Title	A multi-camera approach to image-based rendering and 3-D/Multiview display of ancient chinese artifacts
Author(s)	Ng, KT; Zhu, Z; Wang, C; Chan, SC; Shum, HY
Citation	IEEE Transactions on Multimedia, 2012, v. 14 n. 6, p. 1631-1641
Issued Date	2012
URL	http://hdl.handle.net/10722/169228
Rights	IEEE Transactions on Multimedia. Copyright © IEEE.

A Multi-Camera Approach to Image-Based Rendering and 3-D/Multiview Display of Ancient Chinese Artifacts

King-To Ng, *Member, IEEE*, Zhen-Yu Zhu, Chong Wang, Shing-Chow Chan, *Member, IEEE*, and Heung-Yeung Shum, *Fellow, IEEE*

***Abstract*—This paper proposes an image-based approach for the capturing, rendering and display of ancient Chinese artifacts for cultural heritage preservation. A multiple-camera circular array is proposed to record images of the artifacts, which forms a simplified circular light field (SCLF). A systematic image-based approach and associate algorithms such as segmentation, depth estimation and shape morphing are developed for rendering new views of the Chinese artifacts. An object-based compression scheme is also proposed to reduce the data size for storage and transmission of the texture, depth maps and alpha maps associated with the object-based circular light field. Spatial redundancies among the various images are exploited to improve the coding performance, while avoiding excessive complexity in selective decoding of the light field to support fast rendering speed. To allow the Chinese artifacts to be viewed over the internet, scalable prioritized transmission and rendering schemes of the SCLF with low latency were also developed. The multiple views so synthesized enable the ancient artifacts to be displayed in 3-D/multi-view displays. Several collections from the University Museum and Art Gallery at The University of Hong Kong were captured and excellent rendering results are obtained.**

***Index Terms*—IBR object coding, image-based rendering, object-based coding, plenoptic videos.**

I. Introduction

IMAGE-BASED rendering/representation (IBR) [1]–[14] is a 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 3-D television systems. There has been considerably progress in these areas since the pioneer work of light-field [9] and lumigraph [10]. Other important IBR representations include the 2-D panorama [7], plenoptic modeling [8], the 3-D concentric mosaics [11], ray-space representation [15], etc. Interested readers are referred to [2] and [12] for more details. IBR is also closely related to multi-view videos [14]–[18] and free-viewpoint videos [21], but it puts much emphasis on intermediate view synthesis and the associated processing of the video data.

In previous works [4] and [5], the authors have developed a multiple cameras system for capturing a class of dynamic image-based representation called "plenoptic videos" (PVs). It is a simplified light field for dynamic environment where a linear array of video cameras is used to simplify the hardware requirement. Moreover, the simplified and regular camera geometry allows a continuum of virtual views to be synthesized along the line segments joining the video cameras. In [22], an object-based approach was proposed by the investigators to reduce rendering artifacts by extracting objects with large disparities using semi-automatic segmentation and tracking technique. From the segmented objects, approximated depth information for each object can be estimated so as to render virtual 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 IBR objects are mixed on different backgrounds.

While there has been considerable progress recently in the capturing, storage, and compression of IBR [1], [15], [16], [23]–[40], 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, 3-D and multi-view displays are becoming popular [41] and their costs have been reducing dramatically. This breakthrough motivates us to study 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 [42], the 3-D facial reconstruction and visualization of ancient Egyptian mummies [43], the great Buddha Project [44], to name just a few. To avoid possible damage to the ancient artifacts and speed up the capturing process, we propose to employ the image-based approach instead of using 3-D laser scanners. A circular array consisting of multiple digital still cameras (DSCs) was therefore constructed in this work to capture the simplified light field of the ancient artifacts along circular arcs, which we shall call the simplified circular light field (SCLF) or circular light field (CLF) in short. The circular array is chosen to provide users with a better visual experience,

Manuscript received September 26, 2011; revised January 14, 2012, April 25, 2012, and May 04, 2012; accepted May 04, 2012. Date of publication May 14, 2012; date of current version November 22, 2012. Part of this work was presented at IEEE ISCAS 2010 [40] and IEEE APCCAS 2010 [66]. This work was supported in part by Hong Kong Research Grant Council (RGC) and the Innovation and Technology fund (ITF). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Monica Aguilar.

K.-T. Ng, Z.-Y. Zhu, C. Wang, and S.-C. Chan are with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong (e-mail: ktng@eee.hku.hk; zyzhu@eee.hku.hk; cwang@eee.hku.hk; scchan@eee.hku.hk).

H.-Y. Shum is with the Microsoft Corporation, USA (e-mail: hshum@microsoft.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2012.2199291

because it supports fly over effect and close-up of the artifacts uniformly in the angular domain. We also developed novel techniques for rendering new views of the ancient artifacts from the images captured using the object-based approach. In particular, we developed a novel mutual information-based shape morphing technique to better preserve the boundaries of the object during intermediate view synthesis. The multiple views so synthesized enable the ancient artifacts to be displayed in modern 3-D and multi-view displays. To reduce the data size of the SCLF, an object-based compression scheme is developed for their efficient storage and transmission. In the proposed coding scheme, in addition to the image texture, object-based information such as shape, grayscale alpha maps (for matting) and depth information of the object is also coded to facilitate rendering. This coding scheme may be viewed as a modification of our previous object-based compression technique [4], [5], [45] for PVs to the SCLF setting. Spatial redundancy among the various images is exploited to improve the coding performance, while avoiding excessive complexity in selective decoding of the light field to support fast rendering speed. To further reduce the latency in launching the application and improve the response time, a prioritized transmission scheme using scalable coding technique is proposed. In particular, adjacent light field images are grouped together and coded separately. Therefore, rendering at a certain virtual view angle can begin when the associated compressed data of the group is received. Moreover, by encoding the SCLFs into a lower-resolution base layer and an enhancement layer, the startup time can be greatly reduced compared by transmitting the compressed data using conventional methods. The rendering quality is refined progressively with the reception of the enhancement layer. In the proposed system, the Audio and Video Coding Standard Workgroup of China (AVS) compression algorithm [46] is used for compressing the low-resolution base layer because of its good performance and low complexity. The enhancement layer is compressed using the JPEG-2000 wavelet-based algorithm [47]. 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 in ordinary as well as 3-D/multiview displays are achieved. Please note that we have reported preliminary results of the proposed system in [66]. The main differences from this work are 1) the AVS compression algorithm is used for compression because of its low complexity and good performance, 2) a new shape morphing processing is introduced in order to improve the rendering results, and 3) a new prioritized transmission scheme has been proposed which can support any viewing position of the users as opposed to [66]. Moreover, more rendering, compression and streaming results are included to illustrate the effectiveness of the proposed approach.

In summary, the main contributions of this paper are: 1) the construction of a multi-camera array to capture SCLFs of ancient Chinese artifacts; 2) the development of a systematic image-based approach and associate algorithms such as segmentation, depth estimation, and shape morphing for rendering the Chinese artifacts; 3) the development of a scalable compression algorithm for prioritized transmission and rendering of the simplified circular light field with low latency; and 4) the demonstration of the excellent rendering results of the proposed approach by displaying these ancient Chinese artifacts on ordinary as well as modern 3-D/multi-view TVs. A demonstration video of the proposed algorithm can be found at http://youtu.be/uK1kko7dnMI. 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 the image-based technology.

The organization of the paper is as following: in Section II, the principle of the proposed system is briefly reviewed. The system construction and associated algorithms such as segmentation, depth estimation, sharp morphing and rendering algorithms are described in Section III. This is followed by the scalable compression algorithm for prioritized transmission and rendering of the SCLFs in Section IV. Experimental results are presented at each section to illustrate the design flow and finally conclusions are drawn in Section V.

II. Proposed Object-Based Approach

Central to IBR is the plenoptic function [6], which describes all the radiant energy that can be perceived by the observer at any point in space and time. The plenoptic function is a 7-dimensional function of the viewing position, the azimuth and elevation angles, time, and wavelengths. Traditional images and videos are 2-D and 3-D 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 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 [9], [10] rendering relies on dense sampling 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 3-D 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 and editing.

