Time_lapse: immersive interaction with historic 3-D stereo images

Christa Sommerer, Laurent Mignonneau, Roberto Lopez-Gulliver

ATR Media Integration & Communications Research Laboratories
2-2 Hikaridai, Seika-cho, Soraku gun, 619-02 Kyoto, Japan
{christa, laurent, gulliver}@mic.atr.co.jp

Keywords
Virtual Environment, Human Computer Interaction, Immersion, Stereoscopic Vision

Abstract

Mankind’s fascination with three-dimensional images and stereo vision can be traced back to the ancient Greeks, and it was photography and the invention of the stereoscope during the mid-nineteenth century that helped to spur large-scale public interest in stereo images and stereo viewing techniques. Stereoscopes made it possible to see a pair of slightly different two-dimensional photographs in three dimensions by making use of the human eye’s parallax. With this century’s new possibilities in three-dimensional computing, historic stereoscopic images can now be used for the creation of interactive virtual environments. Driven by the wish to make historic stereo images interactively accessible to the user, we have developed an immersive virtual environment called "Time_lapse." This system allows two remotely located users to enter and interact with historic stereo images through full-body integration and immersion.

1. A Brief History of Stereoscopic Vision

1.1. Early forms of stereo vision

The fascination with the concept of stereo vision goes back to ancient Greece around 300 B.C. Euclid for the first time explained the principle of binocular vision. He demonstrated that the right and left eyes see slightly different versions of the same scene and that the merging of these two images produces the perception of depth. In the sixteenth century, Leonardo Da Vinci experimented with perspective in an effort to create the impression of depth in his paintings. Around the same time, the Florentine painter Jacopo Chimenti created pairs of "stereo" drawings. During the Industrial Revolution, demand for more sophisticated forms of viewing resulted in the developments of new techniques, including magic laterns, the polyorama panoptique, peep shows and the kaleidoscope. [1]

1.2. Brewster’s Stereoscope

However, it was the invention of photography that really made mass-culture 3-D viewing possible. The first patented stereo viewer was Sir Charles Wheatstone’s reflecting stereoscope in 1838. The device was a bulky and complicated contraption that utilized a system of mirrors to view a series of pairs of crude drawings. In 1844 a technique for taking stereoscopic photographs was demonstrated in Germany, and a much smaller and simpler viewer that utilized prismatic lenses was developed in Scotland by David Brewster (Fig. 1).
1.3. Crystal Palace - a breakthrough for the stereoscope

The real breakthrough for Brewster’s Stereoscope came in 1851 with the opening of the Great Exhibition in London’s Crystal Palace located in Hyde Park. Many countries of the world were represented in an extravagant display housed in this huge glass building designed by engineer Joseph Paxton. Its beautiful domed roof (Fig. 2) made it the perfect setting for the stereo photographs taken by the company Negretti and Zambra. The various attractions in the Crystal Palace included "The Medieval Hall," "The Lotus Pond," "Egypt" (Fig. 3), and "Rome," among others. When Queen Victoria took a fancy to the Stereoscope at the Crystal Palace exposition in 1851, stereo viewing became vastly popular in Britain. [2]

1.4. Touring the world from home

The stereoscope slides that were produced allowed people to sit in their own home and tour the world. The most popular slides were travelogues that showed the world from the abbeys and countrysides of Europe to the pyramids and tombs of ancient Egypt, to the Great Wall of China, and to the Taj Mahal (Fig. 4). The great events of the day found their way onto the stereo slides. The building of the Panama Canal, the terrors of war, and the destruction of natural disasters such as earthquakes were brought into homes in much the same way as television does today. By the 1870’s, local commercial photographers had sprung up around the country, and for a fee they would produce stereo slides of one’s farm, family or shop (Fig. 5).
2. "Time_lapse:" immersive interaction with 3-D stereo images

Inspired by the collection of historic stereo images at the Tokyo Metropolitan Museum of Photography [3] and the wish to create an interactive virtual environment where users could interact with these historic images, we developed an immersive virtual environment called "Time_lapse." This system allows two remotely located users to enter and interact with the historic stereo images. To adapt the images for the construction of our three-dimensional immersive interactive environment and to provide the user with an immersive interaction experience, various image preparation processes were needed. These are described below.

