An Object-Based Approach to Plenoptic Videos

Zhi-Feng Gan, Shing-Chow Chan, King-To Ng
Department of Electrical and Electronic Engineering
The University of Hong Kong
Hong Kong
{zfgan, scchan, ktng}@eee.hku.hk

Heung-Yeung Shum
Microsoft Research, Asia
P. R. China
hshum@microsoft.com

Abstract—**This paper proposes an object-based approach to plenoptic videos, where the plenoptic video sequences are segmented into image-based rendering (IBR) objects each with its image sequence, depth map and other relevant information such as shape information. This allows desirable functionalities such as scalability of contents, error resilience, and interactivity with individual IBR objects to be supported. A portable capturing system consisting of two linear camera arrays, each hosting 6 JVC video cameras, was developed to verify the proposed approach. Rendering and compression results of real-world scenes demonstrate the usefulness and good quality of the proposed approach.**

I. Introduction

Image-based rendering (IBR) is an emerging and promising technology for photo-realistic rendering of scenes and objects from a collection of densely sampled images and videos. Since the data size of image-based representations is usually very large, the capturing, storage, transmission and effective rendering are the fundamental problems in IBR research. Interested readers are referred to a recent survey for more information [1]. In [2], a class of dynamic image-based representations called the plenoptic videos is proposed for capturing dynamic scenes. It is based on a simplified light field of dynamic scenes with viewpoints being constrained along a series of line segments instead of a 2D plane as shown in Fig. 5. Hence, it is a four-dimensional plenoptic function [3]. The main motivation of limiting the vertical dimension of the light field is to reduce the hardware complexity of the real-time capturing systems. Plenoptic videos are also closely related to multiview video sequences. However, plenoptic videos usually rely on denser sampling in regular geometric configurations to improve the rendering quality. In addition, the random access to individual pixels in the compressed data stream, so-called the random access problem in IBR, is usually required for rendering intermediate views.

One difficult problem of rendering light fields and plenoptic videos is the excessive artifacts due to depth variations. If the scene is free of occlusions, then the concept of plenoptic sampling [4] can be applied to determine the sampling rate in the camera plane. Unfortunately, because of depth discontinuities around object boundaries, the sampling rate is usually insufficient and significant rendering artifacts due to occlusion are observed. Moreover, appropriate mean depths for objects have to be determined to avoid blurring within the objects and ghosting at the boundaries. Thus, depth segmentation or some kind of depth information is necessary in order to improve the rendering quality. Motivated by Gortler et al's work on lumigraph [5] and the layered depth images of Shade [6], we assume that each image pixel in a light field has a color as well as a depth value. Instead of using a global depth map, this representation, which can be viewed as local depth images between successive cameras, is less sensitive to errors in camera position and depth maps encountered in practical multi-camera systems. Due to the limited amount of information that we can gather from images and videos, a very high-resolution depth map is usually not available. Besides, the data rate of these detailed depth maps sequences is very high. Fortunately, plenoptic sampling tells us that dense sampling of image-based representation will tolerate this variation within the segments by interpolating the plenoptic function. In other words, it is highly desirable to focus on objects with large depth discontinuities. By properly segmenting into objects at different depths, the rendering quality in large environment can be considerably improved using mean depth values [7].

These observations motivate us to develop in this paper an object-based approach to plenoptic videos, where the plenoptic video sequences are segmented into IBR objects, each with its image sequence, depth map and other relevant information such as shape information. Therefore, desirable functionalities such as scalability of contents, error resilience, and interactivity with individual IBR objects can be incorporated. For example, IBR objects can be processed, rendered and transmitted separately. The IBR objects can be obtained by the semi-automatic segmentation tool in [8]. A MPEG-4 like object-based algorithm is also developed for compressing the video texture associated with the alpha maps. A simple method for detecting possible occlusions during rendering and estimating the rendered pixels is also proposed. This can be viewed as a spatially varying reconstruction filter in the frequency domain. Basically, our rendering algorithm explores and observes the physical model and constraints of image formation so that the rendering quality can be improved at lower sampling rate. To verify the proposed approach, a portable plenoptic video system, which consists of two linear arrays each carrying 6 video cameras, for large and dynamic environment scenes was constructed. The rendering and compression results were very satisfactory, which demonstrate the usefulness of the proposed object-based approach.

