Annotating Traditional Chinese Paintings for Immersive Virtual Exhibition

WEI MA and YIZHOU WANG, Peking University
YING-QING XU, Tsinghua University and Microsoft Research Asia
QIONG LI, The Palace Museum, Beijing
XIN MA, Microsoft Research Asia
WEN GAO, Peking University

We propose a new method of annotating a masterpiece of traditional Chinese painting with voice dubbings and environmental sounds. The painting was created with moving focus drawing technique without rigorous perspective. A novel algorithm is proposed to infer the 3D space of the painting according to its layout and embed the audio annotations. For exhibition, the masterpiece is scanned into a high-resolution gigapixel image for presenting the drawing details, and we develop an interactive multimedia system with a panning and zooming interface to enable smooth navigation on the giant painting and exploring the historical culture. During the navigation, the system estimates the 3D position of the user's viewpoint from his/her actions, and subsequently synthesizes a realistic stereo audio field according to the viewer's orientation and distance from the annotations in the 3D space. The proposed system provides an immersive user experience by rendering a visual-audio consistent perception.

Categories and Subject Descriptors: I.3.8 [**Computer Graphics**]: Applications; J.5 [**Computer Applications**]: Arts and Humanities

General Terms: Algorithms

Additional Key Words and Phrases: Chinese paintings, moving focus, immersive experience, virtual exhibition

ACM Reference Format:

Ma, W., Wang, Y., Xu, Y.-Q., Li, Q., Ma, X., and Gao, W. 2012. Annotating traditional chinese paintings for immersive virtual exhibition. ACM J. Comput. Cult. Herit. 5, 2, Article 6 (July 2012), 12 pages.
DOI = 10.1145/2307723.2307725 http://doi.acm.org/10.1145/2307723.2307725

1. INTRODUCTION

Ancient paintings are significant in that they help us to learn about ancient art [Rowley 1947] and historical culture [Hansen 1996]. In order to preserve cultural heritage, in traditional exhibitions they

This work is supported by the National Natural Science Foundation of China (no. 61003105) and Microsoft Research Asia eHeritage theme-based program research funding.
W. Ma is now affiliated with Beijing University of Technology, China.
Authors' addresses: W. Ma (corresponding author), College of Computer Science and Technology, Beijing University of Technology, China; email: rubbymawei@gmail.com; Y. Wang (corresponding author), National Engineering Lab for Video Technology and Key Lab of Machine Perception (MOE), School of Electrical Engineering and Computer Science, Peking University, China; email: yizhou.wang@pku.edu.cn; Y.-Q. Xu (corresponding author), The School of Art and Design, Tsinghua University and Microsoft Research Asia; email: yqzu@singhua.edu.cn; Q. Li, The Palace Museum, Beijing, China; X. Ma, Microsoft Research Asia; W. Gao, National Engineering Lab for Video Technology and Key Lab of Machine Perception (MOE), School of Electrical Engineering and Computer Science, Peking University, China.

DOI 10.1145/2307723.2307725 http://doi.acm.org/10.1145/2307723.2307725

are displayed at a certain distance or inside glass cabinets. In such cases, it is hard to closely view their details and acquire additional background knowledge. As multimedia technology has been more and more developed, exhibiting digitized replicas has become a popular alternative means of displaying cultural heritage, ranging from archeological sites and statues [Carrozzino et al. 2009; Koller et al. 2009; Foni et al. 2010], to paintings [Chu and Tai 2001; Zhu et al. 2004; Lin et al. 2009]. Compared to traditional exhibitions, the digital exhibitions not only provide viewers with a means to interact with an exhibit without actually contacting it, but also are capable of revealing background knowledge [Foni et al. 2010]. Most of the state-of-the-art multimedia exhibitions present a painting or parts of it in the form of animations [Chu and Tai 2001; Zhu et al. 2004] sometimes being triggered by users using intelligent user interfaces [SeoulAliveGallery 2008; Lin et al. 2009]. However, from these exhibitions, it is hard to clearly see the original appearances of the artwork.

Our goal is to design a multimedia exhibition system which allows viewers to freely and closely appreciate a digital version of an original painting and gain its historical knowledge with immersive user experiences. To achieve this goal, the painting is represented in the form of a high-resolution gigapixel image annotated with voice dubbings and sounds from nature. The audio pieces are organized in form of conversations and natural environments. In order to provide a natural user experience when the viewer explores the painting, perception consistent visual-audio rendering is essential [Witmer and Singer 1998]. As we know, when viewing a picture with a 3D scene, people perceive a space with objects' in-plane positions as well as their relative depths (refer to Chapter 5 in Palmer [1999]). We name this space 3D perception space. For spatially visual-audio consistent perception, the output audio field should also render the 3D perception space from the point of the viewer. To achieve this point, we infer the 3D positions of sound sources in the perception space to control the rendering of the audio field.

