INTERACTIVE VIEWPOINT-SPACE NAVIGATION FOR VISUAL-AUDIO EXHIBITION OF PAINTING

Wei Ma[1,2]*, Yang Liu*[1]*, Yizhou Wang*[1,2]*, Yingqing Xu*[3]*, Hongbin Zha*[2]*, Wen Gao*[1,2]

[1]Nat'l Engineering Lab for Video Technology, Sch'l of EECS, Peking University
[2]Key Lab. of Machine Perception (MoE), Sch'l of EECS, Peking University
[3]Microsoft Research Asia
Email: mawei@cis.pku.edu.cn, yang.liu.graphics@gmail.com,
yizhou.wang@pku.edu.cn, yqxu@microsoft.com, zha@cis.pku.edu.cn, wgao@pku.edu.cn

ABSTRACT

In this paper, we present a system for exhibiting a Chinese landscape painting about 900 years old. There are three parts in our system: (1) we allocate a voice dubbing or background music, which is treated as a point sound source, onto the 2D painting and obtain its position in the 2D space. All of the audio data are then located in a 3D hidden space, by projecting their 2D positions to the 3D space through a projection model. (2) A two-layer directed graph structure is proposed to well organize the audio data in a 4D space (with 1D temporal and 3D spatial). (3) The exhibition is defined as an active exploration in a viewpoint space, which faces both the image and the 3D world where the sound sources reside. The 3D space and the two-layer graph structure generate a natural and meaningful stereo audio field. Meanwhile, compared to videos with guided walk through, the active exploration makes the exhibition more attractive.

Keywords— Painting exhibition, viewpoint space, graph structure, 3D recovery, stereo audio field

1. INTRODUCTION

Digital exhibitions of ancient paintings become much popular in recent years. Most exhibitions borrow advanced technologies in computer science, such as 3D animation and voice recognition, to bring compelling user experiences. These exhibitions attract a large amount of population in concerning about the paintings. However, they emphasize so much on the fancy show of technologies that the exhibitions of the paintings as priceless artworks are neglected. An innovation in this area should be made to go further beyond the show of technologies.

In this paper, a new exhibition style is proposed based on a Chinese landscape painting, "Along the River during the Ching-Ming Festival "(Ching-Ming scroll for short). The exhibition is a totally interactive exploration of the painting, which is annotated with environmental sounds and voice dubbings. Different from previous exhibition forms which emphasize on presenting new fashioned experiences, the system given in this paper is designed to: (1) introduce the painting's historical culture (the living environment and daily life of the people in it) by a well-organized audio feast; (2) exhibit the artwork itself by preserving clean screen without any floating or popping-up sub-windows and providing deep zoom-in operations for appreciation of drawing details; and (3) bring people immersive experiences by producing natural visual-audio rendering effect.

To achieve these goals, we pose the interactive exploration as a free navigation in a viewpoint space through a panning and zooming interface. For real-time smooth navigation, the image is organized in a multi-scale pyramid form as done in [1]. According to the study of cognitive psychologists, people perceive the underlying 3D worlds of 2D images when looking at the 2D images. Therefore, for a audio field consistent with human's structural visual perception on the scroll, we position the audio annotations in the 3D hidden space depicted by the 2D painting. Stereo audio players are used to render the audio field. Both the pyramid images and the sound sources positioned in the hidden space are presented in front of the viewpoints. To well organize the sound sources in both the 3D space and the time dimension, we introduce a two-layer directed graph structure. One layer is for natural environmental sounds and the other is for performances happening in certain contexts. Each node in the graph represents a sound source attributed with its position in the 3D space. Connections between the nodes reflect their temporal or causal relations. During the smooth user navigation in the viewpoint space, audio annotations are informed in real time by the varying viewpoints and play by the constraints given in the graph structure.

The contributions of the paper include that: 1) a new exhibition form is presented in consideration of showing the ancient culture and the artwork; 2) a viewpoint space, involving pyramid images and a 3D hidden space, is proposed for interactive navigation; 3) a two-layer directed graph is introduced for data organization in a 4D virtual environment (with 1D temporal and 3D spatial).

