Beating the Limitations of Camera-Monitor Mediated Telepresence With Extra Eyes

author1[1] author2[2]

MIT[1]

CMU[2]

November 12, 2013

Abstract

- In physical presence, you are most aware of your immediate surroundings, such as what is at your feet or who is beside you, and less aware of objects further away.
- In telepresence, almost the opposite is true.
- Due to the nature of the medium, you are most aware of what is in front, often at a distance, as dictated by the limited view of the camera.

Abstract

- Even where remote camera control is possible, the range of exploration is limited and the logistics of control are typically awkward and slow.
- All of this adds up to a pronounced loss of awareness of the periphery in telepresence.
- The research described here attempts to compensate for these problems through two mechanisms.

Abstract

- First, we provide telepresence users with two separate views, one wide-angle and the other, a controllable, detailed view.
- To simplify navigation, the two views are seamlessly linked together, so that selecting a region of one will have an effect in the other.
- Second, we utilize sensor information from the remote location to provide the user with notification of relevant events that may require attention.

Introduction

- Normal human vision can be conceived as consisting of two highly mobile cones of view.
- One is the focused foveal cone, one degree wide, while the second is the peripheral cone, or global field of view, spanning approximately 170 degrees .
- Excellent spatial resolution is provided by the first, while the second, lower resolution view, provides us with stimulus that acts to redirect our attention.

Introduction

- Camera-monitor mediated vision, in contrast, suffers in resolution and due to the size of the display, uses limited azimuth of the visual field.
- Watching television, for instance, typically involves the foveal cone only.

Introduction

- The narrow channel of information, both in the sense of bandwidth and field of view, imposes limitations on the ability to explore, follow conversations, check reactions, and generally sense significant actions in a remote space, such as people passing by or entering.
- In such situations, users must choose between a global and a focused view.

Introduction

- With the former, resolution is sacrificed to permit a wide field of view and easy change of gaze direction.
- If only the focused view is provided, users obtain details but no peripheral awareness.
- This is typical of most videoconference settings .
- One approach to support both the foveal and peripheral cones is with multiple views.
- The problems with this approach are well understood.

Introduction

- The Multiple Target Video (MTV) system of Gaver et. al. first proposed the use of multiple cameras as a means of providing more flexible access to remote working environments.
- Users were offered sequential access to several different views of a remote space.
- However, as the authors noted, a static configuration of cameras will never be suitable for all tasks.

Introduction

- Furthermore, switching between views introduces confusing spatial discontinuities.
- A further study (MTV II) by Heath et. al. attempted to address this latter issue by providing several monitors, so that every camera view was simultaneously available.
- While this new configuration was more flexible, the inability of static cameras to provide complete access to a remote space still remained a problem.

Introduction

- Furthermore, the various views were independent of one another, and the relationship between them was not made explicit.
- Consequently, spatial discontinuities persisted.
- Another approach involved the Virtual Window concept , which uses the video image of a person's head to navigate a motorized camera in a remote location.

Introduction

- Our user experience with this technique revealed a significant improvement to the user's sense of engagement in meetings.
- Unfortunately, when the camera was focused on a small area, the loss of global context often made the user unaware of important activity taking place out of view.
- To compensate for the limitations on vision imposed by camera-monitor mediated telepresence, the work discussed here offers to:.

Introduction

- 1. Provide both a global (peripheral) and a detail (focused) view, simultaneously.
- We note that this approach has already been used extensively in the Ontario Telepresence Project by combining the two views through a picture-in-picture device.
- The same approach with multiple views was also proposed by Kuzuoka et. al. .

Introduction

- However, as will be discussed later, providing a link between the two views is not only critical for usability, but also supports the goal of multiple views while avoiding the pitfalls of spatial discontinuities inherent in the MTV studies .
- 2. Provide a navigation mechanism using these views, allowing users to redirect their view in both direction and scale, through a simple user interface.

Supporting Foveal and Peripheral Cones

- It has been suggested by several vision researchers that a brain mechanism exists to drive foveating saccades of the eye in response to stimulus in the periphery region .
- In the discussion of their model of saccadic eye movement, Tsotsos et al. comment that these saccades play an important role in the exploration of the visual world .
- Supporting evidence for this comes from neurophysiology.

Supporting Foveal and Peripheral Cones

- A region known as PO, which receives a representation of the periphery of the visual field, has been identified in the brains of primates .
- Deprived of this information, individuals suffering from tunnel vision, or a loss of vision outside the fovea, exhibit severe problems navigating through their physical surroundings, even when these surroundings are familiar to them .

Overlaid Multiple Views

- As an initial attempt to provide this support, we developed a prototype system, consisting of a large and small display, as shown in Figure 1.
- The large screen display provides the user with a wide angle view of the remote space while the small display provides a high resolution view of the area of interest.
- With the camera orientations fixed and the proper geometric positioning of the two displays, spatial discontinuities are minimized.

Overlaid Multiple Views

- The sensation of increased peripheral awareness obtained by this system is very powerful.
- We note that this prototype requires two high-resolution displays, one of them quite large, in order to achieve a significant effect.
- As this may be prohibitively expensive for most videoconference users, we would like to unify the two views into a single display.