2.1. Image preparation

The following sections describe the image processing steps performed on the stereo images to make them suitable for interactive viewing and exploration by the user. There are three main processes: image acquisition, depth extraction and virtual views generation for motion parallax. The image’s depth is used to achieve the interposition depth clue, where near objects occlude far objects. Virtual views are synthetically generated to enhance the depth effect provided by the motion parallax clue, where near objects move more than distant ones.

2.1.1. Image acquisition

A total of 15 historic colour stereo images were selected from the collection of the Tokyo Metropolitan Museum of Art. Left and right image pairs were scanned from a book[1] at 300 dpi and then scaled down to a 512x512 pixel resolution. No colour adjustment was made during the scanning process. Due perhaps to their antiquity, the images’ quality is anything but optimal. Judging from the era they were taken, we can conclude that they are indeed black and white images. However, some of them seem to be retouched to give them a coloured appearance. Also, the colours are fading away. These image sets could be viewed with a rather simple stereo-viewer provided with the book. The images were taken using stereo cameras with lenses separation, also called inter-ocular separation or baseline, that is roughly equal to the average human eye’s separation of 55-65 mm. Most of the scenes were carefully chosen to give a relatively good stereoscopic effect in spite of the short lens separation.

2.1.2. Depth extraction

Due to the less than optimal quality of the scanned images, the image colours between left and right images differed significantly, so a general correlation-based stereo matching algorithm to extract a dense depth map was not appropriate. However, we applied a recursive and adaptive stereo matching algorithm [4], based on the Kanade-Okutomi[5] algorithm, to our images. No image
rectification was needed since the camera lenses’ setup assures that epipolar lines coincide with raster scan lines in the image. Depth extraction results were, as expected, not very precise (Fig. 6).

![Fig. 6 Depth extraction results](image)

The extracted dense depth map was too noisy, boundaries of objects were not sharply defined, low-textured areas such as the floors looked messy; these are some of the weaknesses of any correlation-based stereo algorithm. Then we considered using other algorithms in the literature[6,7]. Although some worked better than others in certain areas, the depth map still had to be retouched and cleaned up to be usable. Therefore, we decided to "paint" the dense depth map by hand. Some automation or guide while doing this was strongly suggested before we started our long, meticulous and somewhat tedious task. Given that the image scenes were relatively simple, mostly planar objects with well defined edges, we used the image results of a colour segmentation algorithm in [8,9] and our previous depth map from the automated methods above to serve as guidelines in performing the task. That is, edges in the colour segmented image can delimit objects and thus identify areas of constant depth, or gradation depth, as "blobs." We painted these areas accordingly by referring to the actual depth values in the extracted depth map. A sort of gradient interpolation (gradation) was used in floors/ground between near and far ends and in some other sections to accommodate the non-planar nature of certain objects in the scenes. We used a ramp of gray-scale shades ranging from black to white with pixel values of 0-255. In Fig. 7, image (b) shows a finished dense depth map done this way and image (a) is the colour-segmented image used as reference.

![Fig. 7 Finished depth map (b) and color-segmented image (a) used as guideline.](image)

We realize that this method lacks any precise algorithmic nature and is prone to errors, but it indeed helped us to simplify and solve the problem in a relatively easy and fast way: combining automated algorithmic processes with hand work. However, more complex scenes could never have been processed in a satisfactory way by this method. Since both images of a stereo pairs required a depth map, we implemented a simple algorithm to recreate one depth map from the other. The idea is simply based on the definition of disparity for
stereo images[10]. Let’s suppose a dense depth map corresponding to the right stereo image, say $DM_r(x, y)$, is given. Then the dense depth map for the left stereo image can be calculated as:

$$DM_l(x, y) = DM_r(x_o, y)$$

where: $x_o = x - DM_r(x, y)$

for all possible values of $x$ and $y$ in the image. Pixels where $x_o$ is negative or out of the image size range are discarded. In our case, before applying this method, the painted depth map image had to be rescaled from its gray scale values, 0-255, to the actual min-max disparity values of the image under consideration. For this we needed to scale the min-max disparity values of the image in pixels, accordingly.