The paper is organized as follows. The system configuration for the proposed plenoptic video system is described in Section 2. The sampling and reconstruction issues of the plenoptic video are discussed in Section 3. Section 4 is devoted to the rendering of the IBR objects. The proposed object-based compression algorithm is presented in Section 5. Experimental results are presented in Section 6. Finally, conclusions are drawn in Section 7.

II. The Capturing System

Previous attempts to generalize image-based representations to dynamic scenes are mostly based on 2D panoramas. These include the QuickTime VR [9] and panoramic videos [10]. The panoramic video is a sequence of panoramas created at different locations along a path in space, which can be used to capture dynamic scenes at a stationary location or in general along a path with 360 degrees of viewing freedom. The plenoptic video described in this paper is a simplified light field for dynamic environment, where the viewpoints of the user are constrained along line segments

0-7803-8834-8/05/$20.00 ©2005 IEEE.

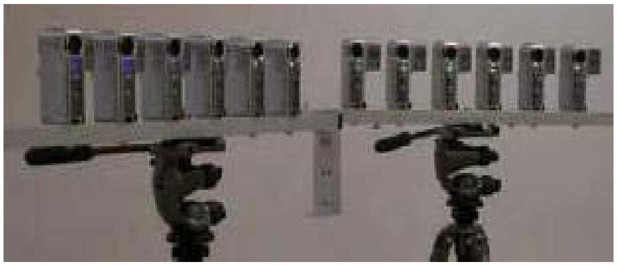

Fig. 1. Two linear camera arrays, each consists of 6 JVC video cameras.

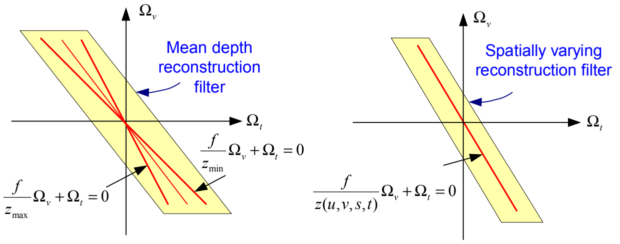
Ω_v
Mean depth reconstruction filter
Ω_t
$\frac{f}{z_{\min}}\Omega_v + \Omega_t = 0$
$\frac{f}{z_{\max}}\Omega_v + \Omega_t = 0$
Ω_v
Spatially varying reconstruction filter
Ω_t
$\frac{f}{z(u,v,s,t)}\Omega_v + \Omega_t = 0$

Fig. 2. Left to right: (a) Mean depth reconstruction filter. (b) Spatially varying reconstruction filter with local mean depth.

instead of a 2D plane in [11]. This greatly reduces the complexity of the dynamic IBR system. However, unlike panoramic videos, users can still observe significant parallax and lighting changes along the line segments. More recently, there were attempts to construct light field video systems for different applications and characteristics. These include the Stanford multi-camera array [12], the 3D rendering system of Naemura *et al*. [13], and the (8×8) light field camera of Yang *et al*. [14]. The Stanford array consists of more than one hundred of cameras and is intended for large environment applications. It uses low cost CMOS sensor and dedicated hardware for real-time compression. The systems in [14] and [15] consist of respectively 16 and 64 cameras and are intended for real-time rendering applications.

Fig. 1 shows the proposed plenoptic video system used to capture dynamic scenes. This system consists of two linear arrays of cameras, each hosting 6 JVC DR-DVP9ah video cameras. The spacing between successive cameras in the two linear arrays is 15cm and the angle between the arrays can be flexibly adjusted. More arrays can be connected together to form longer segments. Because the videos are recorded on tapes, the system is also more portable for capturing outdoor dynamic scenes. Along each linear camera array, a 4D simplified dynamic light field is captured, and the user's viewpoints are constrained along the linear arrays of video cameras. The use of multiple linear arrays allows the user to have more viewing freedom in sport events and other life performance. The proposed system represents a design tradeoff between simplicity and viewing freedom. Other configurations can also be employed. The cameras are calibrated using the method in [16]. In order to use this method to calibrate our camera array, a large reference grid was designed so that it can be seen simultaneously by all the cameras. Using the extracted intrinsic and extrinsic parameters of the cameras, the videos of the cameras can be rectified for rendering. After capturing, the video data stored on the tapes can be transmitted to computers through FireWire interface. All these components are relative inexpensive and they can readily be extended to include more cameras.

