Harnessing Mobile Ubiquitous Video

Neil J. McCurdy, Jennifer N. Carlisle, William G. Griswold
Department of Computer Science and Engineering
University of California, San Diego
La Jolla, CA 92093-0114
{nemccurd,jcarlisl,wgg}@ucsd.edu

ABSTRACT

Realityflythrough is a telepresence/tele-reality system that works in the dynamic, uncalibrated environments typically associated with ubiquitous computing. By opportunistically harnessing networked mobile video cameras, it allows a user to remotely and immersively explore a physical space. Live 2d video feeds are situated in a 3d representation of the world. Rather than try to achieve photorealism at every point in space, we instead focus on providing the user with a sense of how the video streams relate to one another spatially. By providing cues in the form of dynamic transitions, we can approximate photorealistic telepresence while harnessing cameras "in the wild." This paper shows that transitions between situated 2d images are sensible and provide a compelling telepresence experience.

Author Keywords

Telepresence, Ubiquitous video

ACM Classification Keywords

H.5.1 [**Multimedia Information Systems**]: Artificial, augmented, and virtual realities

INTRODUCTION

We are rapidly moving toward a world where networked video cameras are ubiquitous. Already, camera-equipped cell phones are becoming commonplace. Imagine being able to tap into live video feeds to remotely explore the world in real time. RealityFlythrough is a telepresence system that makes this vision possible.

There are numerous applications for such a system, but perhaps the most compelling involves disaster response. Consider, for example, first responders equipped with head-mounted wireless video cameras encountering the chaos of a disaster site. As they fan out through the site, they continuously broadcast their location, orientation, and what they see to a RealityFlythrough server. The responders' central command views each of these video feeds from a first-person perspective, transitioning between them in a manner that reveals the spatial relationships between the source cameras. The resulting situational awareness helps central command direct medics to the injured, firefighters to potential flare-ups, and engineers to structural weaknesses. As more people enter the site and fixed cameras are positioned, the naturalness of the flythrough is enhanced until ultimately the entire space is covered and central command can "fly" around the site looking for hot spots without constraints.

There have been many approaches to creating interactive immersive environments that promote exploration of either a remote or a virtual space. The virtual reality community builds the environments from scratch, using photograph-based texture maps if necessary and where possible [1]; the graphics and vision communities create photorealistic renderings of novel views using photographs (and in some cases video feeds) taken from different angles [5]; and the robotics community achieves the effect by attaching a camera to a remote-controlled robot [6].

Our work starts with a different set of assumptions, and as a result leads to a very different design. The goal of RealityFlythrough is to harness networked ubiquitous cameras. Ubiquitous cameras are everywhere, or at a minimum can go anywhere. They are inside, outside, carried by people, attached to cars, on city streets, and in parks. Ubiquity moves cameras from the quiet simplicity of the laboratory to the harsh reality of the wild. The wild is dynamic—with people and objects constantly on the move, and with uncontrolled lighting conditions; it is uncalibrated—with the locations of objects and cameras imprecisely measured; and it is variable—with video stream quality, and location accuracy varying by equipment being used, and the quantity of video streams varying by location and wireless coverage. Static surveillance-style cameras may be available, but it is more likely that cameras will be carried by people. Mobile cameras that tilt and sway with their operators present their own unique challenges. Not only may the position of the camera be inaccurately measured, but sampling latency can lead to additional errors.

It is a non-trivial challenge to support live and real-time remote exploration of the world. The ideal is to have a camera lense at every possible vantage point so that a photorealistic view can be realized from anywhere. Given the pragmatic limits to ubiquity, this will not be an option in the near term. The solution, then, is to take advantage of the camera lenses that are available, and to either attempt to synthesize a novel view from the available images, or to provide a mechanism for the user's view to transition from one image to another. The synthesis of photorealistic novel views in real-time is

CHI 2005, April 2–7, 2005, Portland, Oregon, USA.
ACM 1-59593-002-7/05/0004.

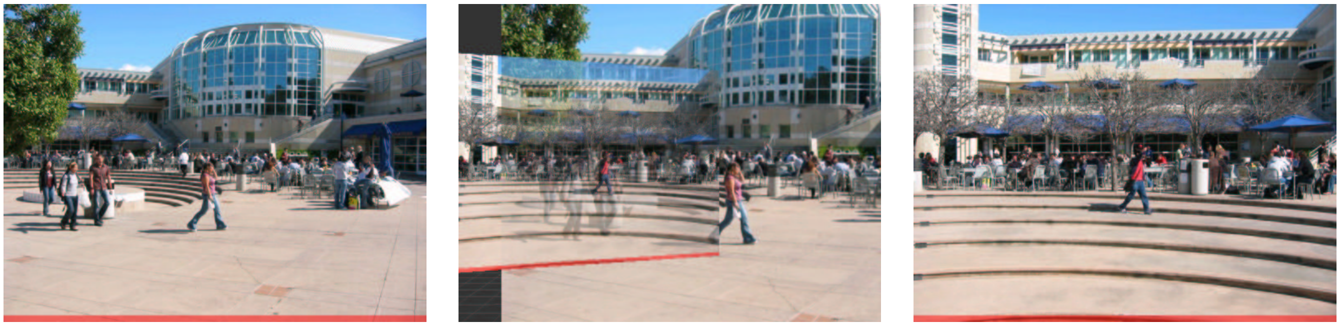
(a) (b) (c)

