Image-Based Rendering of Ancient Chinese Artifacts for Multi-view Displays - a Multi-Camera Approach

Z. Y. Zhu, K. T. Ng , S. C. Chan
Department of Electronic and Electrical Engineering,
The University of Hong Kong.
{zyzhu, ktng, scchan }@eee.hku.hk

H. Y. Shum
Microsoft Corporation
hshum@microsoft.com

***Abstract*—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. This paper proposes an image-based approach to the rendering and multi-view display of ancient Chinese artifacts for cultural heritage preservation. A multiple-camera circular array was constructed to record images of the artifacts. Novel techniques for segmenting and rendering new views of the artifacts from the sampled images are developed. The multiple views so synthesized enable the ancient artifacts to be displayed in modern multi-view displays and conventional stereo systems. Several collections from the University Museum and Art Gallery at the University of Hong Kong are captured and excellent rendering results are obtained.**

I. INTRODUCTION

Image-based rendering/representation (IBR) [1-12,14,17] is an emerging and promising technology for rendering new views of scenes from a collection of densely sampled images or videos. It has potential applications in virtual reality, immersive, advanced visualization and 3D television systems. There has been considerably progress in these areas since the pioneer work of lumigraph [2] and lightfield [3]. Other important IBR representations include the 2D panorama [5], plenoptic modeling [4], the 3D concentric mosaics [7], ray-space representation [8], etc. Motivated by lightfields and lumigraphs, the authors have developed a real-time system for capturing and rendering a simplified dynamic lightfield called the "plenoptic videos (PVs)" [11,12] where videos are taken along line segments instead of a 2D plane to simplify the capturing hardware for dynamic scenes. Please refer to [14,17] for a recent tutorial to the signal processing aspects of IBR.

While there has been considerable progress recently in the capturing, storage and compression of IBR, it is somewhat difficult to fully explore the exceptional experience offered by image-based rendering, since conventional TV displays can only display a single rather than multiple views. With the advance of technology, multi-view displays are becoming available [18] and their costs have been reducing dramatically. It is predicted that 3D or multi-view TVs will be another wave after high-definition TVs. This breakthrough motivates us to study in this paper the important problem of cultural heritage preservation and dissemination of ancient Chinese artifacts using the image-based approach.

Pioneer projects in cultural heritage preservation of large scale structure and sculptures includes the Digital Michelangelo Project [19], the 3D facial reconstruction and visualization of ancient Egyptian mummies [20], the great Buddha Project [21], to name just a few. To avoid possible damages to the artifacts and speed up the capturing process, we propose to employ the image-based approach instead of using 3D laser scanners. A circular array consisting of multiple digital still cameras (DSCs) was therefore constructed in this work. Using this circular camera array, we developed novel techniques for rendering new views of the artifacts from the images captured using the object-based approach [12]. The multiple views so synthesized enable the ancient artifacts to be displayed in modern multi-view displays. A number of ancient Chinese artifacts from the University Museum and Art Gallery at the University of Hong Kong were captured and excellent rendering results are obtained.

In summary, the main contributions of this paper are: 1) the development of a systematic image-based approach and associate algorithms for rendering and display of ancient Chinese artifacts on ordinary as well as modern multi-view TVs and 2) the construction of a multi-camera array to demonstrate the excellent quality of the proposed approach. We wish the present work can serve as a framework for rendering and multi-view display of ancient artifacts so as to facilitate the preservation and dissemination of cultural artifacts using image-based technology. The organization of the paper is as follows: in Section II, the principle of the proposed system is briefly reviewed. The system construction, depth estimation and rendering algorithms are described in Section III. Experimental results are presented in Section IV and finally conclusions are drawn in Section V.