However, unlike photographs and western paintings produced by optical perspectives, most traditional Chinese paintings were created by using "moving focus" [Fong 2003]. This renders ineffective most methods commonly employed to analyze the 3D world [Horry et al. 1997; Furukawa et al. 2009]. We propose a solution to the moving focus problem, which positions the sound annotations in the hidden 3D space. When a user navigates the gigapixel image by panning and zooming, the user interface estimates the 3D position of the user's current viewpoint, and subsequently synthesizes a stereo-audio field according to the viewer's spatial orientation and distance from the annotations.

The rest of the article is organized as follows: Section 2 presents related work. The target painting, "Life Along the Bian River at the Pure Brightness Festival" (Bian River scroll for short) is briefly introduced in Section 3. Section 4 presents the solution to 3D inference from moving focus paintings. The data organization is given in Section 5. Section 6 describes the hardware and considerations on user interface. Finally, conclusions and future work are outlined in Section 7.

2. RELATED WORK

Digital exhibitions proceed along with the development of multimedia, ranging from 3D touring and animation, to intelligent interaction. In 1997, Horry et al. [1997] recovered box structured 3D scenes from pictures of single perspectives, thereby allowing viewers to take a tour into the pictures. Similarly, Chu and Tai [2001] extended the method in Horry et al. [1997] to paintings of two perspectives and provided a virtual walkthrough in a 3D textured virtual scene. The preceding methods are limited to visualization of simply structured scenes. In order to more clearly tell the stories about paintings, Zhu et al. [2004] revived the characters of the "Dunhuang Murals" using animations. Likewise, in the Chinese pavilion of the Shanghai World Expo 2010, a movie of animations, derived by Crystal Digital Technology Co., LTD. [Crystal 2010] from the Bian River scroll, were displayed by twelve projectors to present the lives of the ancient people.

With the development of techniques in human computer interaction, such as multitouch and speech recognition, many researchers have introduced them to digital exhibitions of paintings. For example, in 2007, at an exhibition of a late version of the Bian River scroll, presented by the National Palace Museum in Taibei, the painting was reported to be "alive." Many scenes in the digital painting were enhanced with a set of animations. The painting was projected onto a long sheet of rice paper by three projectors. When tourists visited a specific scene by touching the display, the corresponding movies would jump out of the painting and start to play. In 2008, in Seoul's Alive Gallery, the Mona Lisa was animated and endowed with the ability of speech recognition. This ability helped her to answer questions from visitors, such as "why don't you have any eyebrows" [SeoulAliveGallery 2008]. Lin et al. [2009] designed a system for showing another late version of the Bian River scroll using touch-based navigation. During the navigation, the touched area would become illuminated and animations would begin to play.

The aforesaid exhibitions are all very attractive. However, their main objectives for exhibition are to render 3D scenes and animations interpreted by modern designers, rather than the original art pieces. Recently, Kopf et al. [2007] proposed a method to view gigapixel images via smooth navigation. Their study inspired the work in Luan et al. [2008] for presenting gigapixel images with audio annotations. In particular, the authors in Luan et al. [2008] annotated gigapixel photographs with a set of individual voices. A panning and zooming interface was provided for visitors to navigate the pictures. At the same time, the audio annotations were controlled and adjusted based on viewers' distance from them. This system displays every detail of the image and introduces various aspects of the image with auditory supplementation. These studies inspired us in the design of the exhibition system for showing ancient Chinese paintings, and the daily lives of the ancient people in them.

There are two big differences between our work and the work in Luan et al. [2008]. First, as we mentioned, one key issue in annotating an image for immersive exhibition is how to infer the positions of the annotations in the 3D perception space of the image. In the field of 3D inference, most related work focuses on photographs with pin-hole perspective models [Horry et al. 1997; Furukawa et al. 2009]. Few related works have been presented to deal with irregular images. Luan et al. [2008] manually drew coarse depth maps of panorama pictures. That is a good choice for those images with unknown projection models. However, our target is a painting which contains hundreds of interdependent objects. Manual indication of a 3D layout with object-level resolution is time consuming and labor intensive. An intelligent algorithm is proposed in this article to solve the 3D layout inference problem. Second, we are engaged in developing an artwork exhibition system rather than simply showing the details of a photograph and telling viewers what an object is. Therefore, all our considerations in algorithms and strategies respect the artwork and the historical culture depicted in it. For example, all sound scripts are strictly based on in-depth study of the culture in the painting's dynasty. Voices are well organized to present the real life of the ancient people.

3. ABOUT TRADITIONAL CHINESE PAINTINGS AND THE BIAN RIVER SCROLL

In this work, we mainly target traditional Chinese paintings. These paintings differ much from western paintings in form (e.g., long scrolls versus canvases of regular size), composition techniques for spatial recession (moving focus versus perspective), etc. Rowley presented detailed differences between the two in his book "Principles of Chinese Painting" [Rowley 1947].