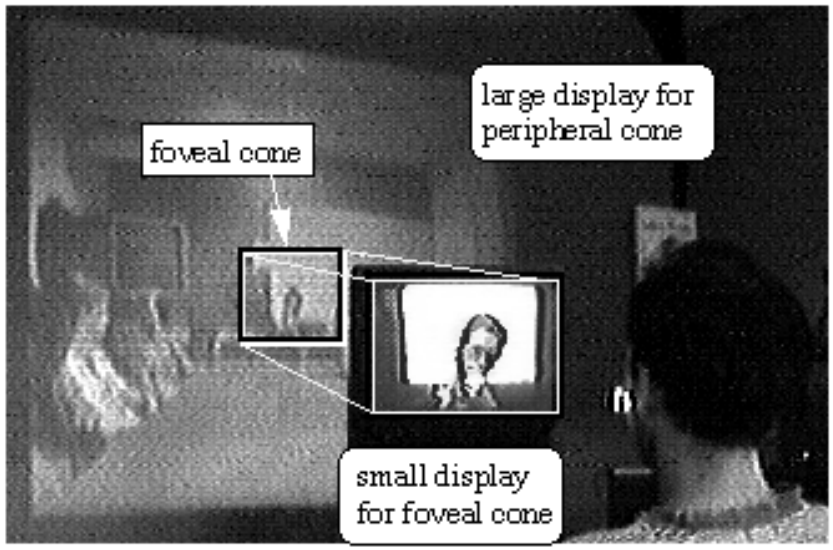


Figure 1. This prototype system uses a large screen display for the peripheral view and a small screen for the detail view.

Disjointed Multiple Views

- Another approach to supporting both the foveal and peripheral views is to display the two separately on the same screen.
- Since the views are disjointed, each can have sufficient size and resolution, even with the limitations of current technology.
- Our implementation of this system is shown in Figure 2.

Disjointed Multiple Views

- The top portion of the display provides a foveal or detail view, obtained from a user-controlled motorized camera, while the lower portion provides the peripheral or global view from a fixed, wide-angle camera.
- Since the views are independent of each another, there is no consistent geometric relationship between the two.

Disjointed Multiple Views

- This can result in an inability to locate the position of the detailed region within the peripheral view, once more bringing us back to the problem of spatial discontinuities.
- Navigation under these conditions is typically difficult and slow.
- This is especially severe when the scene being viewed is relatively homogeneous (e.g. through tele-education, a large class of students).

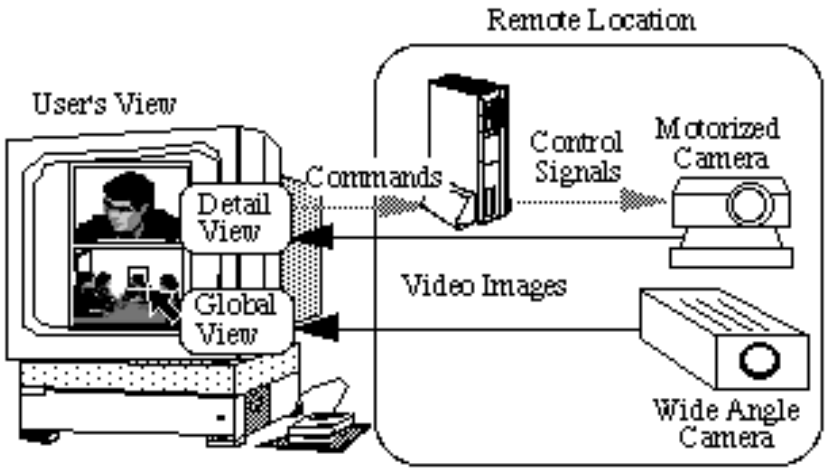


Figure 2. Architecture of the Extra Eyes system.

Linking Multiple Views

- To address the lack of a geometric relationship between the two views, we indicate the detailed region within the global view by means of a yellow bounding box (detail frame), as shown in Figure 3.
- The enclosed region corresponds exactly to what is displayed in the detail view.
- As the detail view changes, the bounding box on the global view adjusts accordingly.

Linking Multiple Views

- Because the two views are logically linked, users can select a desired region by sweeping out a bounding box or simply point-and-click on the global view.
- In the former case, the detail view is defined by the size of the bounding box, while in the latter, the detail view is centered at the selected position and displayed at the maximum zoom.

Linking Multiple Views

- These interaction techniques with the global view permit a far more efficient navigation mechanism than the effectively blind view selection offered by both the original MTV system and the Virtual Window system .
- In addition to control via the global view, the detail view can be manipulated directly through the scroll bars, which provide tilt and pan control of the motorized camera.