II. THE PROPOSED OBJECT-BASED APPROACH

Central to IBR is the plenoptic function [1], which describes all the radiant energy that can be perceived by the observer at any point in space and time. The plenoptic function is an 7-dimensional function of the viewing position, the azimuth and elevation angle, time, and wavelengths. Traditional images and videos are 2D and 3D special cases of the plenoptic function. Depending on the functionality required, there is a spectrum of image-based representations. They differ from each other in the amount of geometry

978-1-4244-5309-2/10/$26.00 ©2010 IEEE

information being used. At one end of the spectrum, like traditional texture mapping in computer graphics, we have very accurate geometric models of the objects in the scenes, but only a few images are required to generate the textures. At the other extreme, light-field or lumigraph [4] rendering relies on dense sampling [4] and very little geometry information such as depth maps for rendering. An important advantage of the latter is its superior image quality and simplicity, compared with 3D model building for complicated real world scenes. Thus, lumigraph or lightfield based representations are ideal for rendering new views, whereas representations with more geometry information are required for more sophisticated operations such as relighting [10].

In this paper, we shall employ the pop-up lightfield [10] or object-based plenoptic videos [12] to synthesize new views of the artifacts. In pop-up lightfields and plenoptic videos, images or videos at cameras located along multiple linear arrays are captured in order to render intermediate or novel views of the scene at other positions. In [12], two linear arrays were used and each array contains 6 JVC DR-DVP9ah video cameras. More arrays can be connected together to form longer segments. To reduce rendering artifacts, the objects are extracted using semi-automatic segmentation and tracking techniques [12]. The operations are illustrated in Fig. 1. From the segmented objects, approximate depth information for each IBR object can be estimated to render new views at different viewpoints. An important advantage of the object-based approach is that natural matting can be adopted to improve the rendering quality when mixing IBR objects on other backgrounds.

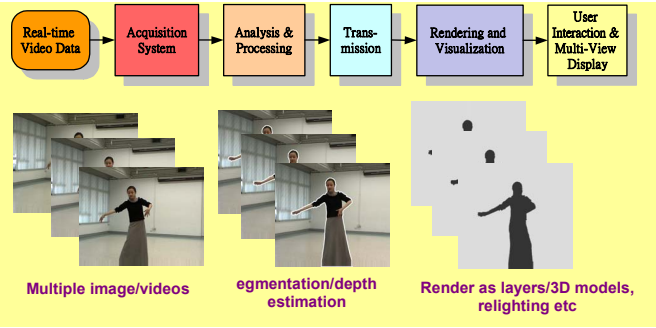

Figure 1. Application of IBR in intermediate view synthesis for multiview TVs.

The main challenge in these approaches is to estimate the depth map. For distance objects, an approximate depth map for each object is sufficient for good renderings as long as the depth discontinuities are well delineated. For objects with fine details such as ancient Chinese artifacts, a more accuracy geometry in form of depth maps is more desirable. Depth estimation using stereo techniques [23,24] is a long standing problem in computer vision and there has been much advanced over the last few years. Techniques using graph cut [27], belief propagation (BF) [22], etc are available. For general scenes with lot of occlusion, reliable depth estimation still poses much difficulty. Fortunately, for capturing ancient artifacts with small to medium sizes, the problem can be considerably simplified as blue screen techniques can be employed. In the following sections, the detailed construction of our system and other technical issues such as camera calibration, object segmentation, depth estimation, and rendering of the ancient Chinese artifacts will be briefly described.

III. SYSTEM CONSTRUCTION AND ALGORITHMS

Fig. 2. Circular camera array constructed.

Figure 2 shows the circular camera array that we have constructed. It consists of 13 canon 450D digital still cameras with an angular spacing of 3 degrees and a radius of 3 meters. The whole array is supported on a tripod for ease of transportation. Before the cameras can be used for depth estimation and 3D reconstruction, they must be calibrated to determine the intrinsic parameters as well as their extrinsic parameters, i.e. their relative positions and poses. This can be accomplished by using a sufficient large checkerboard calibration pattern. We follow the plane-based calibration method to determine the projective matrix of each camera, which connects the world coordinate and the image coordinate. The projection matrix of a camera allows a 3D point in the world coordinate be translated back to the corresponding 2D coordinate in the image captured by that camera.