In this paper, we shall employ the pop-up lightfield [13] or object-based plenoptic videos [14] approach to synthesize new views of the ancient 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 [14], 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. The main advantage of the object-based representation is that by properly segmenting data into objects at different depths, it has been shown that the rendering quality in large environment can be significantly improved. Moreover, by coding plenoptic videos at the object level, desirable functionalities such as scalability of contents, error resilience, and interactivity with individual objects can be achieved. In [14], the videos captured by the cameras are first rectified using the extracted intrinsic and extrinsic parameters of the cameras for further processing. A semi-automatic

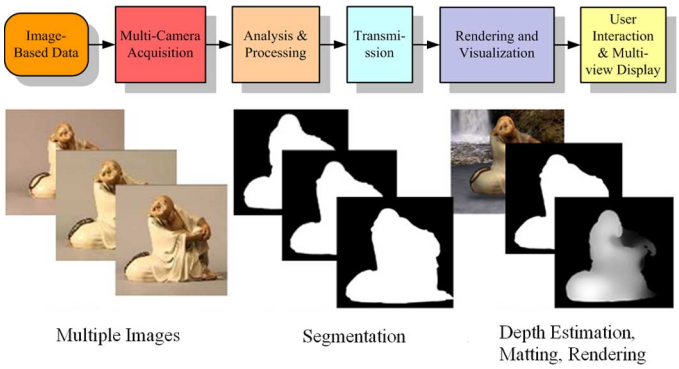


Fig. 1. Application of IBR in intermediate view synthesis for multiview TVs.

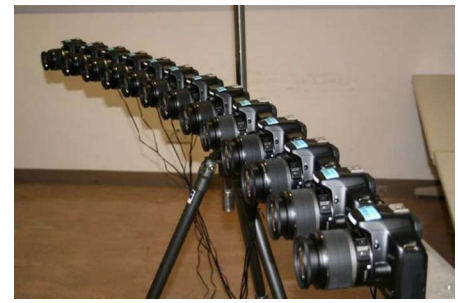


Fig. 2. Circular camera array constructed.

segmentation method called "lazy snapping" [48] is used to segment the object at a particular time instant and view. The objects at other views and time instants are then extracted using semi-automatic segmentation and tracking techniques [14], [22]. The operations are illustrated in Fig. 1. From the segmented objects, approximated depth information for each object can be estimated for rendering new views at different viewpoints. Another important advantage of the object-based approach is that natural matting can be adopted to improve the rendering quality when objects are mixed on other backgrounds. The depth maps and shape information of these IBR objects will be used for virtual view synthesis at the decoder. Therefore, an object will consist of texture, depth, shape and matting information.

One of the main challenges in these approaches is to estimate the depth map. For distance objects, an approximated depth map for each object is sufficient for good rendering as long as the depth discontinuities are well delineated. For objects with fine details such as ancient Chinese artifacts, a more accurate geometry in form of depth maps is more desirable. Depth estimation using stereo techniques is a long standing problem in computer vision and there has been much advancement over the last few years. Techniques using graph cuts [49], belief propagation [50], etc. are available. For general scenes with a 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 shows the circular camera array that we have constructed. It consists of 13 canon 450D digital still cameras with an angular spacing of 3° and a radius of 3 m. The whole array is supported on a tripod for ease of transportation. Before the cameras can be used for depth estimation and 3-D reconstruction, they must be calibrated to determine their intrinsic parameters as well as 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 [51] 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 3-D point in the world coordinate to be translated back to the corresponding 2-D coordinate in the image captured by that camera.

A. Segmentation

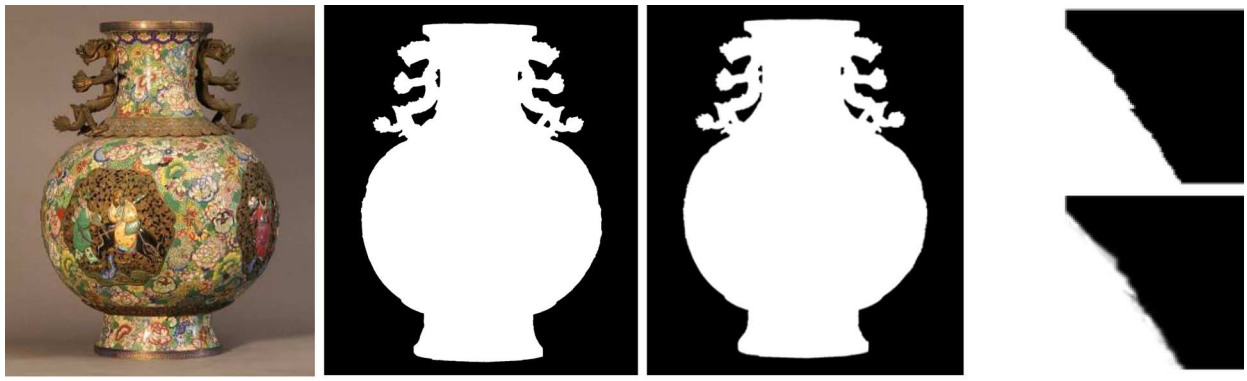
After preprocessing of the captured images to account for differences in color response of the cameras, the object is segmented to facilitate rendering. In the plenoptic video system that we have developed in [14], an initial segmentation of the object is obtained by the semi-automatic segmentation technique lazy snapping [48]. This serves as prior information for level set methods [52], [53] to extract the objects in other images. In this work, we employ the photometric invariant features [54] to extract the foreground from the monochromatic screen background. More precisely, the color tensor describes the local orientation of a color vector $\boldsymbol{f}(x,y)$ as

$$\boldsymbol{T}(x,y) = \begin{bmatrix} \boldsymbol{f}_x^T \boldsymbol{f}_x & \boldsymbol{f}_x^T \boldsymbol{f}_y \\ \boldsymbol{f}_y^T \boldsymbol{f}_x & \boldsymbol{f}_y^T \boldsymbol{f}_y \end{bmatrix} \tag{1}$$

where $\boldsymbol{f}(x,y)$ is a vector which contains the color component 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 [54], the color vector can be seen as a weighted sum of two component vectors: $[R,G,B]^{\mathrm{T}} = e(m_b\mathbf{c}_b + m_i\mathbf{c}_i)$ where $\boldsymbol{c}_b$ is the color vector of the body reflectance, $\mathbf{c}_i$ 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 the light source. Thus

$$[R,B,G]_x^T = em_b(\boldsymbol{c}_b)_x + (e_x m_b + e(m_b)_x)\boldsymbol{c}_b + (e(m_i)_x + e_x m_i)\boldsymbol{c}_i \tag{2}$$

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



(a) (b)

Fig. 3. (a). Extraction results using color-tensor-based method. Left: original, middle: hard segmentation, right: after matting. (b) Close up of segmentations in (a). Upper: hard segmentation, lower: after matting.

$\boldsymbol{f} = em_b\boldsymbol{c}_b$ for matte surfaces. So the projection $\boldsymbol{s}_1$ of the spatial derivative $\boldsymbol{f}_x$ on the shadow-shading axis is

$$\boldsymbol{s}_1 = \left(\frac{\boldsymbol{f}_x^T\boldsymbol{f}}{\|\boldsymbol{f}\|}\right)\cdot\frac{\boldsymbol{f}}{\|\boldsymbol{f}\|}. \tag{3}$$

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

$$\boldsymbol{h} = \frac{(\boldsymbol{c}_i \times \boldsymbol{c}_b)}{|\boldsymbol{c}_i \times \boldsymbol{c}_b|}. \tag{4}$$

The projection of the derivative on the hue direction is the desired shadow-shading-specular quasi-invariant part:

$$\boldsymbol{H} = \left(\frac{\boldsymbol{f}_x^T\boldsymbol{h}}{\|\boldsymbol{h}\|}\right)\cdot\frac{\boldsymbol{h}}{\|\boldsymbol{h}\|}. \tag{5}$$

By replacing $\boldsymbol{f}_x$ in the color tensor (1) by $\boldsymbol{s}_2$ 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. Fig. 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. Matting can then be applied to obtain soft segmentation information, called the matte, of the object. 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

Stereo matching, which estimates 3-D 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 the image-based rendering system. In [55], Scharstein and Szeliski gave an extensive survey on stereo algorithms and provided an online evaluation based on the Middlebury Stereo Evaluation (MSE) data set. Since then, many new and novel approaches to stereo matching algorithms were developed and evaluated online with the MSE data set. Global optimization techniques like graph cuts [49], belief propagation [50], and tensor voting are widely used in top ranked methods.