We take one of the most famous ancient Chinese paintings, the Bian River scroll (Figure 1), for a case study. For a fully unfolded image, refer to ChinaOnlineMuseum [2011]. The painting, the work of an artist in the Northern Song Dynasty, Zeduan Zhang, is now more than 800 years old. It has attracted considerable attention since it was created. Many reinterpretive replicas appeared after Zeduan Zhang's creation of the original due to its great social impact. The painting captures the daily

Fig. 1. The Bian River scroll. (Image is courtesy of the Palace Museum (Beijing)).

lives of people from the Song period at the capital, Pien Jing (today's Kaifeng in Henan Province, China). The entire piece is painted on a long scroll of 24.8 centimeters in height and 528 centimeters in width. It is composed of three scenes, suburb, wharf, and urban, and it accommodates over eight hundred people and tens of houses, trees, and other objects. The content reveals the lifestyle of all stratums of society, from rich to poor, as well as different economic activities in the rural areas and the city. Appreciating this long scroll is a right-to-left gradual process as our system will show. Please refer to Roderick [1965] and Hansen [1996] for a more detailed introduction to the Bian River scroll. All in all, the painting is not only a timeless work of art, but also a great reference to the culture of ancient China. Therefore, it is important to design and implement an exhibition platform which enables visitors to appreciate the artwork and learn about the ancient culture depicted in the painting.

4. ANNOTATING THE PAINTING

Annotating the painting with the voice recordings (more than 500 in all, all in ".wav" format) is a two-step process. In the first step, each audio file is associated with a corresponding object on the painting. For example, a recording of human speech is linked to the head of a human figure. To implement this step efficiently, we design a specific tool for interactive association. The association tool utilizes an interactive interface that enables users to draw a box on a figure of the painting (see the red boxes in Figure 2) and associate an audio file with this figure. The box defines the sound source position and the ground-contact point of the figure. The top-edge center of the box is the sound source position. The bottom-edge center indicates the figure's in-plane ground-contact point, which is essential for inferring the 3D position of the sound source as we will explain in the following subsections. With this tool, the association of all the pairs of audio files and figures (shown in Figure 2) can be finished within an hour.

In the second step, a new algorithm is proposed to compute the 3D locations of the annotations from their coordinates in the 2D image. In the following subsections, first, we touch upon the concept of moving focus. Second, we analyze the 3D inference problem. Third, we present the solution.

4.1 Moving Focus

Western painters in the Renaissance developed the scientific pinhole perspective [Willats 1997] to determine on 2D canvases the layout of objects in 3D scenes. Pinhole perspective was widely known to common people after its development in photography in the 19th century. In contrast to the fixed single viewpoint in drawing western paintings, as stated in Fong [2003], classical Chinese painters

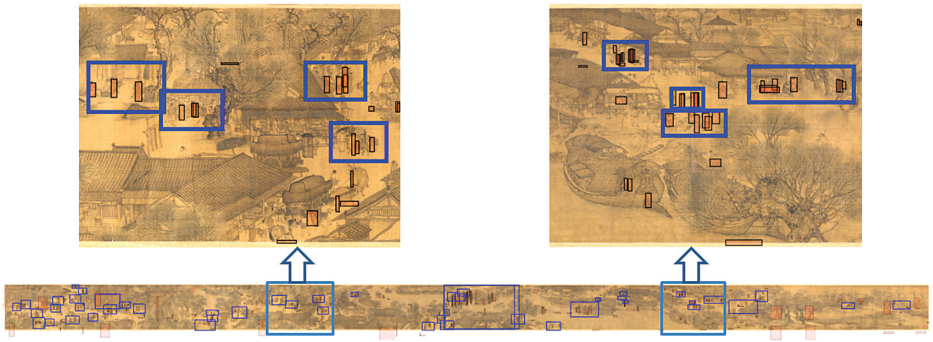

Fig. 2. The distribution of all the audio annotations. Red boxes: annotated figures; blue boxes: stories. (Image is courtesy of the Palace Museum (Beijing)).

position objects on the picture plane additively, in an expanded field of vision, owing to his/her *moving focus* perpendicular to the picture plane and dynamic in parallel.

In this technique, (1) static objects, such as buildings and hills, are aligned in a continuous sequence with a certain angle relative to the horizontal axis of the picture plane. We call this angle inclination angle. The sequence forms oriented roads and other spaces for accommodating dynamic beings, such as humans and carts. The oriented sequences and spaces create the perception of continuous recession on 2D images [Fong 2003]; (2) canvases can be long scrolls typical to traditional Chinese paintings. Each scroll generally has groups of sequences with different inclination angles to form a global space with spatial recession continuously varying across the canvas. Fong, a famous Chinese American artist, gave a graphical illustration of moving focus and its comparison with perspectives [Fong 2003].

4.2 Problem Analysis

Moving focus in paintings have the same advantage of depicting a wide field of view as multiperspective which appears recently in photography [Vallance and Calder 2002; Yu and McMillan 2005; Degener and Klein 2009]. Here, the perspectives are not limited to traditional pinhole models. Numerous nonstandard pinhole projection models, such as parallel, pushbroom, and twisted orthographic projections [Yu and McMillan 2005] are included. This fact motivates us to formulate the moving focus problem using multiperspective. Unfortunately, according to our study, most of the work related to multiperspective focuses on acquisition of images of wide field of view [Vallance and Calder 2002; Yu and McMillan 2005; Degener and Klein 2009]. Few researchers have attempted to infer 3D information from a single multiperspective image. Up to now, only very simple cases are studied [Chu and Tai 2001], in which a couple of independent pinhole perspectives were manually indicated and scenes were assumed to be box-like for depth estimation.