This work is supported by Microsoft Research Asia eHeritage theme-based program research funding. The authors would like to thank Beijing Palace Museum for providing the image and audio scripts, and discussing user interface, Changning Gu and Sang Hu for producing the audio data, Luoqi Liu for implementing the interactive annotation interface, and Tingting Jiang, Bingshu Yang, and Yanhui Liang for revising the paper.

978-1-4244-7493-6/10/$26.00

Fig. 1. "Along the River during the Ch'ing-Ming Festival ", Hand scroll, ink and colors on silk, 24.8 x 528 cm

1.1. Related work

Many projects about painting exhibitions appear in recent years. Chu and Tai [2] rendered landscape paintings as 3D textured virtual scenes. This approach first separates some obvious standing objects from background and then interactively indicates multiple perspective views to recover the 3D scene using the method in [3]. Zhu et al. [4] proposed to tell the stories in the painting, "Dunhuang Murals ", by digitalized animation in 2D images. Analogously, the Ming-dynasty version of "Along the River during the Ch'ing-Ming Festival "preserved in the National Palace Museum, Taibei, became "alive"in 2007. Many scenes in the digital painting are enhanced with a set of 2D and 3D animations. When tourists visit a specific scene by touching the display, those corresponding movies will jump out the painting and start to play. Sankar et al. [5] built a prototype virtual tour of the Sri Andal Temple. They integrate technologies such as Photosynth and HDView to interactively explore visually complex sites. The above works clearly present the stories in the artworks for tourists with computer technologies, and meanwhile bring compelling user-end experiences. However, these forms ignore the show of the artworks. Popping-up windows as explanations interfere with tourists to appreciate the original drawing of ancient great masters. The animations with gross figures produced by modern designers cannot satisfy tourists' artistic appetite, especially when the tourists get familiar with the computer technologies involved.

In this paper, we present a new exhibition platform considering exhibitions of both artworks and historical culture. It is an continuously interactive navigation in the viewpoint space of a visitor. During the navigation, a smooth varying visual-audio feast is presented. The form preserves the original painting and provides close enough viewpoints to study the artwork. The audio data organization in three spatial dimensions and one temporal dimension produces natural stereo audio field.

This form looks similar to the work in [6]. However, our task is much more complicated. First, our target is a landscape painting which is generated by the artist's imagination and totally different from pictures captured by cameras. A special 3D recovery algorithm for positioning sound sources should be developed. Second, we are aiming at revealing the historical culture, especially the living environment and the daily life of the people in the painting, rather than introducing sites as done in [6]. The audio data should be well organized to clearly describe the stories happening in the painting.

1.2. About "Along the River during the Ch'ing-Ming Festival "

"Along the River during the Ching-Ming Festival "(Ching-Ming scroll for short) (Fig. 1) was created in Song Dynasty of ancient China and now is about 900 years old. Its reputation in China is comparable with the Mona Lisa in the Western Society. There appears an academic school of the Ching-Ming scroll to study the historical culture and arts in it. The entire piece is painted on a hand scroll composed of suburb, wharf, and urban scenes with moving focus. It captures the daily life of people from the Song period at the capital, PienJing (today's Kaifeng in Henan province). The painting is valuable as a work of art and as references to historical culture [7]. Therefore, a good exhibition fashion is important to show the artwork and the historical culture (i.e. the living environment and the lifestyle of the people in the painting) in it, and meanwhile attract people to concern with it.

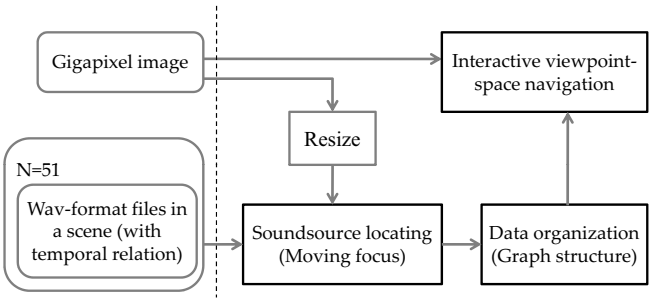


Fig. 2. System overview

1.3. System overview

Our input data include a high-resolution image of the painting for appreciation of drawing details and rich audio data for interpreting the painting. In implementation, the painting is decomposed into 51 scenes. Most of the audio processing steps, including the designing of the audio scripts, recording of the sounds, annotation and organization of the audio, are performed scene by scene. The compiling of the audio scripts and the recording of the sounds refer much to the study on the painting's historical culture.