Linking Multiple Views

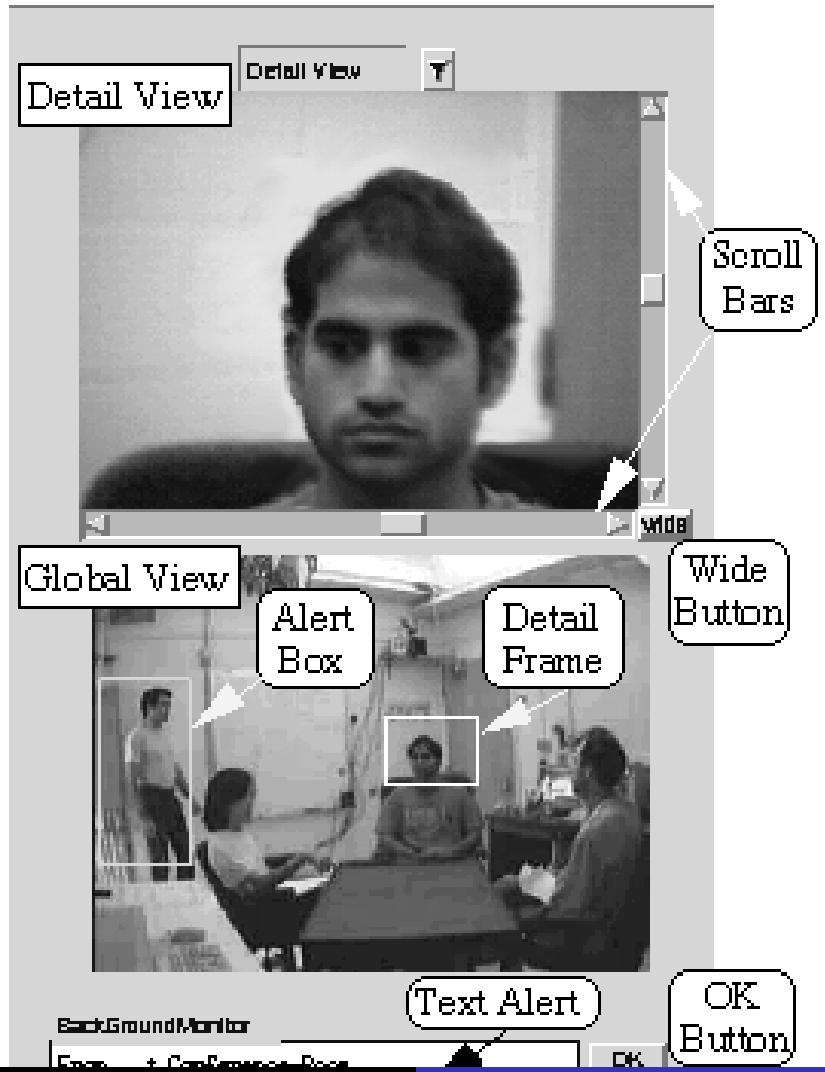
- It is also possible to adjust the zoom factor of the detail view by pressing the left or right mouse button, or obtain a wide view by selecting the wide button.
- To provide a linkage between the global and detail views, we require a mapping between the coordinate systems of each, dependent on the properties of the different cameras.
- We first define a global coordinate system, which covers the entire area visible to both cameras.

Linking Multiple Views

- Next, we define models for each camera, which consist of a view model, and in the case of the motorized camera, a transformation function.
- The models describe the relationship between pixel coordinates of each camera and the global coordinate system.
- In the case of our fixed wide-angle camera, this is simply a one-to-one mapping.
- The transformation function for the motorized camera maps pixel coordinates to the appropriate motor signals.

Linking Multiple Views

- The models and relationships are described in Figure 4.
- When a user selects an area of the global view, the pixel coordinates of this region are first translated into global coordinates through the wide angle view model, and then into pixel coordinates of the detail view.
- The detail pixel coordinates are then mapped into motor signals via the transformation function.
- Finally, the motor signals are sent to the detail camera.


Detail View
Detail View
Scroll Bars
Wide
Global View
Alert Box
Detail Frame
Wide Button
Text Alert
OK Button
BackGroundMonitor
OK

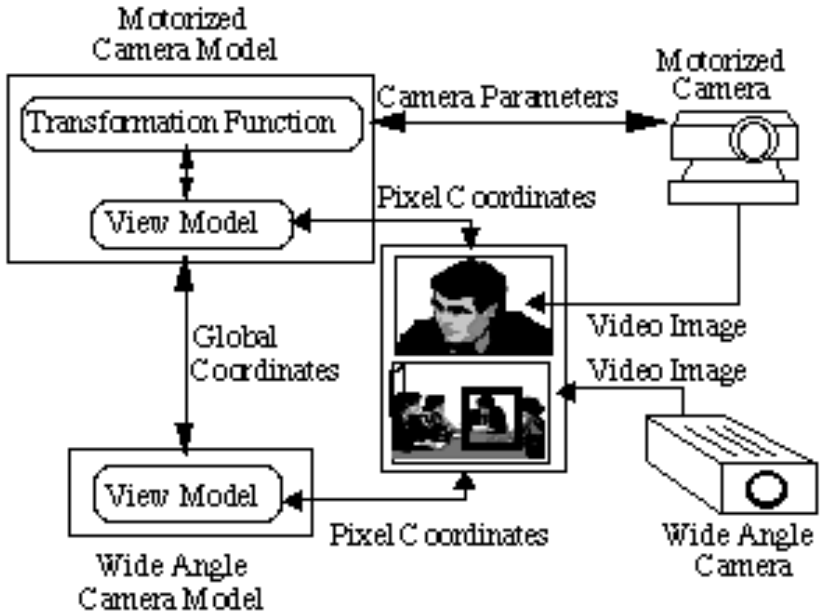


Figure 4. Camera models and their relationships.

Sensory Surrogate for Environmental Awareness

- There exists no substitute for physical presence that offers the fidelity of rapidly directable stereo vision and spatially sensitive binaural audio, as manifested by the human senses.
- To help bridge the gap between physical presence and telepresence in this regard, our Extra Eyes system provides users with a sensory surrogate to increase their awareness of the remote environment.

Sensory Surrogate for Environmental Awareness

- The surrogate monitors background information obtained by sensors and reports on relevant events through the use of sound, text, and graphics, or a combination of the three.
- In this manner, background processing by the computer is used to improve the user's foreground awareness.