III. Sampling and Reconstruction

In plenoptic sampling [4], the number of pictures to render a given scene or the sampling density was studied. For the standard two-plane ray space parameterization, the camera plane and the focal plane are respectively parameterized by the parameters (s,t) and (u,v). Each ray in the parameterization is uniquely determined by the quadruple (u,v,s,t). For fixed values of s and t, we obtain an image taken at a location indexed by (u,v). Interested readers are referred to [5], [11] for more details. Assuming a pinhole camera model, the pixel value observed is the convolution of the plenoptic function $l(v,t)$ with the point spread function. If we know the spectral support of the Fourier transform of $l(v,t)$: $L(\Omega_v, \Omega_t)$, then it is possible to apply the sampling theorem to predict the required sampling density. Assuming that there are no occlusions or depth discontinuities, it was found that the spectral support of $L(\Omega_v, \Omega_t)$ is dependent on the depth of the objects as shown in Fig. 2(a) for a 2D light field. From the figure, it can be seen that objects at a certain depth z will appear as a line in the frequency domain. Thus, if the maximum and minimum depth values are known, a reconstruction filter using the mean depth of the scene can be used to reconstruct the light fields from its sampling and it also defines the sampling density for a given sampling geometry.

This result allows us to determine the sampling rate for proper reconstruction of the light field and to avoid undesirable aliasing effects. Since the Fourier transform of a light field is a global frequency description of the entire light field, it only gives us the frequency components or spectrum in the entire light field, but not its local behavior. For scenes with large depth variations, objects with different depth values will contribute to the entire spectrum. If we window the light field at a particular location (u,v,s,t), and compute its Fourier transform at this location, it will give us its local frequency content. For regions with less depth variations, we would expect a spectrum similar to that shown in Fig. 2(b) with an orientation predicted approximately by plenoptic sampling. A reconstruction filter tailored for this particular mean depth can be used to reconstruct the light field locally. Therefore, the reconstruction filter should be spatially varying and it should depend on the local depth image of the light field. Ideal reconstruction filters with support shown in Fig. 2(a) usually have long filter length and they will cause ringing artifacts in reconstruction. The spatially varying reconstruction allows simpler reconstruction filter such as bilinear interpolation to be used, if the local mean depth of the region is known. Although the quality of rendering will be improved with the amount of depth information or geometric information we have, very accurate depth values are generally not required inside regions with limited depth variations according to plenoptic sampling.

At object boundaries, image pixels cannot be interpolated simply because of a discontinuity generated by the occlusion. Fortunately, the light field looks like a piecewise continuous 2D signal with pixels from foreground covering those from the background. If the objects are near coplanar and if they are at a large distance from the cameras, the magnitude of the discontinuity will be smaller and fewer artifacts will be generated. Therefore, the previous analysis using locally adaptive reconstruction filter and plenoptic sampling can be applied to individual objects except around the boundaries. In sampling theorem, the band-limited signal is reconstructed by interpolation of the data samples. On the other hand, in recovering piecewise continuous signals, the discontinuity has to be identified and interpolation/extrapolation are performed independently on each side of the discontinuity in order to avoid excessive artifacts. In other words, by exploring the structure of the light field or the physical geometry, it is possible to

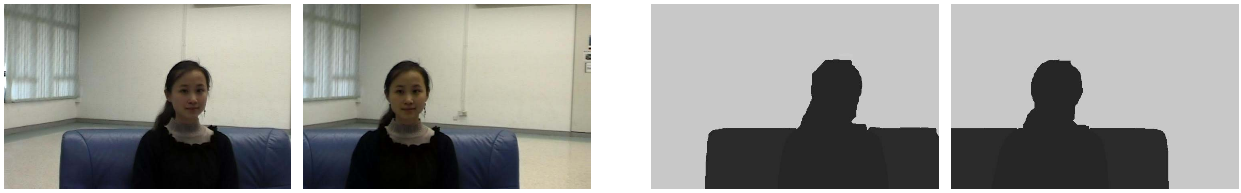

Fig. 3. (a) Snapshots of the sequence *Poem* captured by the proposed system. (b) 8-bit depth maps of a snapshot of this sequence.

