximation of the depth camera spacing for more complex scenes.

For the real data, we generated our own 4-D lightfield testset comprising both color and depth images as shown in Fig. 4(a) and Fig. 4(b), respectiviely. The images were acquired using Microsoft's Xbox Kinect camera mounted to our camera rig, see Fig. 4(c). The camera rig comprises a 10x10 grid with a separation of 5cm, horizontally and vertically, resulting in a 100 color and depth image pairs. Similar to the synthetic scene, a sparse number of color images is fixed beforehand and the number of depth images varied. Using this subset of images, the remainder are rendered using the algorithm detailed in [11]. Fig. 4(d) shows the PSNR between the rendered images and their unused ground truths as the number of depth cameras varies. In line with our theory, the vast majority of the depth cameras are redundant and as few as 16 are necessary for rendering with little drop in the synthesized quality.

5. CONCLUSIONS

We have presented a spectral analysis of multi-view depth images using a framework akin to the plenoptic function. Similar to the initial research on plenoptic sampling, we assume no-occlusions and derived an exact expression for the spectrum of a multi-view depth image set belonging to a slanted plane. From this expression we determined a maximum acceptable depth camera spacing for a slanted plane, which depends on the minimum depth of the scene and the camera characteristics. Therefore, when performing IBR on a slanted plane, fewer depth cameras are required than image cameras. Finally, we validate this statement empirically for more complex scenes using both synthetic and real images. Our results show that our derived maximum depth camera spacing is a good approximation for more complex scenes as the rendering quality saturates beyond that point.



(a) (b) (c) (d)

Fig. 4. The real data. An example of a color image, (a), and a depth image, (b), of the scene captured at a camera position. (c) The Microsoft Xbox Kinect camera mounted on the 10-by-10 grid. Each camera position is 5cm apart, vertically and horizontally. (d) Graph of the rendered image PSNR as the number of depth cameras increases. The number of color cameras used in the rendering is fixed.

6. REFERENCES

[1] J.X. Chai, X. Tong, S.C. Chan, and H.Y. Shum, "Plenoptic sampling," in *Proc. SIGGRAPH*, July 2000, pp. 307–318.

[2] M. Do, D. Marchand-Maillet, and M. Vetterli, "On the bandwidth of the plenoptic function," *IEEE Trans. on Image Process.*, 2011, DOI: 10.1109/TIP.2011.2163895.

[3] C. Zhang and T. Chen, "Spectral analysis for sampling image-based rendering data," *IEEE Trans. on Circ. and Syst. for Video Tech.*, vol. 13, no. 11, pp. 1038–1050, November 2003.

[4] C. Chen and D. Schonfeld, "Geometrical plenoptic sampling," in *Proc. ICIP*, Cairo (Egypt), November 2009.

[5] C. Gilliam, P.L. Dragotti, and M. Brookes, "Adaptive plenoptic sampling," in *Proc. ICIP*, Brussels (Belgium), September 2011.

[6] M.N. Do, Q.H. Nguyen, H.T. Nguyen, D. Kubacki, and S.J. Patel, "Immersive visual communication," *IEEE Signal Processing Mag.*, vol. 28, no. 1, pp. 58–66, January 2011.

[7] C. Gilliam, P.L. Dragotti, and M. Brookes, "A closed-form expression for the bandwidth of the plenoptic function under finite field of view constraints," in *Proc. ICIP*, Hong Kong, September 2010.

[8] E.H. Adelson and J. Bergen, "The plenoptic function and the elements of early vision," in *Computational Models of Visual Processing*, MIT Press, Cambridge, MA, 1991, pp. 3–20.

[9] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, "The lumigraph," in *Proc. SIGGRAPH*, 1996, pp. 43–54.

[10] M. Levoy and P. Hanrahan, "Light field rendering," in *Proc. SIGGRAPH*, 1996, pp. 31–42.

[11] J. Pearson, P.L. Dragotti, and M. Brookes, "Accurate non-iterative depth-layer extraction algorithm for image based rendering," in *Proc. ICASSP*, Prague (Czech Republic), May 2011.