3D inference from a single multiperspective image is challenging since it involves a large joint solution space of variables including the number of viewpoints/cameras, the position, pose, and projection type of each viewpoint, camera intrinsic parameters, and scene depth [Szeliski 2011]. Limited pictorial information usually provides insufficient clues to infer a reasonable solution. To simplify the problem, we utilize the properties of moving focus. Considering that moving focus has viewpoints dynamic in parallel and always perpendicular to canvases, we choose multiple parallel projections as the geometric model. In this case, the projection of viewpoints are known to be parallel which is independent of viewpoint positions and intrinsic parameters [Szeliski 2011]. The unknown variables are the number

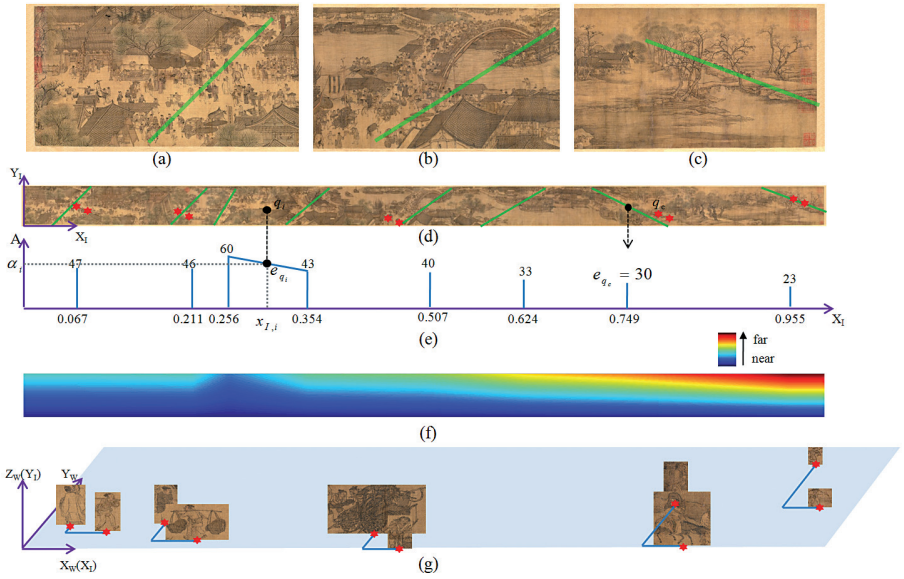

Fig. 3. Structures in the Bian River scroll: (d) slanting lines (the green) in the whole scroll; (a), (b) and (c) close views of the parts in (d); (e) inclination coordinates system recording the inclination angles of the slanting lines; (f) depth map of the ground; (g) illustration of the 3D layout of the objects marked by the red spots in (d). See the text for explanation. (Image is courtesy of the Palace Museum (Beijing)).

of viewpoints, their poses, and scene depth. In the following, we will explain how to determine the unknowns and compute the sound source positions with simple interaction.

4.3 Solution

We position the image coordinates system, with main axes X_I and Y_I, at the left-bottom corner of the painting (see Figure 3(d)). A global 3D coordinates frame is defined, with axis X_W overlapping with X_I and $X_W O_W Y_W$ plane coincident with the 3D ground plane. We use $p = (x_W, y_W, z_W)$ to represent 3D points in the 3D world (in meters) and $q = (x_I, y_I)$ to denote 2D positions in the image space (in pixels). Our goal is to compute a p given a q of a sound source position.

As we observe from traditional Chinese paintings depicting scenes [Yang et al. 1997], ground is an essential component. All the objects are directly or indirectly supported by the ground. Therefore, we factorize the inference of 3D sound source positions into two steps. Given an annotated figure (represented as a box in Figure 2) on the painting, we first infer the 3D position of its ground-contact point (the bottom-edge center of the box), and then compute the 3D position of its sound source (located at the top-edge center of the box) by the 3D ground-contact point, the normal of the 3D ground and the height of the figure.

Assuming q_i is a ground-contact point on the painting, we infer its corresponding point p_i on the 3D ground as follows. According to the definition of moving focus given in Section 4.1, the extent of the recession is determined by the inclination angle. We mark those aligned objects using slanting lines. In all, eight lines are drawn manually on the Bian River scroll (see Figure 3(a), (b), (c), and (d)), along six roads, one bridge, and a wall of a bell tower. These slanting lines vertically go through the painting and determine the alignment of local objects.