In computing the depth maps, we used the squared intensity differences as a cost function, and aggregated the cost in a square window weighted by color similarity and geometric proximity as in Yoon's [56] method. Disparity map is first estimated by the pyramid Lucas-Kanade (LK) feature tracking algorithm, which minimizes the cost/energy by least-square methods. Instead of defining a smoothness term in the energy function, the disparity map is anisotropic diffused after the LK method. Finally, a symmetric stereo model is introduced for occlusion detection and optimized with belief propagation. Fig. 4 shows the typical depth maps and example rendering results of the artifacts computed. If depth capturing devices, such as Kinect, are available, it may be used as initial guess of the depth estimation algorithm to improve the accuracy and computation speed of the algorithm. This will be left for future work. Once the alpha and depth maps have been estimated, virtual views can be synthesized using this information. However, the depth values at object boundaries are usually inaccurate and hence occasionally the rendered point may lie outside the object boundary and holes may appear inside the object. These holes can be filled by inpainting techniques [57]. However, detecting outlying points is more difficult since the object boundary information in forms of the alpha maps is only available at the original views. In what follows, we shall propose a 2-D-shape morphing algorithm based on free-form deformation to synthesis the intermediate shape information so as to detect the outlying points during rendering.

C. Shape Morphing and Rendering

We now propose a free-form deformation method for shape morphing between every two adjacent frames. The free-form deformation method was originally proposed in [58] for 2-D

shape registration in a pattern recognition system. We now extend it to shape morphing. The method is based on maximizing the mutual information (MI) between the original image and the deformed images:

$$MI(A,T(B)) = H(p_A(I)) + H(p_{T(B)}(I)) - H(p_{A,T(B)}(I)) \tag{6}$$

where A and B represent the two input images; $H(p_A(I))$ denotes the entropy of image A; $H(p_{A,T(B)}(I))$ is the joint entropy of image A and $T(B)$, the deformed image of B; $T(*)$ is the deformation function to be determined; I denotes the intensity of the image; $p_A(I)$ denotes the intensity probability density function of image A; and $p_{A,T(B)}(I)$ is the joint intensity probability density between images A and $T(B)$. By maximizing the $MI(A,T(B))$ using the deformation function $T(B)$, the two original images can be registered. In [58], the maximization of (6) was done by the level set method [59] where B is gradually deformed for matching A. In our IBR system, the Chinese artifacts are quite densely sampled and hence successive images only differ slightly in the overall shape. Hence, if the evolution is relatively uniform, then the intermediate deformed results are good approximation to the desired shapes of intermediate views. From definition, the entropies in (6) are given by

$$\begin{aligned} H(p_A(I)) &= -\int_I p_A(i_A)\log p_A(i_A)di_A \qquad (7)\\ &= -\iint_I p_{A,T(B)}(i_A,i_B) \\ &\quad \cdot \log p_A(i_A)di_A di_B \qquad (8)\\ H(p_{T(B)}(I)) &= -\int_I p_{T(B)}(i_B)\log p_{T(B)}(i_B)di_B \\ &= -\iint_I p_{A,T(B)}(i_A,i_B) \\ &\quad \cdot \log p_{T(B)}(i_B)di_A di_B, \\ H(p_{A,T(B)}(I)) &= -\iint_I p_{A,T(B)}(i_A,i_B) \\ &\quad \cdot \log p_{A,T(B)}(i_A,i_B)di_A di_B \qquad (9) \end{aligned}$$

where i_A, i_B are the distance values in the image domain. The distance value is defined as the minimum Euclidean distance between the image pixel and the shape boundary. Hence, (6) can be rewritten as

$$E = \iint_I p_{A,T(B)}(i_A,i_B)\log\frac{p_{A,T(B)}(i_A,i_B)}{p_A(i_A)p_{T(B)}(i_B)}di_A di_B. \tag{10}$$

Meanwhile, the probability density functions can be approximated from the image data by using kernel estimation with, for example, the Gaussian kernel. Hence

$$\begin{aligned} p_A(i_A) &\approx \frac{1}{V}\iint_\Omega G_1(i_A - A(x,y))dxdy \qquad (11)\\ p_{T(B)}(i_B) &\approx \frac{1}{V}\iint_\Omega G_1(i_B - B'(x,y))dxdy \qquad (12)\\ p_{A,T(B)}(i_A,i_B) &\approx \frac{1}{V}\iint_\Omega G_2(i_A - A(x,y), i_B \\ &\quad - B'(x,y))dxdy \qquad (13) \end{aligned}$$

where $G_1(x) = 1/(\sqrt{2\pi\sigma}) \cdot e^{-x^2/(2\sigma^2)}$, $G_2(x,y) = 1/(2\pi\sigma_1\sigma_2) \cdot e^{-1/2\left(x^2/\sigma_1^2 + y^2/\sigma_2^2\right)}$, $B'(x,y) = B(T^1(x,y)), T^{-1}$ is the inverse function of $T,(x,y)$ is the pixel coordinate, Ω is the image domain and V is the area of Ω.

For accurate registration, the deformation is carried out in two steps, namely global and local morphing. In global morphing, which is performed first, the parameters of a global transformation are determined by matching the two images so as to model their relative translation and rotation. In the local morphing step, local deformation is performed and it is defined by a 2-D spline function. The transformation parameters, which are the displacement values at a regular grid to interpolate the spline function, are determined by minimizing the objective function in (6). For matching the object boundaries of two adjacent views, the boundaries of the two views are first converted to their respective distance images using the distance transformation [58]. The zero level set of the distance image gives the object boundary in the original image.

In the present work, the global deformation function T_G is chosen as an affine transformation: $T_G(x,y) = \boldsymbol{M} \cdot [x,y,1]$, where $\boldsymbol{M}$ is the 3×3 affine transformation matrix. The model parameters can be obtained by maximizing the objective function in (10). Let B' be the transformed image obtained by the affine transformation after the first step. The local transformation $T_L(B')$, which is a 2-D spline function, is parameterized by the displacement vector at a uniform grid of control points C, $\boldsymbol{P}_c(m,n) = [P_{c,x}(m,n), P_{c,y}(m,n)]^T$, which is indexed by $m = 1,\ldots,M$, $n = 1,\ldots,N$. If (X,Y) is the resolution of the input image, the spacing of the control points in the x and y direction are $\Delta_x = X/M$ and $\Delta_y = Y/N$, respectively. The deformation of any pixel in the image is obtained by spline interpolation of those at the grid points C. Therefore, the deformation of pixel (i,j), $\boldsymbol{P}(i,j) = [P_x(i,j), P_y(i,j)]^T$, where $1 \le i \le X, 1 \le j \le Y$, can be written as

$$\boldsymbol{P}(i,j) = \Sigma_{x=0}^3 \Sigma_{y=0}^3 \beta(u)\beta(v)\boldsymbol{P}_c(m+x, n+y) \tag{14}$$

where $u = i/\Delta_x - \lfloor i/\Delta_x \rfloor$, $v = j/\Delta_y - \lfloor j/\Delta_y \rfloor$, $\beta(u)$ is the cubic B-spline function, $\{m+x, n+y)|(x,y) \in [0,3]\}$ are the neighbor control points of (i,j). $[P_{c,x}(m,n), P_{c,y}(m,n)]$ are parameters of the transformation function $T_L(B')$ and $T_L(B'(i,j)) = B'(i + P_x(i,j), j + P_y(i,j))$.