A. SEGMENTATION

After preprocessing of the captured images to account for differences in color response etc of the cameras, the object is segmented to facilitate rendering. In the plenoptic video systems that we have developed in [12], an initial segmentation of the object is obtained by the semi-automatic segmentation technique Lazy snapping [13]. This serves as a prior information for level set methods [15,16] to extract the objects in other images. In this work, we employ the photometric invariant features [25] to extract the foreground from the monochromatic screen background. More precisely, the color tensor describes the local orientation of color vector $\boldsymbol{f}(x,y)$ as: $\boldsymbol{T}(x,y) = \begin{bmatrix} \boldsymbol{f}_x^T \boldsymbol{f}_x & \boldsymbol{f}_x^T \boldsymbol{f}_y \\ \boldsymbol{f}_y^T \boldsymbol{f}_x & \boldsymbol{f}_y^T \boldsymbol{f}_y \end{bmatrix}$, where $\boldsymbol{f}(x,y)$ is a vector which contains the *RGB* values at position (x,y) and the subscripts x and y in $\boldsymbol{f}_x(x,y)$ and $\boldsymbol{f}_y(x,y)$ denote respectively the derivative of $\boldsymbol{f}(x,y)$ with respect to x and y, the image coordinates. According to [25], the *RGB* vector $[R,G,B]^T$ can be seen as a weighted sum of two component vectors: $(R,B,G) = e(m_b \boldsymbol{c}_b + m_i \boldsymbol{c}_i)$ where $\boldsymbol{c}_b$ is the color vector of the body reflectance, $\boldsymbol{c}_i$ is the color vector of the interface reflectance (i.e. specularities or highlights), m_b and

m_i are scalars representing the corresponding magnitudes of reflection and e is the intensity of the light source.

Fig. 3 (a). Extraction results using color-tensor-based method. Left: original, middle: hard segmentation, right: after matting.

(b) Close up. Upper: hard segmentation Lower: after matting.

Thus $[R,B,G]_x^T = em_b(\boldsymbol{c}_b)_x + (e_x m_b + e(m_b)_x)\boldsymbol{c}_b + (e(m_i)_x + e_x m_i)\boldsymbol{c}_i$ which suggests that the spatial derivative is a sum of three weighted vectors, successively caused by body reflectance, shading-shadow and specular changes. For matte surfaces, the intensity of interface reflectance is zero (i.e. m_i=0) and the projection of the spatial derivative $\boldsymbol{f}_x$ on the shadow-shading axis is the shadow-shading variant containing all energy which can be explained by changes due to shadow and shading. The shadow-shading axis direction is $\boldsymbol{c}_b$ which is parallel to $\boldsymbol{f} = em_b\boldsymbol{c}_b$ for matte surfaces. So the projection $\boldsymbol{s}_1$ of the spatial derivative $\boldsymbol{f}_x$ on the shadow-shading axis is $\boldsymbol{s}_1 = (\boldsymbol{f}_x^T \boldsymbol{f} / \| \boldsymbol{f} \|) \cdot \boldsymbol{f} / \| \boldsymbol{f} \|$. Subtraction of the shadow-shading variant $\boldsymbol{s}_1$ from the total derivative $\boldsymbol{f}_x$ results in the shadow-shading quasi-invariant $\boldsymbol{s}_2 = \boldsymbol{f}_x - \boldsymbol{s}_1$ .In summary, the derivative of the color tensor can be separated into shadow-shading variant part $\boldsymbol{s}_1$ and shadow-shading invariant part $\boldsymbol{s}_2$. The shadow-shading invariant part does not contain the derivative energy caused by shadows and shading. To construct a shadow-shading-specular quasi-invariant, this part is combined with the hue direction, which is perpendicular to the light source direction $\boldsymbol{c}_i$ and the shadow and shading direction $\boldsymbol{c}_b$. Therefore the hue direction is $\boldsymbol{h} = (\boldsymbol{c}_i \times \boldsymbol{c}_b) / |\boldsymbol{c}_i \times \boldsymbol{c}_b|$. The projection of the derivative on the hue direction is the desired shadow-shading-specular quasi-invariant part: $\boldsymbol{H} = (\boldsymbol{f}_x^T \boldsymbol{h} / \| \boldsymbol{h} \|) \cdot \boldsymbol{h} / \| \boldsymbol{h} \|$. By replacing $\boldsymbol{f}_x$ in the color tensor equation $\boldsymbol{T}(x,y) = \begin{bmatrix} \boldsymbol{f}_x^T \boldsymbol{f}_x & \boldsymbol{f}_x^T \boldsymbol{f}_y \\ \boldsymbol{f}_y^T \boldsymbol{f}_x & \boldsymbol{f}_y^T \boldsymbol{f}_y \end{bmatrix}$ as $\boldsymbol{s}_2$ or $\boldsymbol{H}$, we can get the shadow-shading-specular-quasi-invariant color tensor and the shadow-shading invariant color tensor respectively. By setting a suitable threshold value for the color tensor, we can detect the boundary of the object. Figure 3 shows some segmentation results that were obtained using the color tensor method, followed by Bayesian matting for extracting a foreground from the background. After segmentation, the hard boundary of the object will be obtained (see Figure 3(a)). Matting can then be applied to obtain a soft segmentation information, called the matte, of the object (see Figure 3(b)). The matte, which is an image containing the portion of foreground with respect to the background (from 0 to 1) at a particular location, greatly improves the visual quality of mixing the objects onto other backgrounds.