For immersive experiences, we choose interactive exploration rather than guided walk through. The exploration is defined as a free navigation of a tourist in a continuous viewpoint space. At each viewpoint, corresponding image parts and sounds are presented to the tourist. Considering cognition psychology, people perceive 2D images in 3D forms. We define the user's viewpoint and all the sound sources in a 3D perceptual space and render the stereo audio field surrounding the viewpoint. The stereo helps the tourists perceptually sense the environment by ears [8].

Besides the concept of the viewpoint space, two main steps are involved (as shown in Fig. 2). The first one is the locating of sound sources in the 2D painting and then in the 3D hidden space. In order to annotate the figures in the painting with sound sources, we develop an interactive user interface. Within this interface, we relate each sound to a figure patch in the painting. Since this part involves no algorithm, we skip it in the main body. After annotating a sound source, a 3D inference algorithm is performed to obtain the 3D position of the source. As shown in Fig. 2, we choose a resized image for fast computation in this step. Normalized image coordinates are adopted in our exhibition system for dealing with images with different resolutions.

The second step is the organization of the large image data and rich audio data. The image is organized in a multi-resolution pyramid form as done in [1]. To well organize the audio data for natural and meaningful audio field, we propose a two-layer directed graph structure: one layer for environmental sounds, the other for sounds happening in certain contexts. Each node represents a sound source, attributed with a 3D position. Connections between nodes reflect their temporal or casual relationships. In the second layer, sounds are organized with stories as the base unit in the graph structure, since the voice of an individual object makes no sense.

The remaining is organized as follows. In section 2, we introduce the concept of the viewpoint space. In section 3, we present how to infer the 3D position of a sound source after relating it to its figure in the painting. Then, we describe the two-layer graph structure in details in section 4. Results are given in section 5. Conclusions and future research directions are summarized in section 6.

2. VIEWPOINT SPACE

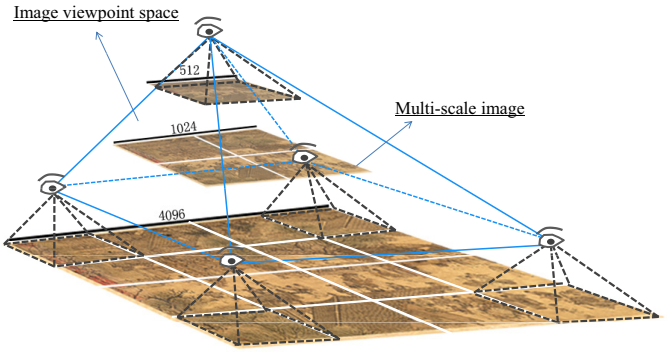


Fig. 3. Image viewpoint space

In this section, we introduce the concept of the perceptual viewpoint space. For clarity, we first describe the viewpoint space in front of the painting. Then, we extend it to the perceptual viewpoint space facing both the 2D painting and the 3D hidden space recovered by the painting to hold sound sources.

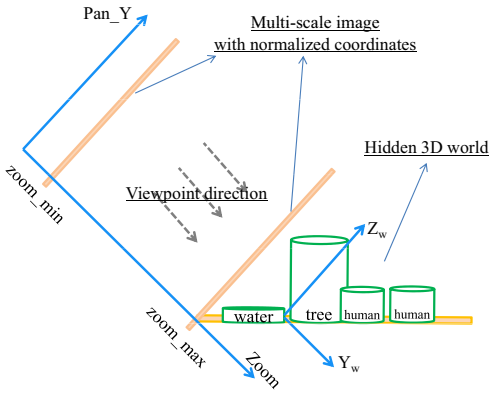


Fig. 4. Perceptual viewpoint space

For detailed appreciation, we adopt a high-resolution image (gigapixel image). A panning and zooming user interface is provided for smooth navigation in the gigapixel image, by using the deep zoom technology which is developed based on the idea in [1] for web applications. The deep zoom technology organizes the gigapixel image in a multi-resolution pyramid form. As illustrated in Fig. 3, the bottommost layer is the full resolution data and a upper level has the resolution of a quarter of its lower one. A viewpoint in front of the images is defined as a position, a direction and a viewport (as illustrated in Fig. 3). Here, the viewport is the display screen. The direction is assumed to be perpendicular to the screen. The position is the center of the screen under the coordinates system of the scroll. It continuously changes during exploring the scroll through the zooming and panning interface. The definition of the above viewpoint properties can be improved, for example by capturing eye-gazing as the viewpoint direction, for much friendlier human computer interface (HCI) in future work. Which layer and local parts in that layer should be presented depends on the properties of the viewpoint (as illustrated in Fig. 3).