By substituting (14) into (10), one gets the local matching data term to be minimized. In order to reduce the variance of the variable, an additional smoothing term and other prior constraints can be added to the data term above. A popular smoothing term is the L_2 norm of the displacement vector, $E_{smooth} = \sum_{(m,n)}(\alpha_1||\boldsymbol{P}_c(m,n)||^2 + \alpha_2||\nabla\boldsymbol{P}_c(m,n)||^2)$, where α_1, $i = 1,2$, are the regularization parameters. Since there are only a few model parameters of the affine transformation, the smoothing term is only needed in local matching.

Let t_i, $i = 1,\ldots,N$, be the parameters of T_G or T_L. The derivative of E with respect to t_i is

$$\begin{aligned} \frac{\partial E}{\partial t_i} = \iint_I \Bigg[&\left(1 + \log\frac{p_{A,T(B)}(i_A,i_B)}{p_A(i_A)p_{T(B)}(i_B)}\right)\frac{\partial p_{A,T(B)}(i_A,i_B)}{\partial t_i} \\ &- \frac{p_{A,T(B)}(i_A,i_B)}{p_{T(B)}(i_B)}\frac{\partial p_{T(B)}(i_B)}{\partial t_i}\Bigg] di_A di_B. \qquad (15) \end{aligned}$$

Fig. 4. Left: Typical depth maps of the artifacts computed. Right: example renderings on a given background.

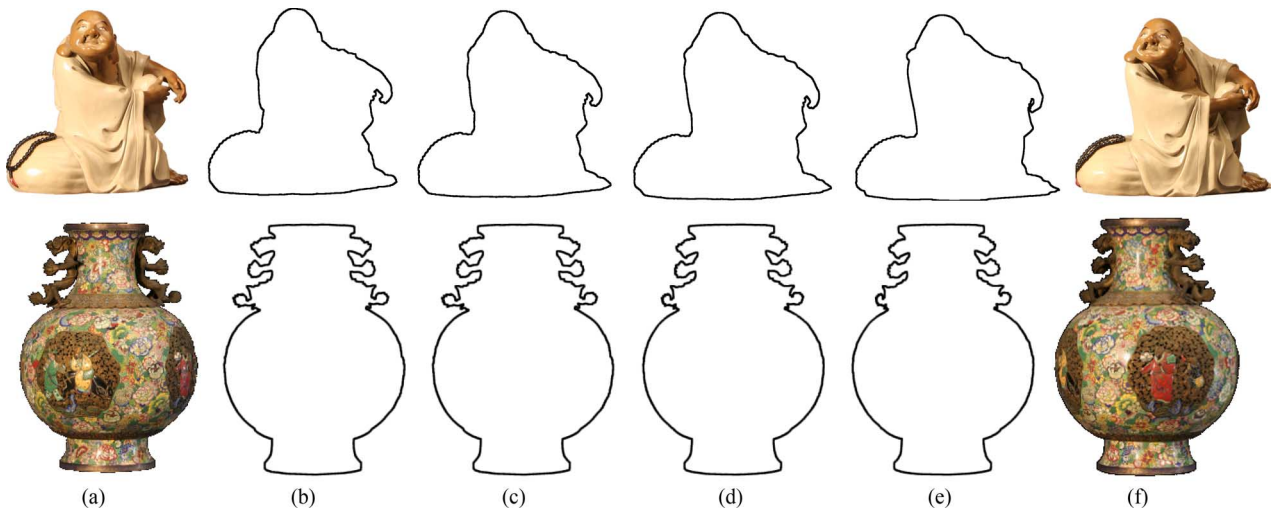


Fig. 5. (a) and (f) Segmented images at adjacent views. (b) to (e) are the boundaries generated by the morphing process.

Using the gradient above, the limited memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm [60] can be used to solve for the unconstrained nonlinear optimization problems. The L-BFGS method has the advantages that the explicit evaluation of the Hessian is not required, since it can be recursively estimated, and it was found to be much faster than the level set method. In order to speed up the optimization process in the local matching step, multiple resolution grids are employed. More precisely, a coarse grid like a (4×4) grid is first used to estimate the transformation parameters. Then, the grid size is progressively doubled until the distance between each control point is less than a certain value ε. In our case, we set ε as 4 pixels. The initial guess of the optimization at each resolution uses the previous result in order to speed up the convergence.

In the global morphing, we can interpolate the intermediate shape boundary linearly because the affine transformation is linear. If K intermediate shape boundaries are required, then the kth, $k = 1, \ldots, K$, intermediate boundary pixel can be approximated as

$$P_{i,j}^{(k)} = \frac{T_G(P_{i,j}^{(k-1)})}{K} + P_{i,j}^{(k-1)}. \tag{16}$$

In the local morphing, we can use each iteration result as the intermediate shape boundary. If the iteration converges in L steps, then the kth interview shape boundary can be approximated as

$$P_{i,j}^{k} = T_L^k(P_{i,j}^{k-1}) + P_{i,j}^{k-1} \tag{17}$$

where $k = 1, \ldots, L$ is the iteration number and T_L^k is the kth iteration local transformation result. The curves can be further interpolated to obtain the K desired views.

The morphed intermediate boundaries are stored and compressed. During rendering, these intermediate boundaries are then interpolated to facilitate real-time checking of outlying points during rendering and obtain a smooth object boundary.

D. Experimental Results

Fig. 5 shows the segmented images of two adjacent views in one of the SCLFs that we have taken. Fig. 5(b)–(e) shows four intermediate object boundaries when morphing from Fig. 5(a) to Fig. 5(f). Fig. 6 shows the rendering results of the proposed shape morphing algorithm. One can notice that the object boundary with shape morphing has better visual quality than the one without shape morphing. From the correspondences of the curves, the matte from the original images can be transferred to obtain those for the intermediate views.

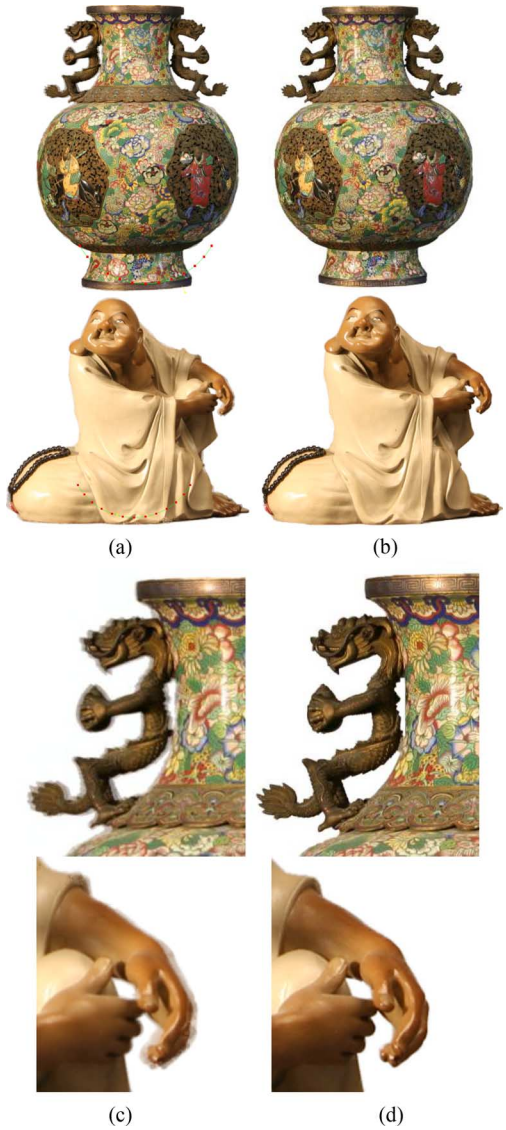


Fig. 6. Rendering examples of shape morphing (top) *Dragon Vase* and (bottom) *Buddha*. (a) is the rendering result without shape morphing, (b) is the one with shape morphing, (c) and (d) are the enlargement part of (a) and (b), respectively.