We record the absolute values of the inclination angles in an inclination coordinates system (refer to Figure 3(e)). Its horizontal axis is X_I and vertical axis is A representing the degrees of inclination

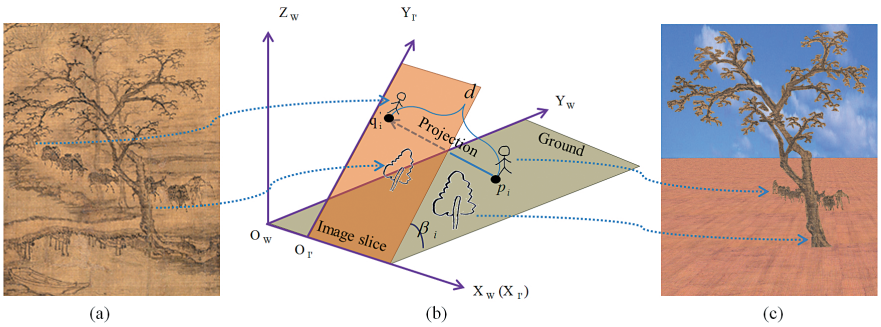

Fig. 4. Illustration of putting sound sources on an image slice (a) in the 3D space (c). (b) $X_{I'}O_{I'}Y_{I'}$ is the local coordinates system of the image slice. q_i' is a point on the slice. $X_W Y_W Z_W O_W$ determines the coordinates system of the 3D world. β_i is the bird's eye-view angle. p_i is the corresponding point of q_i' on the 3D ground. d is the depth of q_i'. (Image is courtesy of the Palace Museum (Beijing)).

angles. Assuming the center of a slanting line is q_c, the corresponding inclination coordinates are recorded as $e_c = (x_{I,c}, \alpha_c)$, where $x_{I,c}$ represents the X_I component of q_c and α_c is the inclination angle of the line. The inclination coordinates of all the slanting lines form a recorded set. Given a point $q_i = (x_{I,i}, y_{I,i})$ on the painting (see Figure 3(e))), we determine α_i of $e_i = (x_{I,i}, \alpha_i)$ by linear interpolation. Denoting e_i's left and right closest neighbors as $e_l = (x_{I,l}, \alpha_l)$ and $e_r = (x_{I,r}, \alpha_r)$, there is

$$\alpha_i = \frac{(x_{I,i} - x_{I,l})\alpha_l + (x_{I,r} - x_{I,i})\alpha_r}{x_{I,r} - x_{I,l}}. \tag{1}$$

Next, we proceed to compute p_i by q_i with inclination angle $e_{q_i} = \alpha_i$. We take an image slice (assuming a constant inclination angle in it) of the same height with the painting at q_i to illustrate the computation (refer to Figure 4). A local coordinate system $X_{I'}O_{I'}Y_{I'}$ is defined for the image slice. We denote the coordinates of q_i in the local coordinates system as $q_i' = (x_{I',i}, y_{I',i})$. There are

$$\begin{bmatrix} x_{I',i} \\ y_{I',i} \end{bmatrix} = \begin{bmatrix} x_{I,i} - x_{I,O'} \\ y_{I,i} \end{bmatrix}, \tag{2}$$

in which $x_{I,O'}$ is the X_I component of $O_{I'}$ in the global image coordinate frame. The bird view angle β_i of the local image patch relative to the 3D ground, that is, the pose of the local viewpoint, is determined by the inclination angle α_i at q_i,

$$\beta_i = 90^o - \alpha_i. \tag{3}$$

Considering that the viewlines of painters and viewers are vertical to moving-focus paintings, the projection from p_i to q_i' takes the form of parallel,

$$\begin{bmatrix} x_{I',i} \\ y_{I',i} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & cos(\beta_i) \end{bmatrix} \begin{bmatrix} x_{W,i} \\ y_{W,i} \end{bmatrix}. \tag{4}$$

By combining Eqs. (2) and (4) and the prior of p_i being a point on the 3D ground, that is, $z_{W,i} = 0$, we can obtain p_i,

$$\begin{bmatrix} x_{W,i} \\ y_{W,i} \\ z_{W,i} \end{bmatrix} = \begin{bmatrix} x_{I,i} \\ \frac{1}{cos(\beta_i)} y_{I,i} \\ 0 \end{bmatrix}. \tag{5}$$

$y_{W,i}$ equals the depth d of the ground-contact point q_i.

The preceding computation assumes the same scales in the painting and its 3D world. For practical use, such as audio field rendering, a world with physical dimensions (in meters) is required. To determine the scale parameter s which scales the physical world to the image space, we use an adult as a reference,

$$h = Ls, \tag{6}$$

where, h denotes the adult's height on the painting. L is a constant representing the average physical height of adults. Every q computed using the previous method should be divided by s for physical dimensions.

We do some preliminary evaluation on the 3D inference results as follows. By treating each pixel in the painting as a ground-contact point, we obtain the depth map of the ground (Figure 3(f)) using the aforesaid method. From Figure 3(f), we see that the depth of the ground is continuous. The variation of the ground in depth is consistent with that we perceive from the image. More importantly, it correctly reflects the relative distances between objects on the ground. We choose a few pairs of objects with equal pairwise *image distances* (see the red spots in Figure 3(d)) for illustration. The pairs are positioned on the 3D ground (as illustrated in Figure 3(g)) by the proposed algorithm. As we can see from Figures 3(g), the relative pairwise *spatial distances* are consistent with that we perceive from the image.