As aforementioned, all sources are positioned in a 3D hidden space depicted by the 2D painting. The navigation in the image viewpoint space should be consistent with the navigation in the 3D hidden space. We name the viewpoint space facing both the images and the 3D space as a perceptual viewpoint space.

Fig. 4 presents a side view of the space. The multi-scale images are projected to be normalized coordinates, which are independent of the image scales. The image viewpoint space is defined by the main axes Pan_X, Pan_Y and $Zoom$. The 3D hidden world is positioned behind the last image layer as shown in Fig. 4, defined by the main axes X_w, Y_w and Z_w. In next section, we will explain the method which relates the last image plane to the 3D space. The difference between the last image plane and the others is their positions along the $Zoom$ axis which is parallel with the Y_w axis. This difference is easily solved by a 1D translation and a scaling operations. Thus, given a viewpoint position in the image viewpoint space, we can obtain its corresponding position in the 3D hidden world. In this

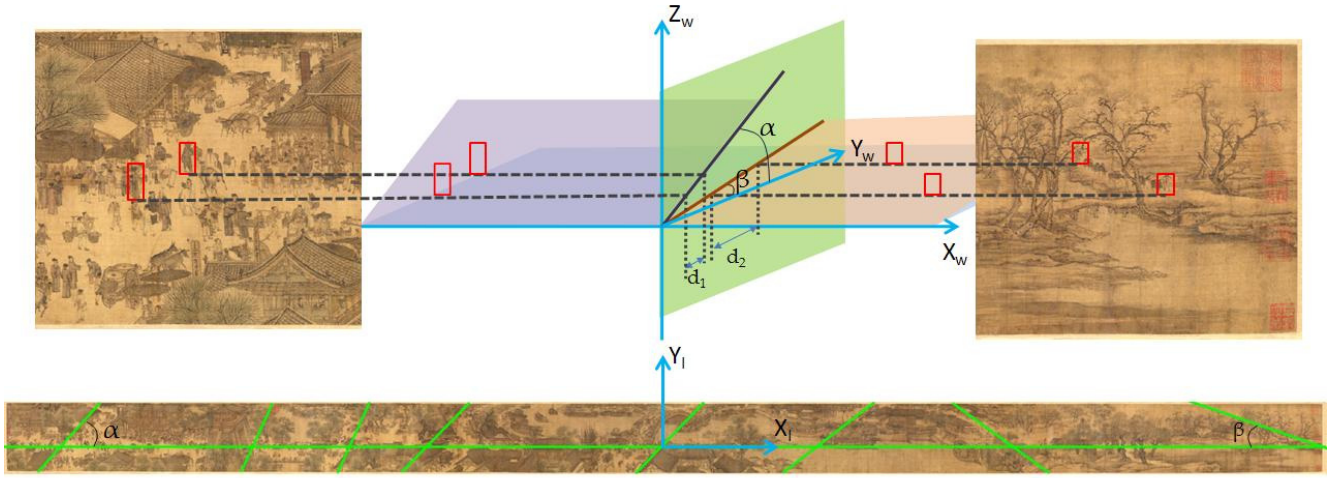


Fig. 5. Intrinsic structure of the painting

way, we build the correspondence between the two spaces. The two together form the perceptual viewpoint space.

3. MOVING FOCUS

In this section, we explain how to compute the 3D position, given a sound source and its corresponding image patch in the painting. Different from western paintings, Chinese landscape paintings are produced through moving focus: painters draw what he want to express on a canvas with continuously varying viewpoints. We cannot directly infer the correspondence between the 3D hidden space and the 2D Ching-Ming scroll using traditional perspective methods [9] [10]. A solution to the moving focus problem is presented as follows.