IV. Compression and Other Issues of SCLFS

Since the high resolution SCLFs involve large amount of data, they need to be compressed before transmission or storage. In the receiver, the images are decompressed and rendered in real-time according to the user's interaction. In our case, we need to transmit the texture and associated alpha and depth maps to the receiver for rendering. Considerable efforts have been devoted to the efficient compression methods of various image-based representations such as the light fields, lumigraphs, and concentric mosaics [4]–[11]. Apart from good coding efficiency, it is important for the compression algorithm to provide efficient access to the compressed light field data so that they can be selectively decoded to reduce the storage requirement and decoding complexity [13], [14], [16]–[20], [45], [61]. Moreover, a prioritized transmission scheme using scalable coding technique is highly desirable in order to reduce the latency in launching the application and improving the response time. In the following sections, we will propose a scalable SCLF compression algorithm, which offers the two desirable properties mentioned above.

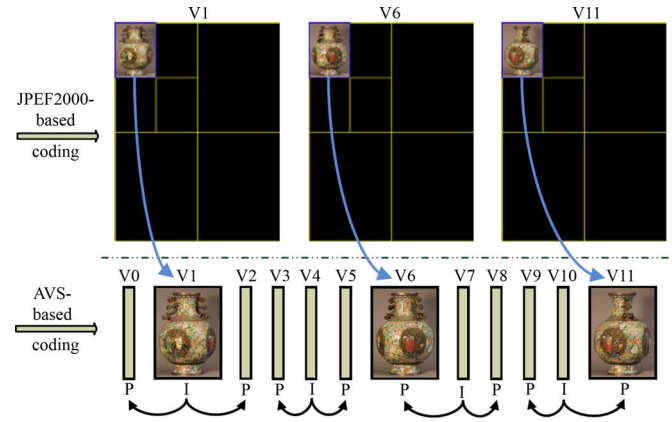


Fig. 7. Proposed compression framework.

A. Spatial Scalability

Scalable coding, in forms of spatial scalability, SNR, and temporal scalabilities [34], is an effective method to prioritize the transmission of compressed information to support users with different bandwidth/decoding complexities and to reduce the latency in progressive viewing. This is particular desirable in IBR compression because the data is in huge size but highly correlated. The encoding scheme incorporates spatial scalability in our previous proposed PV coding scheme [4], [5], [14], [45], [62]–[64] in order to support users with different bandwidth/decoding complexity and to reduce the latency in interactive rendering through prioritized transmission of the compressed SCLF data. We first perform wavelet transformation on the light field images so as to produce a multi-resolution representation of the SCLF images as shown in Fig. 8. The low resolution images are considered as the base layer while the higher frequency wavelet coefficients are used to form the enhancement layer. The base layer is intended to provide a reasonable rendering quality of the Chinese artifacts to users. Together with the enhancement layer, high quality rendering at full resolution can be achieved. The base layer can serve to reduce the transmission bandwidth during initial launching of the application or congested network traffic due to its reduced transmission bandwidth. We now describe the compression of the base and enhancement layers to support selective decoding/transmission. Using these functionalities, we can then perform prioritized transmission to reduce the loading latency and improve response time.

B. Coding of the Layers and Selective Transmission/Decoding

As can be seen from Fig. 7, the base layer consists of a sequence of light field images with reduced resolution. To facilitate efficient access and selective decoding of the image data, we avoid the excessive inter-dependency in predicting the P-pictures to simplify the selective decoding/transmission. Other configurations can be used at the expense of slightly higher decoding complexity, but possibly with a better coding efficiency. Instead of using MPEG-2 as in [64] for coding the

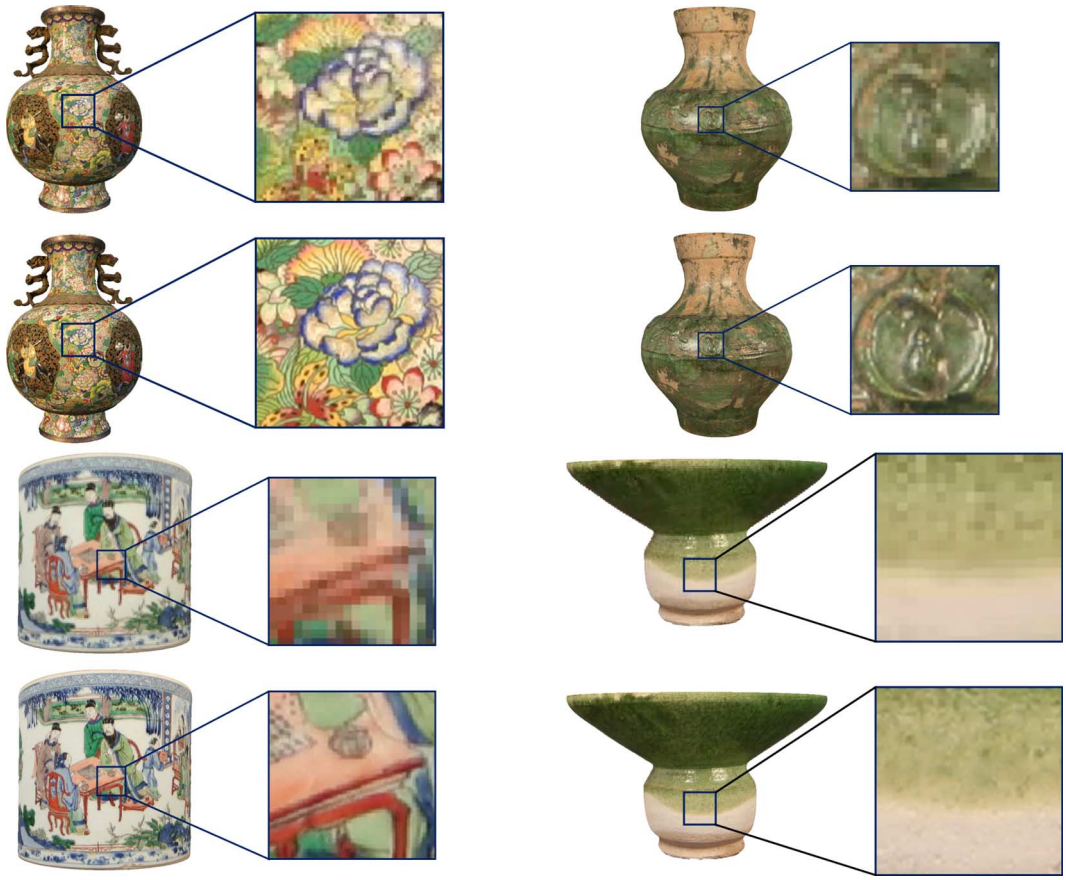


Fig. 8. Progressive rendering examples. The upper one is only rendered by base layer. The lower one is refined by enhancement layer.

P- and B-pictures, we have adopted the AVS standard because of its good coding efficiency. The AVS standard is also closely related to the H.264 video coding standard [65], which was used in our preliminary work [66]. Therefore, our configuration can also support standard H.264 or modified H.264 multiview video coding (H.264 MVC) [67] without major changes. The depth and alpha maps are coded analogously. The shape information is coded as in [64] and the additional morphing information will be transmitted after the base layer. Finally, we note that it is also possible to employ other packetizing systems and transmission protocols to the proposed scheme.

C. Prioritized Transmission

Even for the base layer, the data transmission may take long time before the rendering application can be launched. Usually, the users may wish to start rendering or viewing once the application is invoked. Under these circumstances, it is important to selectively transmit the compressed data according to some priority schemes in order to satisfy the rendering requirement of the users as much as possible. As all data is compressed and stored in a server, the rendering clients should fetch the necessary data for rendering according to predicted successively, then the entire compressed light fields have to be transmitted before rendering can proceed.

Fig. 9. Four artifacts captured by the proposed system. From left to right (with different image size): *Green Bottle* (976×1120), *Wine Glass* (656×480), *Brush Pot* (640×640), and *Dragon Vase* (1104×1488).