To compute the 3D position of a sound source associated with a figure, we first compute the 3D standing point of the figure using the method detailed before. Then, we move the 3D point with a quantity of the figure's height along the normal of the ground. In this way, we obtain the 3D positions of all the sound sources.

5. DATA ORGANIZATION

In this section, we introduce our considerations with regards to the data structure of the image and sounds. The well-designed data structures guarantee natural user experiences of the interactive exhibition.

5.1 Image Data Structure

The digital image is 151587×7469 (3.18G in the ".tiff" format). The painting is organized in a multiresolution pyramid form using Microsoft's ICE tool. This form provides great convenience for viewpoint-based rendering. Refer to Kopf et al. [2007] for details.

5.2 Audio Data Structure

More than 500 audio annotations (860M in all) are used (reflected in Figure 2). The textual scripts of the annotations are compiled by researchers majoring in the painting's historical culture. The audio data is made in a professional recording studio according to the sound scripts and the painting. Human voices are recorded in Mandarin with a Kaifeng accent from the Henan Province to reflect the life of ancient people. Other sounds are simulated by professional foley walkers. The starting time of each audio file in the stories is determined by professional mixers. All sounds are exported as single-channel files in ".wav" format.

We organize the voice dubbings in the form of short stories, encircled by rich sounds from nature. There are 50 stories in all for the whole scroll (as shown in Figure 2, each of them is marked by a blue box). To simulate a realistic auditory environment simply using point sound sources, we assign many virtual environmental sources to the painting. As shown in Figure 5, there are four sources for bird songs (numbered from 1 to 4) in four different places, for which the bottom two are added by assumption. The spatially distributed sounds are mixed to be a stereo audio field by 5.1-channel audio

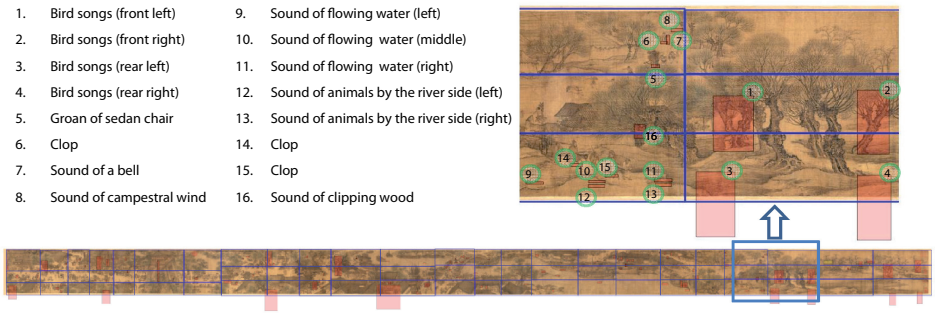

Fig. 5. The sounds from nature positioned on the scroll and the storage structure in blocks. (Image is courtesy of the Palace Museum (Beijing)).

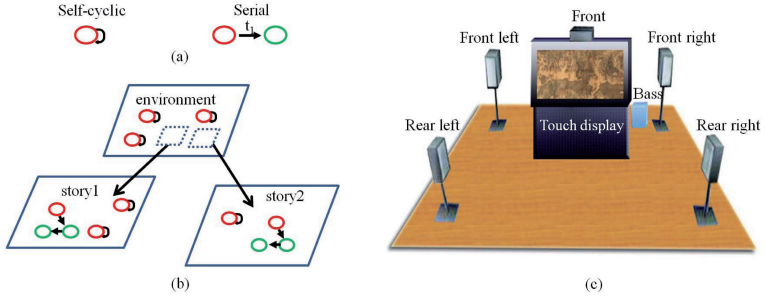

Fig. 6. Graph data structure with (a) two types of nodes and (b) two layers of sound sources. (c) Illustration of the platform (5.1-channel audio players and a touch display).

players. Visitors whose viewpoints enter into that 3D virtual space will be surrounded by the bird songs, just as if they were in a forest. Figure 5 shows all the sources for the sounds from nature that we set. Note that not all the natural objects have sounds, since too many sounds result in noisy sound fields.

We design a two-layer directed graph data structure to organize the two types of sound sources in spatial and temporal dimensions. In this graph data structure, two types of nodes are used (as shown in Figure 6(a)). The first is self-cyclic, denoting sounds from nature. The second type is for serial nodes. Each node represents a single sound source. A serial of connected nodes represents a story. The property of each node is the position of the sound source in the 3D space. A directed connection between two nodes represents their temporal relation. The weight of the connection, for example the t_1 in Figure 6(a), represents the time interval (in the unit of seconds) between the starting times of the two sound sources. The red nodes in Figure 6 denote trigger nodes, that is, entrance nodes. The trigger node of a sound from nature is itself. That of a story is the sound source that first speaks.