Considering the moving viewpoints, we adopt the orthography projection model. In this case, the key problem becomes the depth inference of the scroll. To express the spatial elongation in depth on the 2D canvas, the painter adopted a series of skills. These skills form our basic foundation to infer the correspondence between the 2D and its 3D space, and further locate each sound source in the 3D space.

According to the analysis of the Ching-Ming scroll scholars, the spatial structure is realized by positioning structured objects along a certain angle to create a backward prolonged effect. For example, the road in the top right image in Fig. 5 makes this scene look long extended to the far. On the contrast, the directions of the upward road in the top left image in Fig. 5 indicate that this scene is not so long extended as the top right one. Based on this observation, we draw the slanting lines in the painting interactively (as indicated in Fig. 5). Here, the slanting lines mean apparent cues that the painter adopted to depict the elongation of the painting in depth, such as the roads extending from the bottom to the top in Fig. 5. The lines reflect the slanting angles of grounds in the 3D space(as indicated by the two αs and the two βs in Fig. 5), which is positioned behind the image. In this way, we build the correspondence between the image ground and the 3D ground. By this correspondence, we can position the objects at different scenes in a global coordinate system.

In implementation, as shown in Fig. 5, a speaker in the painting is represented as a bounding box, whose top point is the sound source position and its bottom point is the ground-contact point. To compute the 3D position of the sound source, we first project the ground-contact point to the slanted ground using the orthography projection, and get the ground-contact point in the 3D hidden space. Next, the sound source position is computed by translating the 3D standing point with a quantity of the height of the bounding box along the direction of the ground normal. As shown in Fig. 5, the human figures in the top left and the top right pictures are positioned on their slanted grounds respectively. The contrast of their relative distances, d_1 and d_2, in the 3D space (as shown in Fig. 5) are consistent with our visual perception from the painting. For one object standing on the other one, for example a riding man, we draw the bounding box covering the two, since we use ground-contact points to infer source positions. In this way, we build the 2D to 3D mapping for the painting. Up to now, we use pixels as the metric units. To determine the physical position, we simply multiply the 3D positions by a scaling factor, computed by a standing human's pixel height and a given physical height (1.65 meter). Besides, scale adjusting along axis Y_w (i.e. the depth axis) is also allowed in our system for scalable depth elongation in the 3D world.

4. GRAPH STRUCTURE

For well organization of the sound sources in the 4D space (with 3D spatial and 1D temporal), we introduce a directed graph structure, in which each node represents a sound source attributed with its 3D position. Connections between nodes represent their temporal or casual relations. Here, all sounds are treated as point sound sources. To realize surrounding auditory effects, we use multiple point sources of the same subject, as we did for the rural environments and the bird tweets in Fig. 6.

The sound sources are divided into two categories. One is environmental sounds. The other one is sounds having casual relationship with others. The latter appears in the form of stories, also called performances, since words of individual humans without context makes no sense. One big problem for the second type of sounds is that their repetitiveness brings unnatural auditory effects in virtual scenes. Therefore, we design the graph structure to be two layers, which correspond to the two types of sounds respectively. The first type can be cycled infinitely. The second type is triggered only when it satisfies predefined spatial conditions. When tourists fast look through the painting in a large distance, only environmental sounds are presented. Those who have interests in performances, then draw close to learn the events happening in the environments. Notice that all the sounds in the two layers are positioned in the same physical hidden world.

1. Rural environment
2. Rural environment
3. Rural environment
4. Bird tweet
5. Bird tweet
6. Footstep (Walking human 1)
7. Footstep (Hamal 1)
8. Groans of the sedan chair

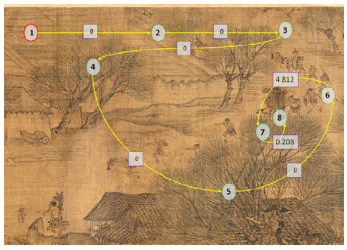


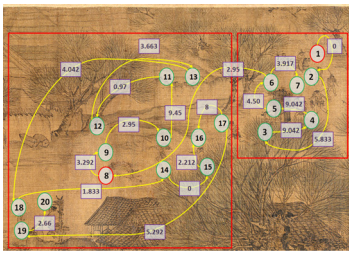
Fig. 6. Environment-layer of the graph structure

Considering that tourists may have no patience to stay at a single scene until its long story finishes, a long story is segmented into several substories. Nodes in each substory are temporally organized. Then, we add a connection between two nodes in two different substories, if the two nodes have casual relationship. Each substory has an entrance node, i.e. the node speaking first. As shown in Fig. 7, this story is divided into two substories. The left substory has no relation with the right one, but the 6th node in the right one relies on the 8th node in the left one. Tourists can enter the story from either of the two entrances (the red nodes in Fig. 7), depending on their viewpoint positions. The way starting from the right entrance will go on with two paths simultaneously from the 6th node (as indicated by the connections in Fig. 7).