B. DEPTH ESTIMATION AND RENDERING

Stereo matching, which infers 3D scene geometry from two images, is an active research area in computer vision. It can be used to improve the rendering quality when synthesizing intermediate views in our image-based rendering system. Recently, Scharstein and Szeliski [23] gave an extensive survey on stereo algorithms and provided an online evaluation based on Middlebury Stereo Evaluation (MSE) data set [24]. Since then, many new and novel approaches to stereo matching algorithms were developed and evaluated online with MSE data set. Global optimization techniques like Graph Cut (GC) [27], Belief Propagation (BP) [22] and Tensor Voting are widely used in top rank methods. In computing the depth maps, we used the squared intensity differences as cost function, and aggregated the cost in a square window weighted by color similarity and geometric proximity as in Yoon's [26] method. Disparity map is first estimated by pyramid Lucas-Kanade (LK) feature tracking algorithm, which minimizes the cost/energy by least-square method. Instead of defining smoothness term in the energy function, the disparity map is anisotropic diffused after LK method. Finally, a symmetric stereo model is introduced for occlusion detection and optimized with Belief-Propagation (BP). Once the alpha and depth maps have been estimated, virtual views can be synthesized. Possible holes are filled by inpainting.

IV. EXPERIMENTAL RESULTS

Using the circular camera array that we constructed in Fig. 2, we have captured 10 ancient Chinese artifacts from the University Museum and Art Gallery at the University of Hong Kong. Fig. 4 shows several examples of the artifacts captured, their segmented outputs, depth maps, and example renderings. The artifacts are also displayed in a Newsight multiview TV which can display 9 views at a time. Using a Graphic processing unit (GPU) running on a NVIDIA GTX260+ graphic card, the rendering and display filtering algorithms are accelerated and interactive rendering at a speed of 2 frames

per second are obtained. Figure 4 shows the segmented images, depth maps and typical renderings with another background of example ancient Chinese Artifacts.

Fig. 4. Left: Typical depth maps of the artifacts computed; right: example renderings on another background.

V. CONCLUSION

An image-based approach to the rendering of ancient Chinese artifacts for cultural heritage preservation is presented. A 13-camera circular array was constructed and novel rendering algorithms were developed to synthesize virtual views of the artifacts from a set of densely sampled images. The artifacts were displayed in a Newsight multi-view TVs interactively with the help of GPU acceleration and excellent rendering results are obtained.

Acknowledgment: This project is supported in parts by the HK RGC GRF grant and the HK Government ITF tier 3 grant on 3D photography and video technology. The generous support of the University Museum and Art Gallery at the University of Hong Kong and the excellent technical support by Mr. James L. C. Koo are gratefully acknowledged.

REFERENCES

[1] 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.

[2] S. J. Gortler, R. Grzeszczuk, R. Szeliski and M. F. Cohen, "The lumigraph," in Proc. of the annual conference on Computer Graphics (SIGGRAPH'96), pp. 43-54, Aug. 1996.

[3] M. Levoy and P. Hanrahan, "Light field rendering," in Proc. of the annual conference on Computer Graphics (SIGGRAPH'96), pp. 31-42, Aug. 1996.