Figure 8a shows an original right image, and Figure 8b is its corresponding right depth map. Applying the above method to Fig. 8b creates the corresponding left depth map image, as shown in Fig. 8c. Unavoidable gaps or holes appear, corresponding to the areas not visible from the right viewpoint but visible from the left viewpoint. Black pixels in Fig. 8d show the same holes with right and left viewpoints interchanged. We then proceeded, by simply raster scanning the image’s lines from left to right and filling the holes, black pixels, with the last non-black pixel encountered and resetting it whenever any non-black pixel is found. Note that this filling method fails to work when the newly exposed sections contain an object different from the ones seen from the right image. There is no general algorithm to solve this problem, mainly because of the lack of information, but some partial solutions have been proposed[11,12,13]. Figure 8e shows the final left image’s depth map for the left image with holes filled. In order to validate our results, the newly calculated left image depth map was passed through an edge-detector filter and superimposed with the original image, blending them in a 50%-50% proportion. Figure 8f is a close-up showing how well the edges of the so calculated left image’s depth match the contour of the people in the scene.
2.1.3. Virtual views generation

This section presents a simplified algorithm that synthetically generates virtual views to achieve motion parallax and enhance the stereoscopy effect. The algorithm assumes that the viewpoint is only moving horizontally, in a plane parallel to that of the image. More general algorithms can be found in the image-based rendering literature[11,13,14,15]. However, this algorithm is easy to implement and does not need the camera calibration and/or multiple views required by others. The idea is to divide the image in a specified number of accumulated depth layers, say l_1, l_2, ..., l_n, by using the depth map. That is, given that Img(x, y) is the original image and DM(x, y) is the images’ disparity map:

\[
    l_k(x,y) = \begin{cases} 
    \text{Img}(x,y) & \text{if } d_m < \text{DM}(x,y) < k \times \text{deltaD} \\
    \text{alpha}=0(\text{transparent!}) & \text{otherwise}
    \end{cases}
\]

where:
\[
\text{deltaD} = (d_m - d_M) / n, \\
\text{n = number of layers} \\
\text{d_m = minimum of DM(x,y)} \\
\text{d_M = maximum of DM(x,y)} \\
\text{k in [1, n]}
\]

Fig. 9 shows the l_6 and l_13 of one of the sample images with n=15.

![Fig. 9](image_url)

**Fig. 9** Two depth layers created from the original image.

For each virtual view we want to render, the layers are perspectively shifted according to their depths and rendered in a back-to-front fashion on top of each other. The resulting image shows a perspective of near objects moving more than distant ones. Figure 10 shows one such a generated view, where we have exaggerated the shifting for explanatory purposes. Note how the statues on the center right moved less than front black "bench" in the right lower corner of the image.
Objects shifting themselves drastically compared with the background would produce holes between themselves and the background. By accumulating the depth in every layer, the object simply repeats itself in its boundaries filling the holes. The error introduced this way is hardly noticeable and acceptable for our purposes, given that in our case the maximum variation is no more than twenty pixels. Up to 15 such virtual views are presented to the user as he/she walks from the extreme right side to the left side of our studio setup. The small animation in Fig. 11 better illustrates this. The depth map was processed in exactly the same way and was synchronised with the virtual view shown during real-time interaction to avoid depth-scene discrepancies. Note that the real setup used a stereo version of this.

Given that our camera setup, up to 3 meters from the user, makes it impossible to track the user’s head as in [16], we use his/her current position to provide the appropriate view.

2.2. User’s real-time interaction in "Time_lapse"

In section 2.1., we explained how we constructed a three-dimensional virtual environment by extracting the depth information of the historic stereo images. Having now obtained all the depth values of these stereo image pairs, we can begin integrating the user into the system to make the environment interactive and to provide the user with an immersive feeling.