D. Experimental Results

We now evaluate the performance of the proposed system with four different ancient artifacts as shown in Fig. 9. The resolution of the light field image of the largest artifact is 1104×1488, 2-level of wavelet transform is performed. The enhancement layer consists of the compressed high frequency wavelet coefficients and they are organized as packets to facilitate interactive prioritized transmission. The coding results of the base layer are shown in Fig. 10, where the peak signal-to-noise ratio (PSNR) of Y component versus bits per pixel per frame (BPP) is plotted. For comparison, the experimental results of AVS1-P2 (Jizhun Profile) and H.264 jm18.2 (main profile) are also presented. It can be seen that our proposed method has similar performance with the original

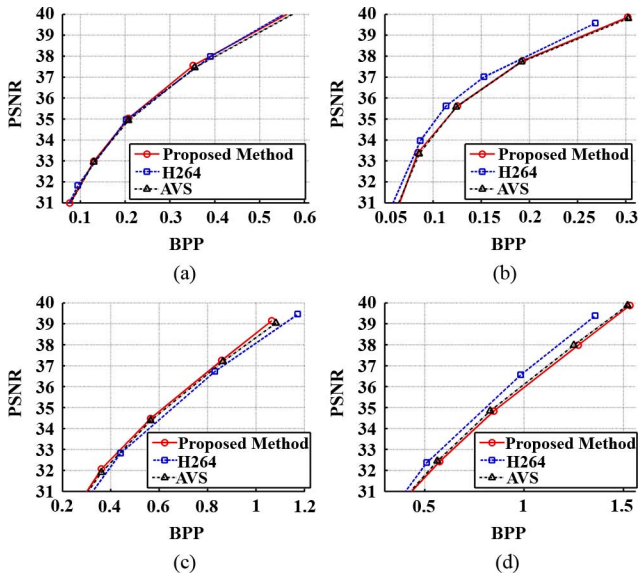


Fig. 10. Coding results for the base layer: (a) *Green Bottle*, (b) *Wine Glass*, (c) *Brush Pot*, and (d) *Dragon Vase* of SCLFs.

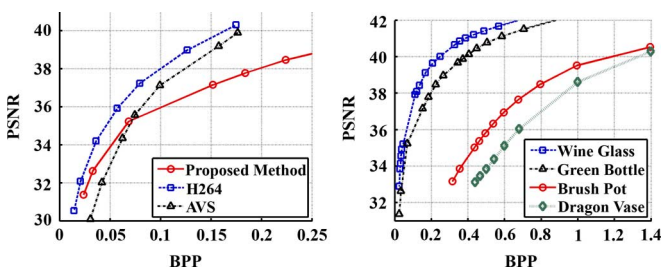


Fig. 11. Left: Comparison for the full frames of *Green Bottle* SCLFs. Right: Coding results for the full frames of SCLFs.

AVS algorithm and H.264. Fig. 11 shows the comparison of full-frame results of *Green Bottle* and the result of all four different artifacts. For high bitrate, the proposed scheme has some PSNR degradation for supporting the scalability functionality. Further improvement in coding performance may be obtained by using B-frames in the data stream but it will introduce more correlation among the data and making selective decoding more difficult. Therefore, there is tradeoff between coding efficient and simplicity in decoding.

Our application only requires two images of the base layer (one I- frame and one P-frame) of SCLFs to start the rendering process. Therefore, it only takes less than one second to transfer the necessary data under 1 Mbps network bandwidth. Then, users can start interacting with the launched application immediately, while the other images of the base layer will be transferred to users' terminal without interfering with the rendering process. Once all images of base layer are received, users can change to any view freely without any latency. Meanwhile, the enhancement layers with detailed information of current view will be transferred to refine the visual quality. An example of the progressive rendering result is given in Fig. 8. A demonstration video of the proposed prioritized transmission can be found

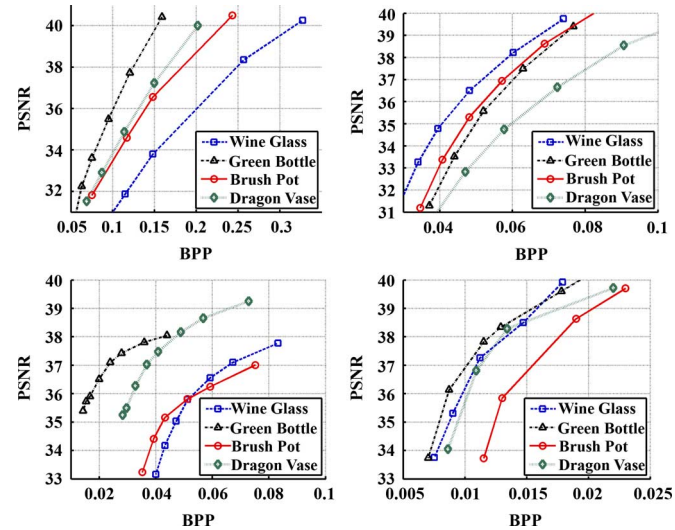


Fig. 12. Coding results for (upper row) the base layer and (lower row) the full frame of (left) alpha maps and (right) depth maps.

at [68]. If we render the scene after receiving all the SCLFs compressed by conventional methods, it will cost about two minutes under the same condition of network bandwidth. Our application can achieve over 200 frames per second (FPS) on an Intel i7 990X PC with an AMD Radeon HD 6950 or NVIDIA GeForce GTX 580 GPU. Coding results for the alpha and depth maps of the base layer and the full frame are also presented in Fig. 12. They have considerably better coding efficiency compared to the texture image, as more redundancies exist in the alpha and depth maps. In our system, both Samsung 42" shutter-based LCD 3DTV and Newsight 42" parallax-based autostereoscopic multiview AD3 TV are supported.

V. Conclusions

An image-based approach for the capturing, rendering and display of ancient Chinese artifacts for cultural heritage preservation has been presented. A multiple-camera circular array was constructed to record the SCLF of the Chinese artifacts. A systematic image-based approach and associate algorithms were developed for rendering new views of the artifacts. Moreover, an object-based compression scheme was employed to reduce the data size for storage and transmission of the texture, depth map and alpha maps associated with the object-based SCLF. Spatial redundancies among images are exploited to improve the coding performance, while avoiding excessive complexity in selective decoding to support fast rendering speed. To allow the ancient Chinese artifacts to be viewed over the internet, two-layer scalable prioritized transmission and rendering schemes of the compressed SCLF are developed with low latency. The frame rate for rendering is over 200 FPS on an Intel i7 990X PC with an AMD Radeon HD 6950 or NVIDIA GeForce GTX 580 GPU. The multiple views so synthesized enable the ancient artifacts to be displayed in 3-D/multi-view displays. Several collections from the University Museum and Art Gallery at The University of Hong Kong were captured and excellent rendering results are obtained.

REFERENCES

[1] H. Y. Shum, S. B. Kang, and S. C. Chan, "Survey of image-based representations and compression techniques," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 11, pp. 1020–1037, Nov. 2003.

[2] H. Y. Shum, S. C. Chan, and S. B. Kang, *Image-Based Rendering*. New York: Springer-Verlag, 2007.

[3] S. C. Chan, H. Y. Shum, and K. T. Ng, "Image-based rendering and synthesis: Technological advances and challenges," *IEEE Signal Process. Mag.*, vol. 24, no. 7, pp. 22–33, Nov. 2007.

[4] 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. IEEE Int. Symp. Circuits Syst.*, May 2004, vol. 3, pp. 905–908.

[5] S. C. Chan, K. T. Ng, Z. F. Gan, K. L. Chan, and H. Y. Shum, "The plenoptic video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 12, pp. 1650–1659, Dec. 2005.

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

[7] S. E. Chen, "Quicktime VR-an image-based approach to virtual environment navigation," in *Proc. Annu. Conf. Comput. Graph (SIGGRAPH'96)*, Aug. 1996, pp. 29–38.

[8] L. McMillan and G. Bishop, "Plenoptic modeling: An image-based rendering system," in *Proc. Annu. Conf. Comput. Graph (SIGGRAPH'95)*, Aug. 1995, pp. 39–46.