Sensory Surrogate for Environmental Awareness

- Sensors in the room monitor the status of presentation technology such as the VCR, document camera, and digital whiteboard, as well as the entry of new individuals as depicted in Figure 5.
- When an event occurs, it triggers an alert-action sequence.

Sensory Surrogate for Environmental Awareness

- The alert corresponds to the screen message displayed (e.g. "Someone has entered the room. Do you wish to view the doorway?"), as well as the appearance of a blue bounding box (alert box) in the corresponding region of the global view, as shown in Figure 3.

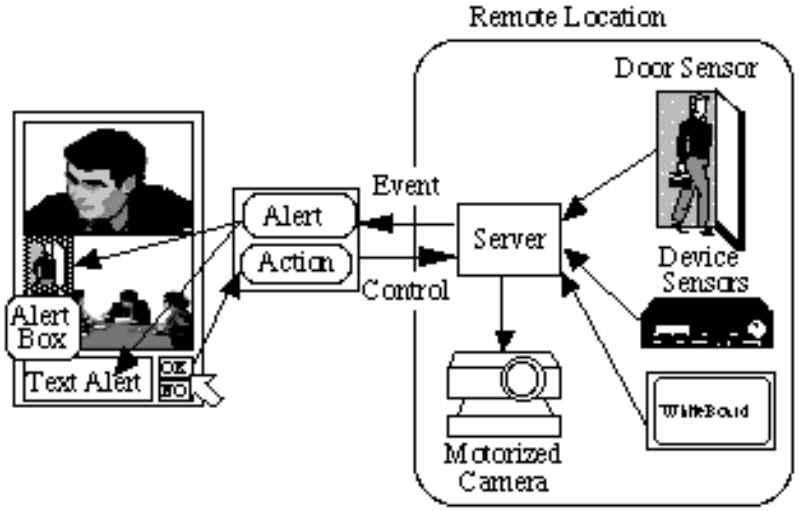
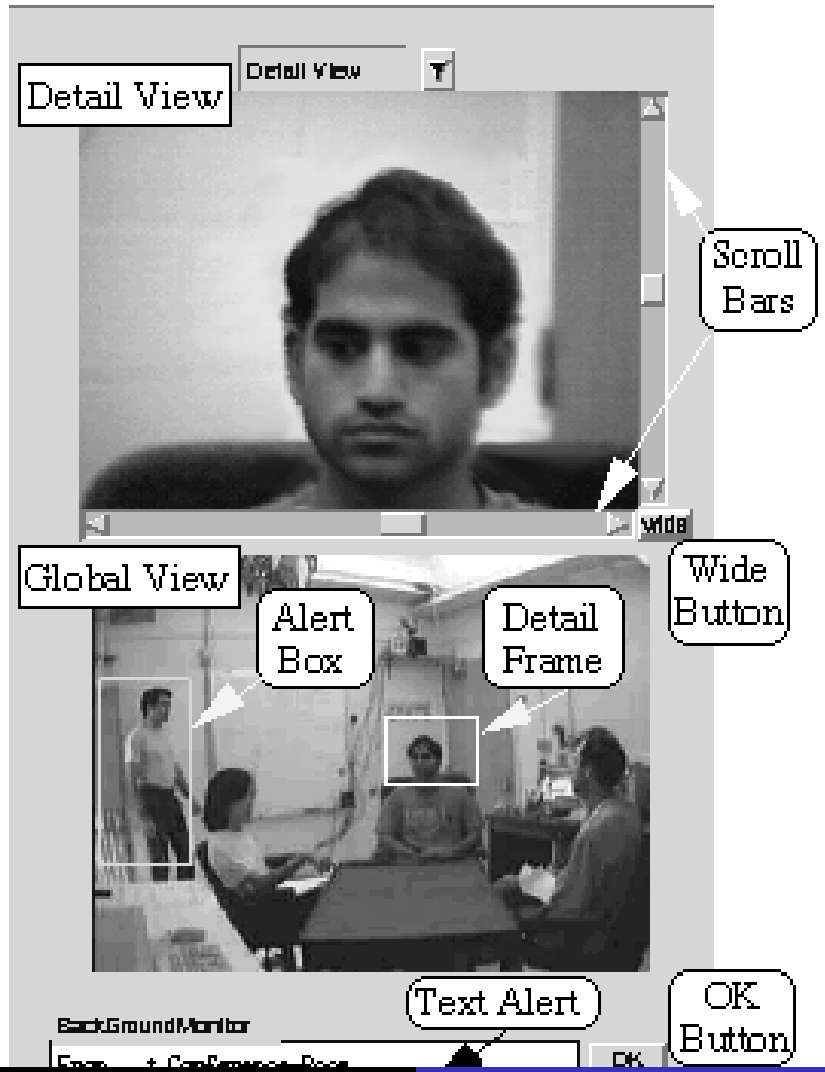


Figure 5. The sensory surrogate in operation.


Detail View
Detail View
Scroll Bars
Wide
Global View
Alert Box
Detail Frame
Wide Button
Text Alert
OK Button
BackGroundMonitor
OK

Sensory Surrogate for Social Awareness

- We have also applied the sensory surrogate concept to increasing social awareness among individuals sharing the media space of the Ontario Telepresence Project .
- The Postcards system (see Figure 6), based on Rank Xerox EuroPARC's Portholes , captures snapshots from each user's office at set intervals and distributes these to members of the media space.