The two-layer graph data structure is defined as shown in Figure 6(b). Sounds from nature are associated to the image layer with $zoom = zoom_{min}$ and set to repeat at all times. We define a trigger area for each story (as indicated by the blue box in Figure 2), which covers the objects in that story. As shown in Figure 6(b), sounds from nature are also allowed in the story layer. A story is triggered once the visitors go near that story, that is,

(1) the center of the display lies in the trigger area of the story, and

(2) the trigger area of the story occupies 75% of the display in width.

Once it is triggered, the story will not stop until it reaches the end. Moreover, in order to prevent the unnatural-sounding repetition of human dialogues in the virtual space, stories are set to be single-round playing. Once a story is accessed, it will not play again unless the viewer goes away and returns.

In terms of storing audio data, the painting is coarsely divided into 24×3 (column $\times$ row) blocks (as shown in Figure 2). The sounds from nature and the stories are stored by blocks for view-dependent data prefetching and rendering. Two windows, auditory perception window w_a, and prefetching window w_p, are defined. Assuming w_v is the display screen, w_a is defined as $1.2w_v$, and $w_p = 1.8w_v$. At each moment, only audio data of the blocks whose centers lie in the prefetching window are loaded into the memory. Those out of that window are removed from the memory in time.

6. SYSTEM

The main hardware includes Intel Core2 Q8200 2.33G CPU, 4G DDR memory, HD4870 1G graphics card, Creative X-Fi Elite Pro audio card, touch display, and 5.1-channel audio players in the front, front left, rear left, rear right, and front right of the viewer, as illustrated in Figure 6(c).

The platform provides a panning and zooming interface to the viewer based on the touch display for smooth navigation of the image and auditory world. The head of the viewer is supposed to be at the center of the display. The viewer's viewpoint in the 3D hidden space is defined by: (1) the coordinates of the center of the touch display in the image coordinates system, denoted as q_v, and (2) the current value of *zoom*. First, q_v is treated as a ground-contact point in the image and its corresponding 3D ground point p_v is computed using the algorithm proposed in Section 4. Then, the viewpoint's 3D position is obtained by translating p_v along the projection line with a physical quantity T proportional to the current value of *zoom*, that is $T = Kzoom$, where K is an adjustable constant. Therefore, an object is nearest to us when we put it in the center of the display and zoom in to the maximum. The background process then activates the computation of a stereo audio field from the relative orientations and distances between the viewer and the sound sources in the 3D hidden space. The audio field is presented by the 5.1-channel audio players surrounding the viewer. The preceding computation on the viewpoint and the visual-audio rendering are completely real time.

Even if visitors observe the platform without interacting at all, we add an automatic display function in our system. In this display process, the scroll with $zoom = zoom_{min}$, automatically exhibits itself as the viewer is appreciating the Chinese scroll (from right to left) at a certain speed, along with a music clip describing the painting. Once the display is touched, the automatic display stops. The music fades out and then is replaced by sounds from nature. The automatic display starts again from the image position where the last visitor stopped, when the idle state (without interaction) lasts for a predefined time (40s in our system).

At the beginning of the project, we consulted historians, archeologists, and artists about the means of exhibition, in order to highlight the prosperous daily life of people during that ancient dynasty depicted by the painting. Without modifying, the masterpiece, we decided to use audial—including short dialogue stories and environmental sounds—to aid the visual during interactive user navigation of the painting. To ensure the correctness of the content of the dialogues, we invited historians and archeologists from the Palace Museum (Beijing) to write the script of dialogues. To ensure the exhibition system has the visual-audio consistency perception during user navigation, we invited several HCI and sound editing professionals to evaluate the system and immersive effect during the development of the system. The brightness and contrast of the displays are adjusted to make the digital version look close to the original painting as much as possible. The system has been in service in two halls of the Palace Museum, Hall of Martial Valor and the Southwestern Lofty Pavilion.

There are two limitations in our system. First, in case that a story has several speakers crowding together, it's hard to distinguish which person is talking. Second, as the painting has to be kept intact,

it is hard to provide a clue for a user to find remote stories that have not been loaded in the auditory perception window.

7. CONCLUSION AND FUTURE WORK

This article presented a system for exhibiting the painting of moving focus, the Bian River scroll, and its cultural backgrounds. The system utilized a high-resolution image of the painting and enhanced it with audio annotations. A solution to the moving focus problem was proposed. It puts the sound sources in the 3D world depicted by the painting, thereby producing stereo audio fields spatially consistent with viewers' visual perception. The audio and image data were well-organized for real-time rendering. A panning and zooming interface was provided for viewers to navigate the painting and the auditory world.

There are two significant things to do next. First, currently, the system is designed to serve for individual appreciation. In the future, we intend to target our system to allow simultaneous interaction of multiple users. To achieve this point, more sophisticated HCI technologies, such as attentive interface [Jaimes and Sebe 2007], can be explored. The second thing is to apply the proposed algorithm and system to the other ancient scrolls, which is of great significance in the appreciation of a variety of traditional Chinese artwork and the spread of information about ancient Chinese culture.