Figure 1. A transition from image (a) to image (c). Image (b) shows the transition in progress as image (a) moves off the screen to the right and image (c) moves in from the left. This transition represents rotating to the left while moving forward.

not possible with today's technology given the conditions of the wild, but it is possible to generate sensible transitions between camera feeds.

RealityFlythrough uses these transitions to convey spatial context. Transitions are a dynamic, real-time blend from the point of view of one camera to the point of view of another, and are designed to help the user generate an internal conceptual model of the space. Fig. 1 shows a transition in progress. Transitions provide a first-person immersion that is natural and comfortable. Other interfaces could be used to display the relationships between the cameras (a birdseye map, for example), but these have the effect of cognitively removing the user from the scene. There is an inherent tension between the uncalibrated nature of the environment and the first person immersion. Because the true relationships between images are not known, the transitions can only provide a hint of how the images are related to oneanother. This hint is enough to allow the human visual system to piece together the relationship between the images because the brain is adept at commiting closure—filling in the blanks when given incomplete information [2]. Closure is a constant in our lives; closure, for example, conceals from us the blind spots that are present in all of our eyes.

The contribution of this paper is a user study that shows that dynamic transitions between imprecisely positioned images can be comprehended. The intelligiblity of these transitions enables a telepresence system that is effective even given the conditions of the wild [4].

In the next section we will describe the user experience in more detail. This will be followed by a discussion of why transitions work. We conclude by presenting the results of an experiment that suggest that transitions help users make sense of the spatial relationships between images.

USER EXPERIENCE

A large element of the user experience in RealityFlythrough is dynamic and does not translate well to the written word or still photographs. We encourage the reader to watch the companion video submitted with this paper, or a more comprehensive video that can be downloaded from the web [3]. We do our best to convey the subtlety of the experience in this section. When observing the images in Fig. 1, keep in mind that the transformation between the images is occurring within about one second, and the one transitional frame represents only about 1/20th of the transition sequence.

The user's display is typically filled with either an image or a video stream taken directly from a camera. When a new vantage point is desired, a short transition sequence is displayed that helps the user correlate objects in the source image stream with objects in the destination image stream. These transitions are shown in a first person view and provide the user with the sensation that she is walking from one location to another. The illusion is imperfect, but the result is sensible and natural enough that it provides the necessary contextual information without requiring much conscious thought from the user.

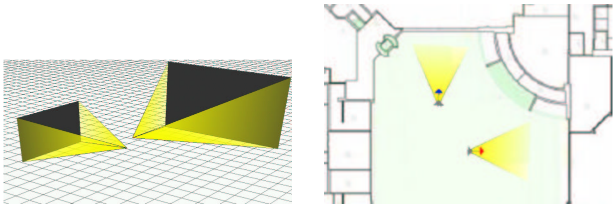
(a) (b)

Figure 2. (a) An illustration of how the virtual cameras project their images onto a wall. (b) A birdseye view that shows the position and direction of two cameras. This is an example of one of the multiple choice answers used in the experiment. For each question the subjects were presented with four images like this and had to decide which one best represents the position and orientation of the two cameras.

RealityFlythrough works by situating 2d images in 3d space. Because the position and orientation of every camera is known, a representation of the camera can be placed at the corresponding position and orientation in virtual space. The camera's image is then projected onto a virtual wall (see Fig. 2a). When the user is looking at the image of a particular camera, the user's position and direction of view in virtual space is identical to the position and direction of the camera. As a result, the entire screen is filled with the image. Referring to Fig.1, a transition between camera A (image (a) in the figure) and camera B (image (c) in the figure) is achieved by smoothly moving the user's position and view from camera A to camera B while still projecting their images in perspective onto the corresponding virtual walls. By using OpenGL's standard perspective projection matrix to render the images during the transition, the rendered view situates the images with respect to each other and the viewer's position in the environment. By the end of the transition, the user's position and direction of view are the same as camera B's, and camera B's image fills the screen. The duration of a transition depends on how far apart the cameras are, but one to two seconds is a comfortable interval and is typical in our setups.