[4] L. McMillan and G. Bishop, "Plenoptic modeling: An image-based rendering system," in Proc. of the annual conference on Computer Graphics (SIGGRAPH'95), pp. 39-46, Aug. 1995.

[5] R. Szeliski and H. Y. Shum, "Creating full view panoramic image mosaics and environment maps," in Proc. of the annual conference on Computer Graphics (SIGGRAPH'97), pp. 251-258, Aug. 1997.

[6] J. Shade, S. Gortler, L.-W. He, and R. Szeliski. Layered depth images. In Computer Graphics (SIGGRAPH'98) Proceedings, pages 231–242, Orlando, July 1998. ACM SIGGRAPH.

[7] H. Y. Shum and L. W. He, "Rendering with concentric mosaics," in Proc. of the annual conference on Computer Graphics (SIGGRAPH'99), pp. 299 - 306, Aug. 1999.

[8] T. Fujii, T. Kimoto, and M. Tanimoto, "Ray space coding for 3D visual communication," in *Proc. Picture Coding Symp. '96*, Mar. 1996, pp. 447–451.

[9] K. Zhou, Y. Hu, S. Lin, B. Guo, and H. Y. Shum, "Precomputed shadow fields for dynamic scenes," in SIGGRAPH'05.

[10] 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, issue 2, pp. 143 -162, April 2004.

[11] S. C. Chan, K. T Ng, Z. F. Gan, K. L. Chan and H. Y. Shum, "The plenoptic videos," IEEE Trans. on Circuits and Systems for Video Technology, vol. 15, no. 12, pp. 1650-2659, Dec., 2005.

[12] S. C. Chan, Z. F. Gan, K. T. Ng and H. Y. Shum, "An Object-Based Approach to Image/Video-Based Synthesis and Processing for 3-D and Multiview Televisions," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 19, no. 6, pp. 821-831, June 2009.

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

[14] H. Y. Shum, S. C. Chan and S. B. Kang. Image-based rendering. Springer, 2007.

[15] S. Osher and N. Paragios. Geometric Level Set Methods in Imaging, Vision, and Graphics. Springer Verlag, 2003.

[16] T. F. Chan and L. A. Vese, "Active Contours Without Edges," IEEE Trans. Image Processing, vol. 10, no. 2, pp. 266-277, 2001.

[17] S. C. Chan, H. Y. Shum, and K. T. Ng, "Image-based rendering and synthesis: technological advances and challenges," IEEE Signal Processing Magazine: Special Issue on MVI and 3DTV, Nov, 2007.

[18] J. Konrad, and M. Halle, "3-D Displays and signal processing," IEEE Signal Processing magazine, vol. 24, no. 6, pp. 97-111, Nov. 2007.

[19] M. Levoy et al, "The digital Michelangelo project: 3D scanning of large statues, Proc. Siggraph 2000, pp. 131-144.

[20] G. Attardi, M. Betrò, M. Forte, R. Gori, A. Guidazzoli, S. Imboden, and F. Mallegni, "3D facial reconstruction and visualization of ancient Egyptian mummies using spiral CT data : Soft tissues reconstruction and textures application," in Proc. Siggraph 1999.

[21] K. Ikeuchi, A. Nakazawa, K. Hasegawa, and T. Ohishi, "The Great Buddha Project: Modeling Cultural Heritage for VR Systems through Observation Proceedings of the 2nd IEEE/ACM International Symposium on Mixed and Augmented Reality, 2003.

[22] J. Sun, N. Zheng, and H. Y. Shum, "Stereo Matching Using Belief Propagation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 7, pp. 787-800, Jul., 2003

[23] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms,"International Journal of Computer Vision, vol. 47, no. 1/2/3, pp.7-42, 2002.

[24] http://www.middlebury.edu/stereo/

[25] J. van de Weijer, T. Gevers, and A. D. Bagdanov, "Boosting color saliency in image feature detection," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, pp. 150-156, Jan 2006.

[26] K. J. Yoon and I. S. Kweon, "Adaptive support-weight approach for correspondence search," IEEE PAMI, 28(4), pp. 650-656, 2006.

[27] V. Kolmogorov and R. Zabih, "Multi-camera scene reconstruction via graph cuts," in Proc. of ECCV, pp. 82-96,2002.