ACKNOWLEDGMENTS

The authors would like to thank the Palace Museum (Beijing) [2012] for providing the image and audio data and discussing the user interface; Luoqi Liu and Yang Liu for implementing the interactive association interface; and Miss Liz Carter and Mr. Matt Callcut for proofreading the article.

REFERENCES

CARROZZINO, M., EVANGELISTA, C., AND BERGAMASCO, M. 2009. The immersive time-machine: A virtual exploration of the history of livorno. In *Proceedings of the 3rd ISPRS International Workshop (3D-ARCH'09)*. 1–5.

CHINAONLINEMUSEUM. 2011. Painting gallery of zhang zeduan. http://www.chinaonlinemuseum.com/gallery-zhang-zeduan.php.

CHU, N. AND TAI, C. 2001. Animating Chinese landscape paintings and panorama using multi-perspective modeling. In *Proceedings of Computer Graphics International*. 107–112.

CRYSTAL. 2010. *Who Animate the Chinese Symphonic Picture Riverside Scene at Qingming Festival*. Ming Bao Publishers, Hongkong.

DEGENER, P. AND KLEIN, R. 2009. A variational approach for automatic generation of panoramic maps. *ACM Trans. Graph. 28*, 28, 1, Article 2.

FONG, W. C. 2003. Why Chinese painting is history. *Art Bull. 85*, 258–280.

FONI, A. E., PAPAGIANNAKIS, G., AND MAGNENAT-THALMANN, N. 2010. A taxonomy of visualization strategies for cultural heritage applications. *ACM J. Comput. Cultu. Herit. 3*, 1, Article 1.

FURUKAWA, Y., CURLESS, B., SEITZ, S. M., AND SZELISKI, R. 2009. Reconstructing building interiors from images. In *Proceedings of International Conference on Computer Vision*. 80–87.

HANSEN, V. 1996. The beijing qingming scroll and its significance for the study of Chinese history. *J. Sung-Yuan Studies*.

HORRY, Y., ANJYO, K. I., AND ARAI, K. 1997. Tour into the picture: using a spidery mesh interface to make animation from a single image. In *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*. 225–232.

JAIMES, A. AND SEBE, N. 2007. Multimodal human computer interaction: A survey. *Comput. Visi. Image Understand. 108,* 1-2, 116–134.

KOLLER, D., FRISCHER, B., AND HUMPHREYS, G. 2009. Research challenges for digital archives of 3d cultural heritage models. *ACM J. Comput. Cult. Heri. 2*, 3, Article 7.

KOPF, J., UYTTENDAELE, M., DEUSSEN, O., AND COHEN, M. F. 2007. Capturing and viewing gigapixel images. *ACM Trans. Graph. 26*, 3, Article 93.

LIN, J.-Y., CHEN, Y.-Y., KO, J.-C., AND ET AL. 2009. I-m-Tube: An interactive multi-resolution tubular display. In *Proceedings of the 17th ACM International Conference on Multimedia*. 253–260.

LUAN, Q., DRUCKER, S., KOPF, J., XU, Y. Q., AND COHEN, M. 2008. Annotating gigapixel images. In *Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology*. 33–36.

PALMER, S. E. 1999. *Vision Science: Photons to Phenomenology*. MIT Press.

RODERICK, W. 1965. Chang Tse-Tuan's Ch'ing-Ming Shang-Ho t'u. Doctoral dissertation, Princiton University.

ROWLEY, G. 1947. *Principles of Chinese Painting*. Princeton University Press.

SEOULALIVEGALLERY. 2008. Talking to mona lisa. http://wn.com/Alive_Gallery, Seoul.

SZELISKI, R. 2011. *Computer Vision: Algorithms and Applications*. Springer.

THE PALACE MUSEUM (BEIJING). http://www.dpm.org.cn/index1024768.html.

VALLANCE, S. AND CALDER, P. 2002. Multi-Perspective images for visualization. In *Proceedings of the Pan-Sydney Area Workshop on Visual Information Processing*. 1253–1260.

WILLATS, J. 1997. *Art and Representation: New Principles in the Analysis of Pictures*. Princeton Academic Press.

WITMER, B. G. AND SINGER, M. J. 1998. Measuring presence in virtual environments: A presence questionaire. *Presence 7*, 225–240.

YANG, X., BARNHART, R., NIE, C., CAHILL, J., LANG, S., AND WU, H. 1997. *Three Thousand Years of Chinese Painting (The Culture & Civilization of China)*. Yale University Press.

YU, J. AND MCMILLAN, L. 2005. Multiperspective projection and collineation. In *Proceedings of the International Conference on Computer Vision*. 580–587.

ZHU, Y., LI, C., AND SHEN, I.-F. 2004. A new style of ancient culture: Animated Chinese Dunhuang murals. In *ACM SIGGRAPH 04 Sketches*. 130.

Received October 2010; revised July 2011; accepted October 2011