Sensory Surrogate for Social Awareness

- A sensory surrogate in the Postcards system compares every two consecutive frames from each office to determine if there is activity there.
- This is done by counting the number of pixels that have changed by more than a certain threshold amount between the two frames.
- Although the algorithm is susceptible to false detection of activity due to camera perturbations, it has worked reasonably well in our environment.

Sensory Surrogate for Social Awareness

- Stored knowledge of activity allows Postcards to determine whether individuals are in or out, or have recently entered or vacated their offices.
- Users can take advantage of this background monitoring feature by asking the system to sense activity and notify them when any number of individuals are simultaneously present in their offices.

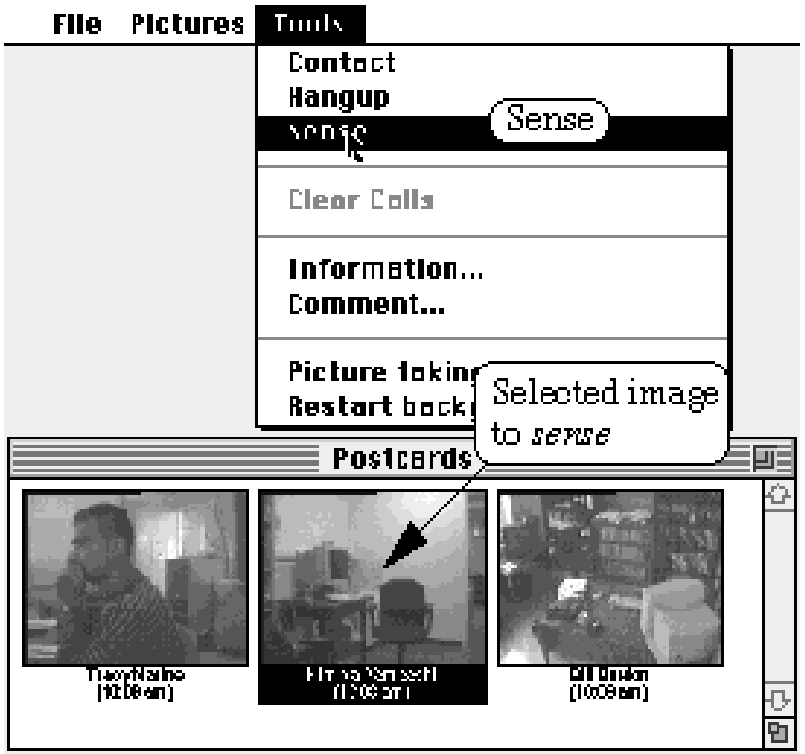


Figure 6. Screen layout of Postcards. Images from each room are captured periodically.

User Study

- We evaluated the performance of Extra Eyes through the following user study.
- Three television monitors were arranged in a remote location, as shown in Figure 7.
- Letters of the alphabet were displayed on a randomly chosen monitor, one at a time.
- The user's task was to use the Extra Eyes system to identify these letters as they appeared, as quickly as possible, while minimizing the number of errors.

User Study

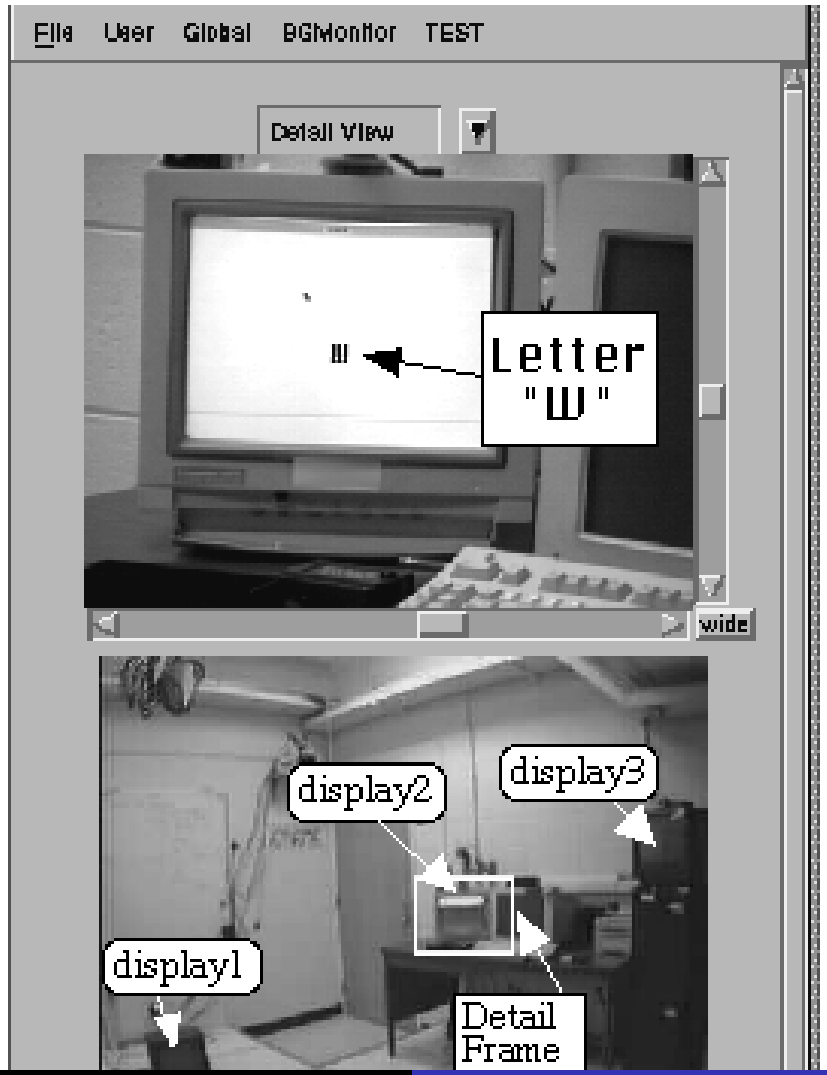
- Each letter would remain on the monitor until the user had identified it, by typing its corresponding key.
- Once the letter was identified, it would be replaced by another letter on a different monitor.
- The font size was sufficiently small so that a zoom factor near the maximum was required for legibility.
- We tested each of our seven subjects on the following conditions, the order being randomly varied, with 20 repetitions per condition:.