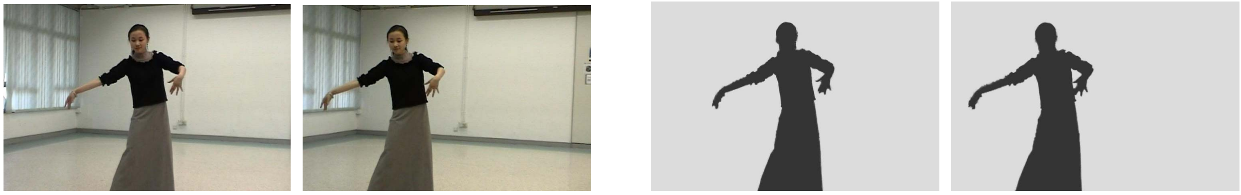

Fig. 4. (a) Snapshots of the sequence *Dance* captured by the proposed system. (b) 8-bit depth maps of a snapshot of this sequence.

reduce the sampling rate in order to get an acceptable reconstruction, provided that depth information, especially the location of the depth discontinuity, can be identified. Therefore, methods for detecting and handling occlusion are important issues. Rendering algorithms using this concept and the pixel-depth representation that we have mentioned earlier will be described in the next section. It is shown that the rendering quality can be significantly improved with additional depth information.

IV. Rendering of IBR Objects

As mentioned earlier, an accurate global depth map is rather difficult to compute. Therefore, we assume that each image pixel in a light field has a color as well as a depth value. This is reasonable, especially when the light field is obtained from a 3D range scanner. The depth information is estimated for each IBR objects after they have been segmented using Lazy snapping [8]. Using this information, it is possible to detect occlusion and interpolate the image pixels during rendering. In [17], a depth matching algorithm for rendering and post-processing of plenoptic video with depth information is proposed. This algorithm brought satisfactory rendering results, but the arithmetic complexity of this algorithm is very high. Here, an improved rendering algorithm with a much lower computational complexity is proposed.

More precisely, instead of finding the depth value of the image pixel to be rendered from the adjacent light field images, the two images are projected using the depth values of each pixel to the current viewing position. Considering the reconstruction of a pixel V in the viewing grid. If the two pixels obtained from projecting the left and right images to the position of pixel V have the same depth values, then there is no occlusion and the value of V can be interpolated from these pixels according to bilinear interpolation. On the other hand, if their depth values differ considerably (say larger than a threshold), then occlusion is said to be occurred. The projected pixel with a small depth value will then occlude the other. Therefore, the value of pixel V should be equal to the one with smaller depth value. Furthermore, if multiple pixels are projected to the location of pixel V, the intensity of pixel V is assigned to the one with smallest depth value. If only one pixel from the left or right image is projected to the position of pixel V, the intensity of pixel V is set to the intensity of this pixel. Finally, due to occlusion, pixel V might not have any projected pixels from adjacent light field images. In this case, we employ the image consistency concept to "guess" the intensity of these pixels [17] from neighboring rendered pixels using interpolation. A linear interpolation from the two image pixels just before and after this occlusion region is performed to cover all undetermined pixels.

V. Object-Based Compression

After the IBR objects have been extracted, they can be compressed individually to provide functionalities such as scalability of contents and interactivity with individual IBR objects. By sharing many useful concepts with the MPEG-4 standard, multiple video streams in the plenoptic video are encoded into user defined IBR objects. Moreover, we exploit both the temporal redundancy and spatial redundancy among video streams in the plenoptic video to achieve higher compression efficiency. In the proposed compression scheme, the texture coding of each IBR object is similar to the modified MPEG-2 algorithm for plenoptic videos [18]. Basically, the algorithm divides the video streams into groups where one of the video streams inside a group is encoded by the MPEG-4 algorithm. It is then used to provide spatial prediction to adjacent secondary video streams, which employs both temporal and spatial predictions to better explore the redundancy in the video streams. Fig. 6 shows the compression results for texture and shape coding in different bit rates achieved by using VM rate control algorithm. The curves denoted by "MPEG-4" represent the results using MPEG-4 without spatial prediction, while the ones denoted by "SP-3" and "SP-5" represent the coding results using the proposed algorithm with spatial prediction using 3 and 5 video streams, respectively. It can be seen that the real-scene IBR object *Dancer* has a considerable improvement in PSNR performance (~1 dB) of the proposed object-based coding scheme over the direct application of the MPEG-4 to individual video object streams. The coding performances of SP-3 are slightly better than that of SP-5 because when the disparity between two video streams increases, spatial prediction becomes less effective. Since the data for each IBR object is self-contained, they can either be rendered individually or together.