We have a specially designed user interface to finish the above tasks. First, given the wav-format files each with a starting time, we interactively put each wav file to its image position and give it its layer label: environment layer or performance layer. Each wav file forms a node and its attributes, i.e. its 3D position, is computed by the moving focus algorithm. Next, we manually indicate the substories by drawing their triggering areas, as shown in Fig. 7 (the red wireframes). The nodes in each substory and the environmental layer are ordered automatically. At the same time, the entrance nodes are determined as the node speaking first in each substory. Finally, we manually relate the substories by their casual relation.

One entrance will be valid if its corresponding substory satisfies its spatial trigger condition (being almost full of the screen). A tourist can go away during the playing of the story. All the voices would become weak due to their gradually increasing distances to the viewpoint. Notice that not all objects in the scene are assigned with sounds, since too many sounds will make the audio field noisy. Another thing we want to point out is that for consideration of memory, if an object speaks intermittently, its voice is divided into several segments to omit blank parts. For example, the neigh of the crazy horse is divided into two segments as shown in Fig. 7.

1. Horse bell
2. Hoofboot
3. “Excuse me, give us the way” (Hamal 1)
4. “My lady, my lord passed a word from the back: we have finished the task of visiting the grave. Before we go home, let’s go and buy you and your girl some clothes ” (Hamal 2)
5. “That’s OK, the lord makes the decision” (Madame in the sedan chair)
6. “What’s happening?” (Madame in the sedan chair)
7. “There darts a crazy horse. It’s running like hell!” (Hamal 2)
8. Hoofboot (Crazy horse)
9. Neigh (Crazy horse)
10. “There is a kid! Watch out” (Horse tracer 1)
11. “Hurry up, little horse, hurry up! Do not let them catch up you! ”(Child 1)
12. Neigh (Crazy horse)



13. “Come on, horse!”(Child 2)
14. Running footstep(Tracer 1)
15. Running footstep (Tracer 2)
16. “Be careful of the kid!” (Tracer 2)
17. “Crazy horse, stand still!” (Tracer 3)
18. Mirth (Kid)
19. “Mom, Mom” (Kid)
20. “Come here, baby, come here” (Mom)

Fig. 7. Performance-layer of the graph structure

5. RESULTS

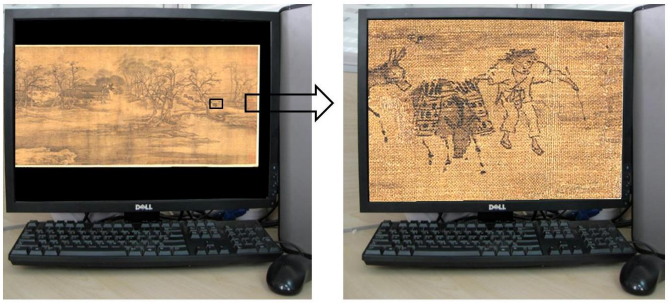


Fig. 8. Left: initial screen; right: zoom in to the detail (around 20 times larger than its corresponding part in the left image)

The used image has 1.08 gigapixel in all. It is provided by Beijing Palace Museum, China. The image, even the canvas texture (as shown in Fig. 8), becomes very clear when the

viewpoint is close enough. This provides great opportunity for tourists to observe the artistic details. The audio scripts of the 51 scenes are derived from the professional study on the painting's historical culture by the researchers in the Beijing Palace Museum. The audio data is made in a professional recording studio. Human voices are recorded in Putong hua with the accent of Kaifeng in Henan Province to reflect the life in that dynasty. Other sound effects are simulated by professional foley mixers. The starting time of each audio file is determined by audio mixers. All sounds are exported as single-channel point sound sources with the starting time as theirs names. The audio data is more than 100M per scene in average.