[9] M. Levoy and P. Hanrahan, "Light field rendering," in *Proc. Annu. Conf. Comput. Graph (SIGGRAPH'96)*, Aug. 1996, pp. 31–42.

[10] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, "The lumigraph," in *Proc. Annu. Conf. Comput. Graph (SIGGRAPH'96)*, Aug. 1996, pp. 43–54.

[11] H. Y. Shum and L. W. He, "Rendering with concentric mosaics," in *Proc. Annu. Conf. Comput. Graph (SIGGRAPH'99)*, Aug. 1999, pp. 299–306.

[12] *Encyclopedia of Computer Vision*, K. Ikeuchi, Ed. New York: Springer, to be published.

[13] H. Y. Shum, J. Sun, S. Yamazaki, Y. Li, and C. K. Tang, "Pop-up light field: An interactive image-based modeling and rendering system," *ACM Trans. Graph.*, vol. 23, no. 2, pp. 143–162, Apr. 2004.

[14] S. C. Chan, Z. F. Gan, K. T. Ng, K. L. Ho, and H. Y. Shum, "An object-based approach to image-based synthesis and processing for 3D and multiview televisions," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 6, pp. 821–831, Jun. 2009.

[15] T. Fujii, T. Kimoto, and M. Tanimoto, "Ray space coding for 3D visual communication," in *Proc. Picture Coding Symp.*, Mar. 1996, vol. 96, pp. 447–451.

[16] M. E. Lukacs, "Predictive coding of multi-viewpoint image sets," in *Proc. IEEE Int. Conf. Acoustics, Speech and Singal Processing*, Apr. 1986, pp. 521–524.

[17] J. Lu, H. Cai, J. G. Lou, and J. Li, "An effective epipolar geometry assisted motion-estimation technique for multi-view image coding," in *Proc. IEEE Int. Conf. Image Processing*, Atlanta, GA, Oct. 2006, pp. 1089–1092.

[18] C. Zhang and J. Li, "Compression of lumigraph with multiple reference frame (MRF) prediction and just-in-time rendering," in *Proc. IEEE Data Compression Conf.*, Mar. 2000, pp. 253–262.

[19] H. Kimata, M. Kitahara, K. Kamikura, and Y. Yashima, "Multi-view video coding using reference picture selection for free-viewpoint video Communication," in *Proc. Int. Picture Coding Symp.*, San Francisco, CA, 2004, pp. 499–502.

[20] ISO/IEC JTC1/SC29/WG11, Survey of Algorithms Used for Multi-View Video Coding (MVC), Doc. N6909. Hong Kong, China, 2005.

[21] J. Carranze, C. Theobolt, M. A. Magnor, and H. P. Seidel, "Free-viewpoint video of human actors," in *Proc. Annu. Conf. Comput. Graph (SIGGRAPH'03)*, Jul. 2003, pp. 569–577.

[22] Z. F. Gan, S. C. Chan, and H. Y. Shum, "Object tracking and matting for a class of dynamic image-based representations," in *Proc. IEEE Advanced Video and Signal-Based Surveillance*, Sep. 2005, pp. 81–86.

[23] B. Wilburn, M. Smulski, K. Lee, and M. Horowitz, "The light field video camera," in *Proc. SPIE Electronic Imaging: Media Processors*, Jan. 2002, vol. 4674, pp. 29–36.

[24] B. Goldlücke, M. Magnor, and B. Wilburn, "Hardware-accelerated dynamic light field rendering," in *Proc. Vision, Modeling, Visualization*, Nov. 2002, pp. 455–462.

[25] T. Naemura, J. Tago, and H. Harashima, "Real-time video-based modeling and rendering of 3D scenes," *IEEE Comput. Graph. Appl.*, pp. 66–73, Mar.-Apr. 2002.

[26] J. C. Yang, M. Everett, C. Buehler, and L. McMillan, "A real-time distributed light field camera," in *Proc. Eurographics Workshop Rendering*, Jun. 2002, pp. 77–86.

[27] J. Li, H. Y. Shum, and Y. Q. Zhang, "On the compression of image based rendering scene: A comparison among block, reference and wavelet coders," *Int. J. Image Graph.*, vol. 1, no. 1, pp. 45–61, 2001.

[28] M. Magnor and B. Girod, "Data compression for light-field rendering," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 3, pp. 338–343, Apr. 2000.

[29] X. Tong and R. M. Gray, "Coding of multi-view images for immersive viewing," in *Proc. IEEE Int. Conf. Acoustics, Speech and Singal Processing*, Jun. 2000, vol. 4, pp. 1879–1882.

[30] J. R. Ohm, "Stereo/multiview encoding using the MPEG family of standards," in *Proc. Electronic Imaging*, San Diego, CA, Jan. 1999, pp. 242–253.

[31] A. Puri, R. V. Kollarits, and B. G. Haskell, "Basics of stereoscopic video, new compression results with MPEG-2 and a proposal for MPEG-4," *J. Signal Process.: Image Commun.*, vol. 10, pp. 201–234, 1997.

[32] T. Naemura, M. Kaneko, and H. Harashima, "Compression and representation of 3-D images," *IEICE Trans. Inf. Syst.*, vol. E82-D, no. 3, pp. 558–567, 1999.

[33] J. R. Ohm, "Encoding and reconstruction of multiview video objects: Looking at data compression in the context of the MPEG-4 multimedia standard," *IEEE Signal Process. Mag.*, vol. 16, no. 3, pp. 47–54, May 1999.

[34] A. Smolic, K. Müeller, N. Stefanoski, J. Ostermann, A. Gotchev, G. B. Akar, G. Triantafyllidis, and A. Koz, "Coding algorithms for 3DTV-A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 11, pp. 1606–1621, Oct. 2007.

[35] P. Merkle, A. Smolic, K. Müller, and T. Wiegand, "Efficient prediction structures for multiview video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 11, pp. 1461–1473, Oct. 2007.

[36] K. Müller, P. Merkle, and T. Wiegand, "Compressing time-varying visual content," *IEEE Signal Processing Mag.*, vol. 24, no. 7, pp. 58–67, Nov. 2007.

[37] M. Flierl and B. Girod, "Multiview video compression," *IEEE Signal Processing Mag.*, vol. 24, no. 7, pp. 66–76, Nov. 2007.

[38] Y. Wang, J. Ostermann, and Y. Q. Zhang, *Video Processing and Communications*. Englewood Cliffs, NJ: Prentice-Hall, 2002.

[39] W. Yang, Y. Lu, F. Wu, J. Cai, K. N. Ngan, and S. Li, "4-D wavelet-based multiview video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 11, pp. 1385–1396, Nov. 2006.

[40] Z. Y. Zhu, K. T. Ng, S. C. Chan, and H. Y. Shum, "Image-based rendering of ancient Chinese artifacts for multi-view displays—A multi-camera approach," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2010, pp. 3252–3255.

[41] S. C. Chan, H. Y. Shum, and K. T. Ng, "Image-based rendering and synthesis: Technological advances and challenges," *IEEE Signal Process. Mag.*, vol. 24, no. 7, pp. 22–33, Nov. 2007.

[42] M. Levoy *et al.*, "The digital Michelangelo project: 3D scanning of large statues," in *Proc. Annu. Conf. Comput. Graph (SIGGRAPH'02)*, Aug. 2002, pp. 131–144.

[43] 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. Annu. Conf. Comput. Graph (SIGGRAPH'99)*, Aug. 1999, pp. 223–239.

[44] K. Ikeuchi, A. Nakazawa, K. Hasegawa, and T. Ohishi, "The great Buddha project: Modeling cultural heritage for VR systems through observation," in *Proc. 2nd IEEE/ACM Int. Symp. Mixed Augmented Reality*, Oct. 2003, pp. 7–16.

[45] 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. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Hong Kong, Apr. 2003, vol. 3, pp. 653–656.

[46] GB/T 20090.2-2006, The Information Technology Advanced Audio and Video Coding Part 2: Video, 2006.