VI. Experimental Results

Two plenoptic videos are captured by our system: *Poem* and *Dance*. The resolution of these videos is 720×576. The *Poem* sequence represents a typical head and shoulder scene frequently encountered in 3D videophone, and the *Dance* sequence shows a fast movement scene. Fig. 3(a) and 4(a) show the snapshots of two sequences *Poem* and *Dance* captured by this system. The corresponding depth maps are shown in Fig. 3(b) and Fig. 4(b). Fig. 7 and Fig. 8 show the rendering results obtained by proposed rendering algorithm. In our study, we also compared the execution

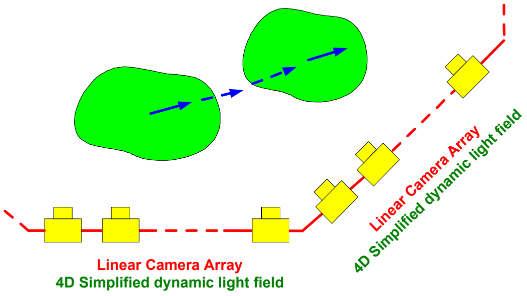
Linear Camera Array
4D Simplified dynamic light field
Linear Camera Array
4D Simplified dynamic light field

Fig. 5. Plenoptic videos: Multiple linear camera array of 4D simplified dynamic light fields with viewpoints constrained along line segments.

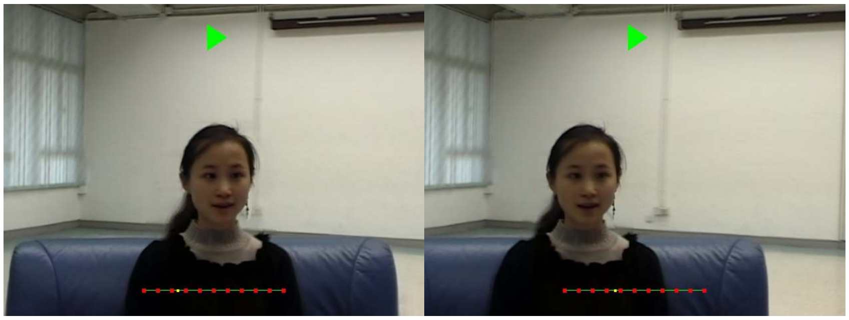

Fig. 7. Rendering results of the sequence *Poem* obtained by the proposed algorithm.

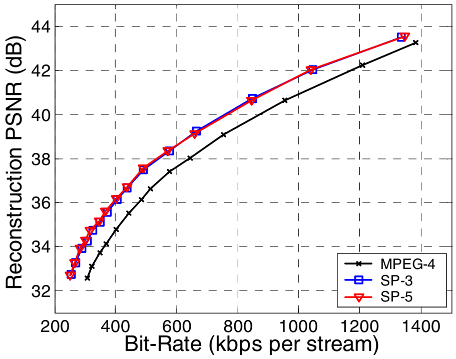

Fig. 6. Coding results for IBR objects *Dance*.

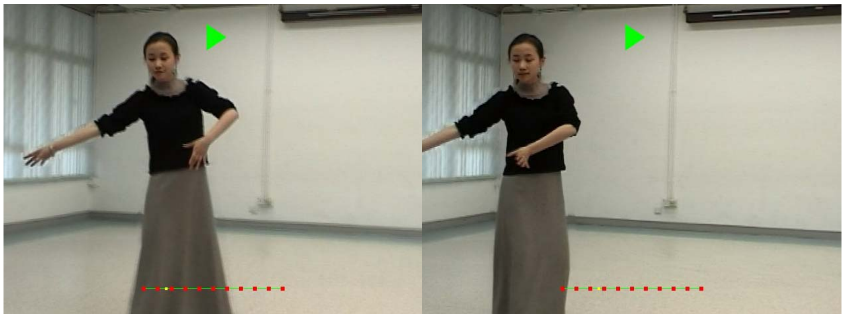

Fig. 8. Rendering results of the sequence *Dance* obtained by the proposed algorithm.

time of the depth matching algorithm [17] and the proposed rendering algorithm. For rendering raw RGB video data in a Pentium 4 2.4 GHz computer, the depth matching algorithm requires on the average 1309.1ms and 2622.6ms to render respectively one frame of the synthetic plenoptic video *Ball* at a resolution of (320×240) [17] and the real-world plenoptic videos at a resolution of (720×576). For the improved algorithm, these two values are reduced respectively to 27.5ms and 128.3ms. Real-time rendering of plenoptic videos can thus be realized.