User Study

- 1. No Global: Only the detail view is visible.
- This situation is equivalent to typical telepresence systems.
- 2. No Global + Text: Same as 1.
- In addition, a text alert indicates the display on which the current letter appears.
- 3. Unlinked: Both the global and detail views are simultaneously visible, but the two views are not linked (i.e. neither view has effect on the other).
- This is equivalent to the MTV system.

User Study

- 4. Linked: Both the global and detail views are simultaneously visible and linked.
- 5. Linked + Text: Same as 4.
- In addition, a text alert indicates the display on which the current letter appears.
- 6. Linked + Action: Same as 5.
- In addition, an alert box appears, and the user can invoke the action corresponding to the alert by pushing the OK button or by clicking anywhere within the alert box.



Discussion and Results of User Study

- For the first three conditions, users exhibited two strategies to identify the various letters.
- When no information beyond that of the detail view was available, users consistently zoomed out to obtain a wide angle view, then panned and tilted the camera to center the letter, before zooming in again.

Discussion and Results of User Study

- This zoom-out strategy, represented by the solid line in the space-scale diagram of Figure 8a, requires over three camera operations, on average, to identify each letter.
- When an alert message was added, indicating the display on which the new letter appears, users tended to change their strategy.

Discussion and Results of User Study

- Knowing the approximate location of the desired monitor from past experience gathered during the study, users often tried to find this monitor by repeatedly panning and tilting the camera, as shown by the solid line in Figure 8b.
- This strategy is quite similar to searching for an object in a familiar room, while in the dark.

Discussion and Results of User Study

- Because users cannot accurately select a desired position with the pan-tilt strategy, this method often requires more operations than the zoom-out strategy.
- The same pan-tilt strategy was used when the global view was provided, but not linked to the detail view.
- For the remaining three conditions, users were able to identify the letters with only a single camera operation.

Discussion and Results of User Study

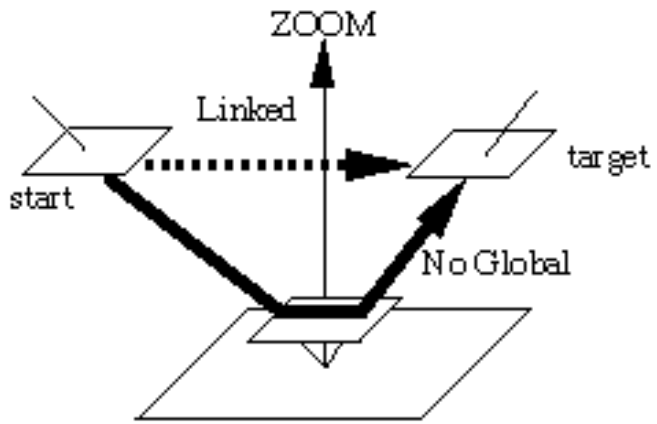
- Figure 9 and Figure 10 present the results of our user study, indicating the average number of camera operations users required to identify each letter, as well as the average completion time with 95% confidence error bars, with each of the six experimental conditions.
- Analysis of variance (ANOVA) showed that both number of operations and trial completion times were significantly affected by the experimental conditions.

Discussion and Results of User Study

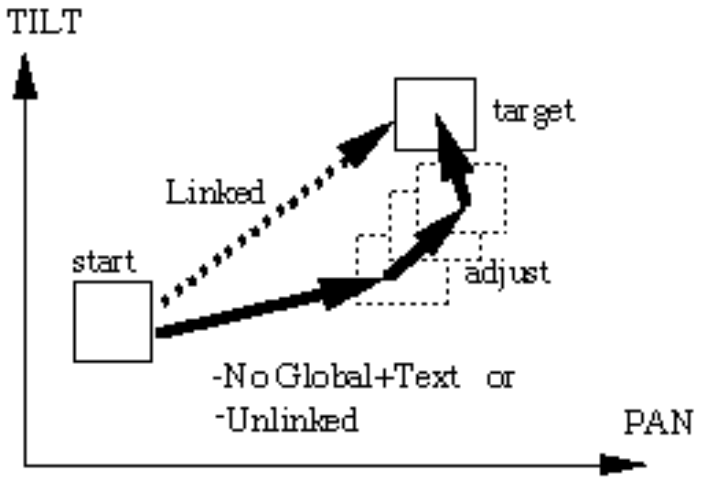
- For number of operations, F(5, 30)=55.2, p¡0.001.
- For completion time, F(5, 30)=40.1, p¡0.001.
- As measured by number of operations (Table A1 in the Appendix), Fisher's protected LSD posthoc analyses showed that all linked conditions were significantly different from the Unlinked and NoGlobal conditions (p¡0.05).
- However, there is no significant difference among linked conditions.