2.2.1. Real-time three-dimensional integration

This interaction environment consists of a 4 x 2.1 meter white floor in front of a light box background. A luminance key technique was used to extract the user’s image and contour, as shown in Fig. 12.
The integration process consists of the following steps: 1) the user’s image is subtracted from the background using colour/luminance keying, 2) the user’s position is tracked by using a camera tracking system called "Pfinder"[19], making the user a flat image with constant depth, and 3) the depth map obtained in the previous sections acts as the so-called z-buffer. After this, the user’s image, as a plane, is correctly positioned in depth according to his/her current position. As a result, the user finds him/herself displayed in depth within the historic stereo image scene and perceives a feeling of immersion. The user can interact with the system by driving it with body gestures as shown in Fig. 13.

Four basic commands are provided for the user: inviting the remotely located person into one’s image scene by bowing towards him/her; taking a snapshot of both image scenes by "virtually" standing beside each other and changing the displayed image scene forward/backward by raising his left/right hand, respectively.

2.2.2. Implementation details

The system consists of three main modules: The virtual environment manager (VEM), the gesture recognition module, and the user image capturing module. The VEM is in charge of: 1) mapping
the user’s body gestures to actions in the virtual world, 2) using the user’s current position to select
the appropriate virtual view of the scene and position the user in the appropriate depth plane in the
scene, and 3) rendering the displayed image by composing the user’s image and the virtual view
scene. The VEM is implemented in C using the OpenGL API. One SGI OnyxIR and two Indys are
used for the rendering and gesture recognition modules. They communicate via rpc and are ethernet
connected. The stereo shutter glasses (StereoGraphics’ CrystalEyes) and emitter are connected to
the Onyx, and the screen projector is synchronised with it at 120 Hz for stereo displaying. Many
useful implementation considerations were taken from[17].

3. Telematic Interaction, Ubiquity and Virtual Sightseeing

Because the image scenes in the two remote interaction environments are different from each other,
users who experience these environments do not really know in which other image scenes they are
being simultaneously displayed. While they can see the other user within their own interactive
environment (given that they "invited" the other user through the "bowing" gesture described in
Section 2.2.1.), they cannot see themselves in the remote image scene. This implies a certain sense
of disembodiment, enabling the user to be visually present at several locations at the same time.
Presence is usually understood as "both a subjective and objective description of a person’s state
with respect to an environment," [20] and the notion of being present in a remote place is
commonly called telepresence [21] [22]. "Time_lapse" deals with the additional concept of
"ubiquity," a distributed and multiple presence of the user’s image in several distinctively different
image environments. This idea of "ubiquity" seems especially suitable for applications in games or
interactive art exhibits as they widen the user’s conventional perception of space, presence and
telepresence.

4. Conclusions and Future Applications

Our interactive immersive environment "Time_lapse" allows users to experience historic stereo
images by virtually stepping into them and interacting with them. It also enables remotely located
users to telematically interact with each other in three-dimensional image scenes. As the images at
the remote sites are not the same, the user will experience a sense of disembodiment and ubiquity.
In the future, several such systems on remote locations, each with its own image scenario, could be
linked together via data connection to provide interactive and telematic experiences. Similar to the
idea of "touring the world from home through stereo images in the mid-19 century," (Chapter 1.4.)
one could imagine a "Time_lapse" system for virtual and interactive sightseeing. Further
applications could include entertainment, interactive art exhibits and telecommunicative
environments. Possible future enhancements to the systems include using movies instead of static
images and adopting light and reflectance effects [18] to add more realism to the integration.

References

[1] T. Moriyama, "Whither Love of 3D - 3D Love Afterwards", In 3D-Beyond the Stereography,
Tokyo Metropolitan Museum of Photography, 1996, pp. 17.-23

[2] Parts of this chapter refer to: http://www.bitwise.net/~ken-bill/stereo.htm

[3] Tokyo Metropolitan Museum of Photography:
http://www.tokyo-photo-museum.or.jp/eng/index.html

3-D Panoramic Photographic Scene", ICAT’98 International Conference on Artificial Reality and