It may be easier to understand how RealityFlythrough works by envisioning the following concrete example. Imagine standing in an empty room that has a different photograph projected onto each of its walls. Each image covers an entire wall. The four photographs are of a 360 degree landscape with one photo taken every 90 degrees. Position yourself in the center of the room looking squarely at one of the walls. As you slowly rotate to the left your gaze will shift from one wall to the other. The first image will appear to slide off to your right, and the second image will move in from the left. Distortions and object misalignment will occur at the seam between the photos, but it will be clear that a rotation to the left occurred, and the images will be similar enough that sense can be made of the transition. RealityFlythrough operates in a much more forgiving environment: the virtual walls are not necessarily at right angles, and they do not all have to be the same distance away from the viewer.

A user of RealityFlythrough can select any position within a scene as the destination of a transition. Since the transitions described so far only move between two cameras, there are many cases where large gaps will appear between the images. Consider, for example, a 180 degree rotation where the majority of the transition will consist of a gap. We fill these gaps by displaying images from other cameras that cover the intervening space. Conceptually, a transition is divided into sub-transitions; each of which is a transition between two cameras. In a real environment the camera density will probably not be high enough for all of these gaps to be filled. To handle this situation, we capture still images from the live video feeds and use these to provide additional imagery. An age indicator bar reveals the age of the image to the user. When no images are available, a floor grid is used instead.

WHY IT WORKS

RealityFlythrough works in the wild because no pre-processing of the imagery is required. All of the information necessary to do a transition can be obtained in real-time. The position of the camera can be retrieved from whatever locationing technology is desired (we use WAAS-enabled consumer GPS's for outdoor positioning), and the tilt, roll, and yaw of the camera can be obtained from a tilt sensor (we use the EZ-Compass-3 produced by AOSI). Since no pre-processing is required, the system works equally well with static images and live video streams. This makes real-time exploration of a live scene possible.

Since the camera optics are not calibrated and the locations of cameras cannot be precisely measured, the transitions between camera views are not perfect. Ghosting, tears, and object misalignment is common. It is important to reveal these defects rather than conceal them with blurring because their presence helps the user make sense of the transition. A double image of a tree during a transition, for example, reveals to the user how the tree has moved between images. This knowledge helps the user understand the relative locations of the view points. We have found that an alpha blend between the overlapping portions of the images creates a transition that is pleasing while being sufficiently revealing.

RealityFlythrough is successful as a real-time live telepresence system because it taps into the power of the human visual system, offloading much of the processing requirements to a tool that is very adept at drawing conclusions from scanty information. The key design decision is determining how to divide the labor between the computer and the brain so that both components are being used to their full potential. How much computer processing is possible given the real-time constraints, and does the resulting visualization provide enough clues to the brain to allow it to draw the correct conclusions? The human visual system provides opportunities, but it also imposes constraints. The visual system evolved in a three dimensional immersive environment, and it uses cues such as stereo perception and parallax to determine depth. This suggests that if we imitate the same 3d environment, the visual system has the best chance to process and comprehend the scene in a natural way. Information about depth is missing in our environment so we provide the vision system with other clues in the context of a 3d environment, and let the vision system make inferences about depth and relative position. These clues come in the form of motion, zooming, and object overlap in the images. Many depth cues are already apparent in the images; what the brain needs to do is resolve the ambiguous depth cues between the two images given the motion suggested by the transition.

EXPERIMENT

To ascertain how well our transitions convey additional information about the spatial relationships between cameras, we constructed an experiment that compared simple two-camera transitions with a no-transition alternative. We will call these two scenarios *transition* and *no-transition* respectively. We assumed that the ideal would be perfect, seamless transitions that could convey spatial relationships with 100% accuracy. Our target was 100% accuracy.

For the *transition* tests, a short video was played that showed a transition between two still photographs. The subject could watch the transition multiple times and could control the playback speed. While watching the video the subjects had to choose the best of four possible birdseye depictions of the scene that showed the relative positions of the cameras (Fig. 2b). The *no-transition* tests were similar, only instead of a video the subjects viewed two photographs while making the selection. The transition represented in Fig 1 is an example of a transition that might have been shown in a *transition* test, and Figs. 1a and 1c are examples of still photos that might have been used in a *no-transition* test.

We had 30 subjects participate in the study. The majority were university students, but their experience with computers varied. Each subject was tested on both *transition* and *no-transition* questions. The scenes depicted in the photographs fell into two categories: *familiar* and *unfamilar*. The *familiar* location was a campus foodcourt that all participants were very familiar with. The *unfamilar* location was a disaster scene that no one was familar with and was difficult to interpret even when familiar with it. Twenty questions were asked of each participant—five in each category. We attempted to make the questions increase in difficulty based on our experience with which motions are difficult to visualize. Rotations were considered simple. Rotations combined with motion were considered more difficult. Questions were randomly interleaved from the four categories, but each participant was asked the questions in the same order.