Discussion and Results of User Study

- The difference between Unlinked and NoGlobal, as well as Unlinked and NoGlobal+Text is also insignificant.
- As measured by completion times (Table A2 in the Appendix), Fisher's protected LSD posthoc analyses showed that all conditions were significantly different from each other ($p¡0.05$), except Linked+Action vs. Linked+Text condition ($p=0.64$) and NoGlobal vs. Unlinked ($p=0.66$).

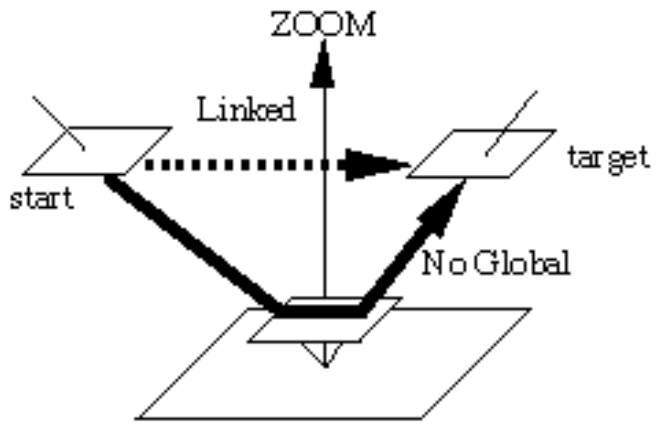


(a) ZOOM-OUT strategy

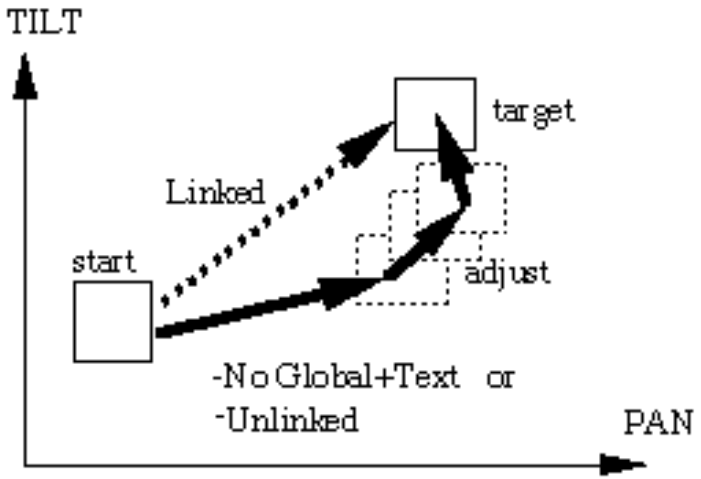


(b) PAN-TILT strategy

Figure 8. Space scale diagram of camera movement



(a) ZOOM-OUT strategy



(b) PAN-TILT strategy

Figure 8. Space scale diagram of camera movement

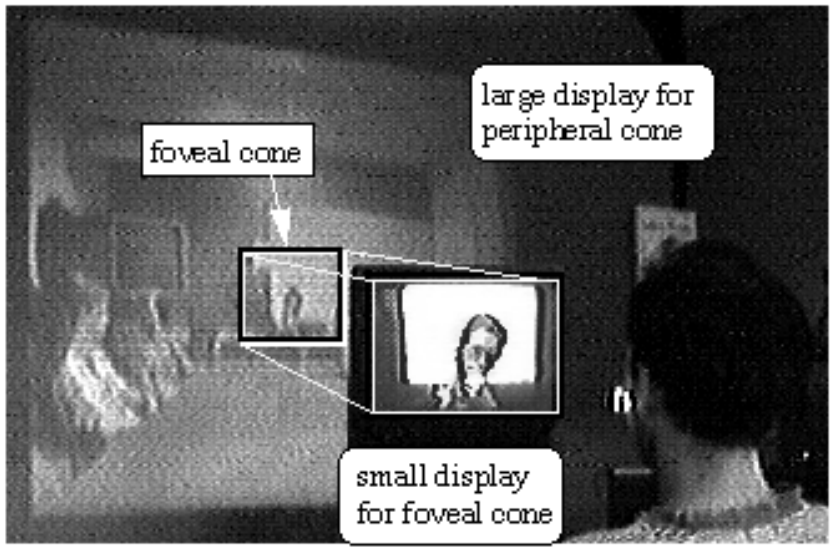


Figure 1. This prototype system uses a large screen display for the peripheral view and a small screen for the detail view.