The main hardware includes Intel Core2 Q8200 2.33G CPU, 4G DDR memory, Creative X-Fi Elite Pro audio card, and stereo audio players. For real-time navigation, we define three windows, as shown in Fig. 9. Visual window means the screen. Sounds in the scenes which lie in the auditory perception window are played according to the temporal constraints and spatial trigger conditions. Only audio data of the scenes in the dispatch window are loaded into the memory.

The interactive navigation is totally in real time. We have asked around twenty people knowing little about the Ching-Ming scroll for test. They show great interest in this exhibition style. The visual-audio feast amazes most of them. The auditory surrounding effects help the users find the locations of the sound sources, since some parts are visually hard to recognize due to the long time. Our platform will be put in Beijing Palace Museum after a period of improving on both algorithms and user interface, and stability testing.

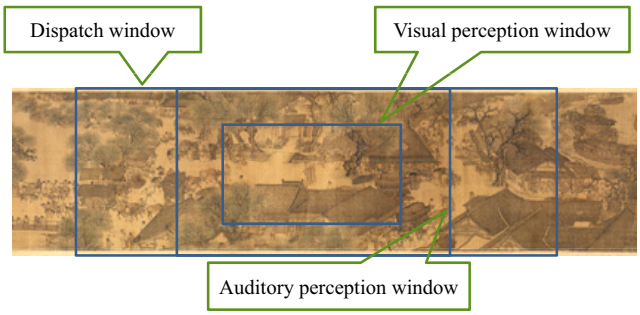


Fig. 9. Perception and dispatch windows

6. CONCLUSION AND FUTURE WORK

We presented an exhibition platform and its implementation procedures. The interactive exhibition is considered to be a free navigation in a viewpoint space, which unifies the multi-resolution pyramid form of the painting and the sound sources residing in the 3D hidden space depicted by the painting. A two-layer directed graph structure is proposed for well-organizing the sounds in three spatial and one temporal dimensions.

The viewpoint space and the graph structure provide great conveniences for future extension. Special equipments and algorithms can be used for more friendly HCI-based navigation. Which figure or story attracts more attention of tourists can be statistically estimated based on the graph structure. This function provides valuable references for improving the audio scripts. Before the implementation, we investigated tens of people as references to the system design. After finishing the early form of the system, we also did pilot test on it. In the near future, we will do further investigation and improve the system by the feedback.

7. REFERENCES

[1] J. Kopf, M. Uyttendaele, O. Deussen, and M. F. Cohen, "Capturing and viewing gigapixel images," *ACM Transactions on Graphics*, vol. 26, pp. 93:1–10, 2007.

[2] N. Chu and C. Tai, "Animating Chinese landscape paintings and panorama using multi-perspective modeling," in *Proceedings of Computer Graphics International*, 2001, pp. 107–112.

[3] Y. Horry, K. I. Anjyo, and K. Arai, "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*, 1997, pp. 225–232.

[4] Y. Zhu, C. Li, and I-F Shen, "A new style of ancient culture: animated Chinese dunhuang murals," in *ACM SIGGRAPH 2004 Sketches*, 2004, p. 130.

[5] A. Sankar, J. Joy, A. Prasad, and N.Datha, "Digital heritage," in *the Video Showcase at CHI 2009*.

[6] Q. Luan, S. Drucker, J. Kopf, Y. Q. Xu, and M. Cohen, "Annotating gigapixel images," in *Proceedings of the 21st annual ACM Symposium on User Interface Software and Technology*, 2008, pp. 33–36.

[7] V. Hansen, "The beijing qingming scroll and its significance for the study of Chinese history," *http://www.yale.edu/history/faculty/materials/hansen-qingming-scroll.html*.

[8] K. K. P. Chan and R. W. H. Lau, "Distributed sound rendering for interactive virtual environments," in *IEEE International Conference on Multimedia and Expo*, 2004, pp. 1823–1826.

[9] D. Hoiem, A. A. Efros, and M. Hebert, "Putting objects in perspective," in *Proceedings of the 2006 IEEE Conference on Computer Vision and Pattern Recognition*, 2006, pp. 2137–2144.

[10] B. C. Russell and A. Torralba, "Building a database of 3d scenes from user annotations," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 2711–2718.