We hypothesized that the *transition* responses would be quicker and more accurate. Given the difficulty we had with determining the locations of the cameras at the *unfamiliar* location, we also hypothesized that the *no-transition* answers would do no better than random guessing, and the *transition* answers would do much better.

	No-Transitions	**Transitions**
Unfamiliar Location	1.63/5 (σ = 1.10)	2.60/5 (σ = 1.19)
Familiar Location	2.73/5 (σ = 1.01)	4.10/5 (σ = 0.87)

Table 1. Mean of correct responses for all 30 participants.

Results

As Table 1 shows, the mean scores for the *transition* questions exceeded those of the *no-transitions* questions. Furthermore, of the 30 subjects 26/30, 86.67% achieved a greater or equal score on the *transitions* questions. This indicates that the transitions provide the user with additional information that is beneficial in determining the spatial relationship between cameras.

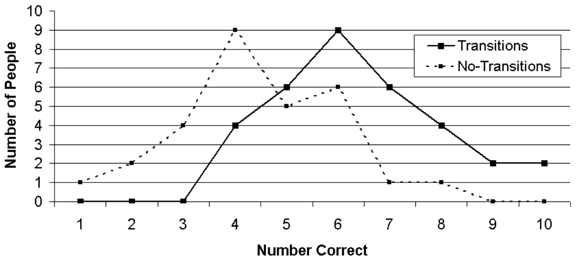

Figure 3. Number of people who answered x number correct. Compares *transitions* to *no-transitions*.

Fig. 3 demonstrates that subjects attained a higher level of success answering transition questions. Scores in the 50th percentile and higher demonstrated a greater rate of success for the transition questions.

When drilling into the data we determined that the success rate on the *transition* questions increased as the experiment progressed. This suggests that as the users become more familiar with transitions, the transitions become easier to interpret. Most notably, the second to last and the last *transition* questions at the *familiar* location were answered correctly 93.33% and 100% of the time respectively. We also looked at the relative increase in speed between the *transition* and *no-transition* questions and noted general patterns that indicated that the *transition* questions were answered more quickly as transition interpretation was learned. In fact, by the end of the experiment all *transition* questions were answered faster than the *no-transition* questions. These results are supported by a comment from one of our subjects in a post-experiment interview: "[transitions are] different than anything I had really seen. At first it seemed very strange and took me by surprise. By the middle or end of the test I had really gotten the hang of it and the transition questions seemed much easier."

Our hypothesis that *no-transition* questions in the *unfamiliar* location would be answered randomly was supported by the data and by user comments. Random guessing would produce an average score of 1.25 out of 5; the 1.67/5 average score obtained in the study is not much better than random. The average *transition* score of 2.73/5 is better, but not quite as good as we had hoped. The subjects lack of experience with transitions may explain this. The transitions were much more difficult at this location because the images and the subject's knowledge of the space provided little additional help. Complete trust had to be placed in the transition, and some subjects were not ready to extend that trust. Referring to Fig. 3, notice that two subjects scored perfectly on the transition questions, and two scored in the 90th percentile. All four of these individuals reported having a great deal of experience playing 1st person shooter games, suggesting that cognition of image-to-image transitions is a skill that is honed through exposure. The subjects who were not able to trust the transitions reported no such experience.

CONCLUSION

We have presented a novel interface for navigating between ubiquitous video feeds. We have shown experimental results that suggest that this interface provides additional information about the spatial relationships between the source cameras, and that comprehension of the interface increases with use. Comprehension for expert users was shown to approach the 100% that would be expected had perfect transitions between images been possible.

REFERENCES

1. Leigh, J., Johnson, A. E., DeFanti, T. A., and Brown, M. D. A review of tele-immersive applications in the CAVE research network. *VR*. 180–.
2. McCloud, S. *Understanding comics: The invisble art.* Harper Collins Publishers, New York, 1993.
3. McCurdy, N. J., and Griswold, W. G. Tele-reality in the wild. UBICOMP'04 Adjunct Proceedings, 2004. `http://activecampus2.ucsd.edu/˜nemccurd/tele_reality_wild_video.wmv`.
4. McCurdy, N. J., and Griswold, W. G. A systems architecture for ubiquitous video. Tech. Rep. CS2005-0813, University of California, San Diego, 2005.
5. Neumann, U., You, S., Hu, J., Jiang, B., and Lee, J. Augmented virtual environments (ave): Dynamic fusion of imagery and 3d models. *VR '03: Proceedings of the IEEE Virtual Reality 2003*. IEEE Computer Society (2003), 61.
6. Paulos, E., and Canny, J. Ubiquitous tele-embodiment: Applications and implications. *International Journal of Human-Computer Studies/Knowledge Acquisition*.