VII. Conclusion

We have presented an object-based approach to plenoptic videos, where the plenoptic video sequences are segmented into IBR objects each with its image sequence, depth map and other relevant information such as shape information. A portable capturing system consisting of two linear camera arrays, each hosting 6 JVC video cameras, was developed to verify the proposed approach. Rendering and compression results of real-world scenes demonstrate the usefulness and good quality of the proposed approach.

Acknowledgements This work was supported by Hong Kong Research Grant Council.

References

[1] H. Y. Shum, S. B. Kang and S. C. Chan, "Survey of Image-based Representations and Compression Techniques," *IEEE Trans.* CSVT, vol. 13, no. 11, pp. 1020-1037, Nov, 2003.

[2] S. C. Chan, K. T Ng, Z. F. Gan, K. L. Chan and H. Y. Shum, "The Plenoptic Videos: Capturing, Rendering and Compression," in *Proc. of IEEE ISCAS*, vol. 3, pp. 905-908, May, 2004.

[3] E. H. Adelson and J. Bergen, "The Plenoptic Function and The Elements of Early Vision," in *Computational Models of Visual Processing*, pp. 3-20, MIT Press, Cambridge, MA, 1991.

[4] J.X. Chai, X. Tong, S.C. Chan and H.Y. Shum, "Plenoptic sampling," in *Proc. of SIGGRAPH'00*, pp. 307–318, July 2000.

[5] S. J. Gortler, R. Grzeszczuk, R. Szeliski and M. F. Cohen, "The Lumigraph," in *Proc. of SIGGRAPH'96*, pp. 43-54, Aug. 1996.

[6] J. Shade, S. Gortler, L. W. He and R. Szeliski, "Layered depth images," in *Proc. of SIGGRAPH'*98, pp. 231-242.

[7] H. Y. Shum, J. Sun, S. Yamazaki, Y. Lin and C. K. Tang, "Pop-Up Light Field: An Interactive Image-Based Modeling and Rendering System," *ACM Trans*. on Graphics, vol. 23, no. 2, pp. 143-162, 2004.

[8] Y. Li, J. Sun, C. K. Tang and H. Y. Shum, "Lazy snapping," in *Proc. of SIGGRAPH'04*, pp.303-308, 2004.

[9] S. E. Chen, "QuickTime VR–An Image-based Approach to Virtual Environment Navigation," in *Proc. of SIGGRAPH'95*, pp. 29-38, 1995.

[10] K. T. Ng, S. C. Chan, H. Y. Shum and S. B. Kong, "On the Data Compression and Transmission Aspects of Panoramic Video," in *Proc. of IEEE ICIP,* vol. 2, pp. 105-108, Oct. 2001.

[11] M. Levoy and P. Hanrahan, "Light Field Rendering," in *Proc. of SIGGRAPH'96*, pp. 31-42, Aug. 1996.

[12] B. Wilburn et al., "The Light Field Video Camera," in *SPIE Proc. Electronic Imaging: Media Processors*, vol. 4674, Jan. 2002.

[13] T. Naemura, J. Tago and H. Harashima, "Real-time Video-based Modeling and Rendering of 3D Scenes," *IEEE Computer Graphics and Applications*, pp. 66-73, Mar.-Apr. 2002.

[14] J. C. Yang, M. Everett, C. Buehler and L. McMillan, "A Real-time Distributed Light Field Camera," in *Proc. of Eurographics Workshop on Rendering*, pp. 77-86, 2002.

[15] B. Goldlücke, M. Magnor and B. Wilburn, "Hardware-accelerated Dynamic Light Field Rendering," in *Proc. of VMV*, pp.455-462, 2002.

[16] Z. Zhang, "A Flexible New Technique for Camera Calibration," *IEEE Trans. on PAMI*, vol. 22(11), pp. 1330-1334, 2000.

[17] Z. F. Gan, S. C. Chan, K. T. Ng, K. L. Chan and H. Y. Shum, "On The Rendering and Post-Processing of Simplified Dynamic Light Fields with Depth Information," in *Proc. of IEEE ICASSP,* vol. 3, pp. 321-324, May, 2004.

[18] S. C. Chan, K. T Ng, Z. F. Gan, K. L. Chan and H. Y. Shum, "The Compression of Simplified Dynamic Light Fields," in *Proc. of IEEE ICASSP,* vol. 3, pp. 653-656, April, 2003.