[47] ISO/IEC 15444-1:2004, Information technology—JPEG 2000 image coding system: Core coding system, 2004.

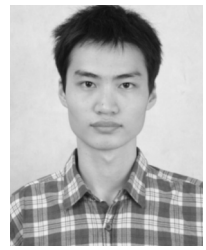
[48] Y. Li, J. Sun, C. K. Tang, and H. Y. Shum, "Lazy snapping," in *Proc. Annu. Conf. Comput. Graph (SIGGRAPH'04)*, Aug. 2004, pp. 303–308.

[49] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, Nov. 2001.

[50] J. Sun, Y. Li, S. B. Kang, and H. Y. Shum, "Symmetric stereo matching for occlusion handling," in *Proc. IEEE Comput. Soc. Conf. CVPR*, Aug. 2005, vol. 2, pp. 399–406.
[51] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1330–1334, Nov. 2000.
[52] S. Osher and N. Paragios, *Geometric Level Set Methods in Imaging, Vision, and Graphics*. New York: Springer-Verlag, 2003.
[53] T. F. Chan and L. A. Vese, "Active contours without edges," *IEEE Trans. Image Process.*, vol. 10, no. 2, pp. 266–277, Feb. 2001.
[54] J. van de Weijer, T. Gevers, and A. D. Bagdanov, "Boosting color saliency in image feature detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 1, pp. 150–156, Jan. 2006.
[55] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vision*, vol. 47, no. 1/2/3, pp. 7–42, 2002.
[56] K. J. Yoon and I. S. Kweon, "Adaptive support-weight approach for correspondence search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 650–656, Apr. 2006.
[57] A. Criminisi, P. Pérez, and K. Toyama, "Object removal by exemplar-based inpainting," in *Proc. Conf. Computer Vision Pattern Recognition*, Madison, WI, Jun. 2003, vol. 2, pp. 721–728.
[58] X. L. Huang, N. Paragios, and D. N. Metaxas, "Shape registration in implicit spaces using information theory and free form deformations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 8, pp. 1303–1318, Aug. 2006.
[59] J. Sethian, *Level Set Methods and Fast Marching Methods: Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision, and Materials Science*. Cambridge, U.K.: Cambridge Univ. Press, 1999.
[60] R. H. Byrd, P. Lu, and J. Nocedal, "A limited memory algorithm for bound constrained optimization (1995)," *SIAM J. Sci. Statist. Comput.*, vol. 16, no. 5, pp. 1190–1208, 1994.
[61] M. G. Strintzis and S. Malasiotis, "Object-based coding of stereoscopic and 3D image sequences: A review," *IEEE Signal Processing Mag.*, vol. 16, pp. 14–28, May 1999.
[62] K. T. Ng, S. C. Chan, and H. Y. Shum, "The data compression and transmission aspects of panoramic videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 1, pp. 82–95, Jan. 2005.
[63] Q. Wu, K. T. Ng, S. C. Chan, and H. Y. Shum, "On object-based compression for a class of dynamic image-based representations," in *Proc. IEEE Int. Conf. Image Processing*, Sep. 2005, pp. 405–408.
[64] K. T. Ng, Q. Wu, S. C. Chan, and H. Y. Shum, "Object-based coding for plenoptic videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 4, pp. 548–562, Apr. 2010.
[65] ISO/IEC 14496-10:2010, Information Technology—Coding of Audio-Visual Objects Part 10: Advanced Video Coding, 2010.
[66] X. Z. Yao, S. C. Chan, Z. Y. Zhu, K. T. Ng, and H. Y. Shum, "Image-based compression, prioritized transmission and progressive rendering of circular light fields (CLFS) for ancient Chinese artifacts," in *Proc. IEEE Asia Pacific Conf. Circuits Syst.*, Dec. 2010, pp. 340–343.
[67] ISO/IEC 14496-10:2010, Information technology—Coding of Audio-Visual Objects Part 10: Advanced Video Coding, Annex H, 2010.
[68] [Online]. Available: http://youtu.be/uK1kko7dnMI.

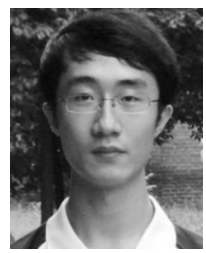
King-To Ng (S'96–M'02) received the B.Eng. degree in computer engineering from the City University of Hong Kong, Hong Kong, China, in 1994, and the M.Phil. and Ph.D. degrees in electrical and electronic engineering from The University of Hong Kong, in 1998 and 2003, respectively.

From 2004 to 2012, he was a Postdoctoral Fellow in the Department of Electrical and Electronic Engineering, The University of Hong Kong. His research interests include visual communication, image-based rendering, and video broadcast and transmission.



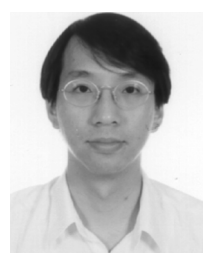
Zhen-Yu Zhu received the B.Eng. degree from Huazhong University of Science and Technology in 2008, Wuhan, China. He is currently pursuing the Ph.D. degree at the department of Electrical and Electronic Engineering, The University of Hong Kong.

His main research interests are in digital image processing, multiple view geometry, and 3-D reconstruction.



Chong Wang received the B.Eng. degree from Zhejiang University of Technology, Hangzhou, China, in 2007, and the M.Eng. degree from the University of Science and Technology of China, Hefei, in 2010. He is currently pursuing the Ph.D. degree at the Department of Electrical and Electronic Engineering, The University of Hong Kong.

His main research interests are in digital signal processing, video de-noising, and super-resolution.



Shing-Chow Chan (S'87–M'92) received the B.Sc. (Eng) and Ph.D. degrees in electrical engineering from The University of Hong Kong, Hong Kong, China, in 1986 and 1992, respectively.

He joined City Polytechnic of Hong Kong in 1990 as an assistant Lecturer and later as a University Lecturer. Since 1994, he has been with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong, and is now a Professor. He held visiting positions in Microsoft Corporation, Redmond, WA, Microsoft Research Asia, University of Texas at Arlington, and Nanyang Technological University. His research interests include fast transform algorithms, filter design and realization, multirate and array signal processing, communications and biomedical signal processing, and image-based rendering.

Dr. Chan is currently a member of the Digital Signal Processing Technical Committee of the IEEE Circuits and Systems Society and an Associate Editor of the *Journal of Signal Processing Systems*. He was an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS I from 2008 to 2009 and the Chairman of the IEEE Hong Kong Chapter of Signal Processing from 2000 to 2002. He was in the organizing committees of several international conferences including the special session co-chair of IEEE International Conference on Acoustic Speech, Signal Processing, 2003, the Technical program co-chair of IEEE International Conference on Field-Programmable Technology (FPT), 2002, the technical program committees of IEEE APCCAS'2006, IEEE APCCAS'2008, and IEEE BioCAS'2006, and the tutorial chair of the IEEE International Conference on Image Processing (ICIP), 2010.

Heung-Yeung Shum (M'90–SM'01–F'06) received the Ph.D. degree in robotics from the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA.

Currently, he is the Corporate Vice President responsible for search product development with Microsoft Corporation, Redmond, WA. Previously, he oversaw research activities at Microsoft Research Asia, Beijing, China, as well as the lab's collaborations with universities in the Asia Pacific region. He was responsible for the Internet Services Research Center, an applied research organization dedicated to long-term and short-term technology investments in search and advertising at Microsoft, and he was a Researcher with Microsoft Research in 1996, based in Redmond. He moved to Beijing, China, as one of the founding members of Microsoft Research, China (later renamed Microsoft Research Asia). There he began a nine-year tenure as a Research Manager, subsequently moving on to become Assistant Managing Director, Managing Director of Microsoft Research Asia, Distinguished Engineer, and Corporate Vice President. He has published more than 100 papers about computer vision, computer graphics, pattern recognition, statistical learning, and robotics. He holds more than 50 U.S. patents.

Dr. Shum is a Fellow of the Association for Computing Machinery.