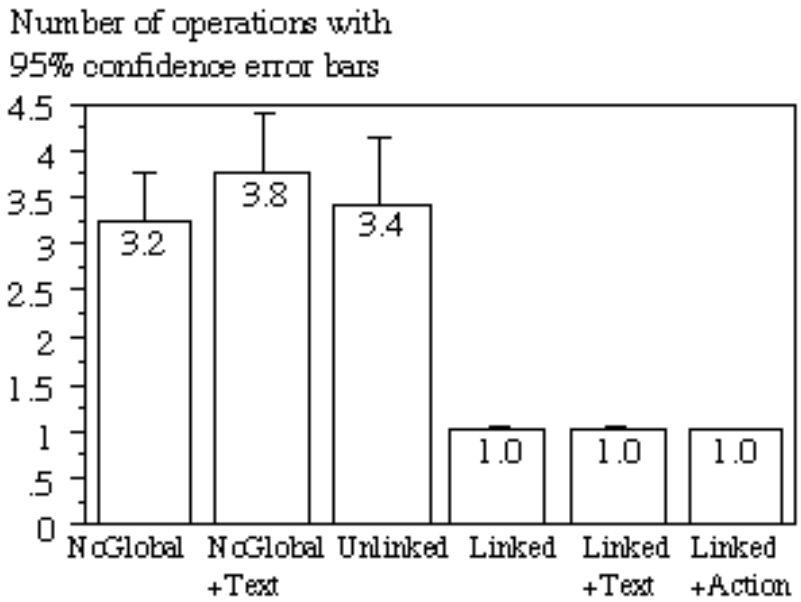


Figure 9. Means of number of operations in each experimental condition.

Figure 9 and Figure 10 present the results of our user study, indicating the average number of camera operations users required to identify each letter, as well as the average completion time with 95% confidence error bars, with each of the six experimental conditions.

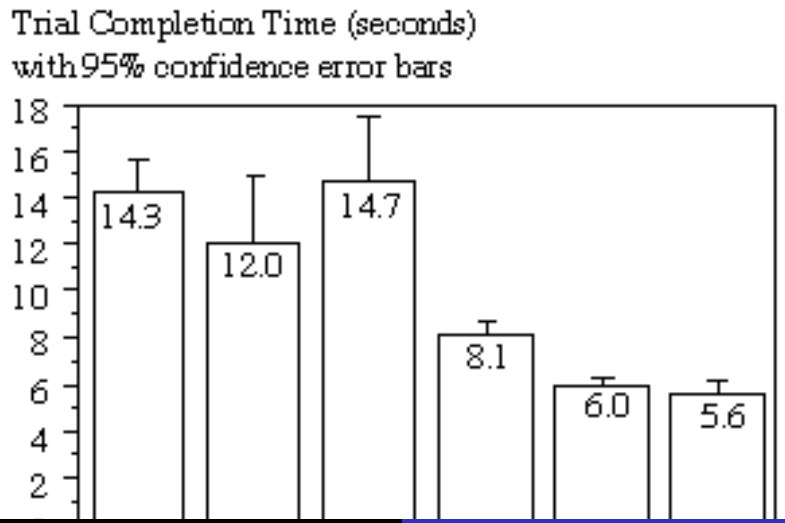


Table A1. Posthoc analysis of six experimental conditions: number of operations.

	Vs.	Diff	Crit. diff	P-Value	
Linked +Action	Linked+Text	.008	.528	.9757	
	Linked	.008	.528	.9758	
	NoGlobal+Text	2.753	.528	.0001	S
	NoGlobal	2.226	.528	.0001	S
	Unlinked	2.416	.528	.0001	S
Linked +Text	Linked	4.2E-5	.528	.9999	
	NoGlobal+Text	2.746	.528	.0001	S
	NoGlobal	2.218	.528	.0001	S
	Unlinked	2.408	.528	.0001	S
Linked	NoGlobal+Text	2.746	.528	.0001	S
	NoGlobal	2.218	.528	.0001	S
	Unlinked	2.409	.528	.0001	S
NoGlobal +Text	NoGlobal	.528	.528	.0500	S
	Unlinked	.337	.528	.2022	
NoGlobal	Unlinked	.191	.528	.4660	

S = Significantly different at the level:0.05.

Table A2. Posthoc analysis of six experimental conditions: time (seconds).

	Vs.	Diff	Crit. diff	P-Value	
Linked +Action	Linked+Text	.440	1.872	.6350	
	Linked	2.554	1.872	.0092	S
	NoGlobal+Text	6.423	1.872	.0001	S
	NoGlobal	8.762	1.872	.0001	S
	Unlinked	9.172	1.872	.0001	S
Linked +Text	Linked	2.115	1.872	.0282	S
	NoGlobal+Text	5.983	1.872	.0001	S
	NoGlobal	8.323	1.872	.0001	S
	Unlinked	8.732	1.872	.0001	S
Linked	NoGlobal+Text	3.868	1.872	.0002	S
	NoGlobal	6.208	1.872	.0001	S
	Unlinked	6.617	1.872	.0001	S
NoGlobal +Text	NoGlobal	2.340	1.872	.0160	S
	Unlinked	2.749	1.872	.0054	S
NoGlobal	Unlinked	.409	1.872	.6585	

S = Significantly different at the level:0.05.

Conclusion 1: Linkage Between Views is Very Important

- When the two views were linked, navigation in the remote environment via selection in the global view was effortless.
- Any desired (visible) target could be selected directly with a single camera operation, as indicated by the dashed lines of Figures 8a and 8b (see also Figure 9).
- In this case, the previous indirect strategies of zoom-out and pan-tilt, which require almost twice as much time as direct selection, were never used.

Conclusion 1: Linkage Between Views is Very Important

- Users expressed their opinion that the direct selection mechanism was more natural than the indirect methods.
- Indeed, all linked conditions were significantly better than the unlinked one in terms of both number of operations and trial completion time.
- Further user feedback was also highly informative.
- Some commented that the detail frame was useful as an indication of direction of camera motion.

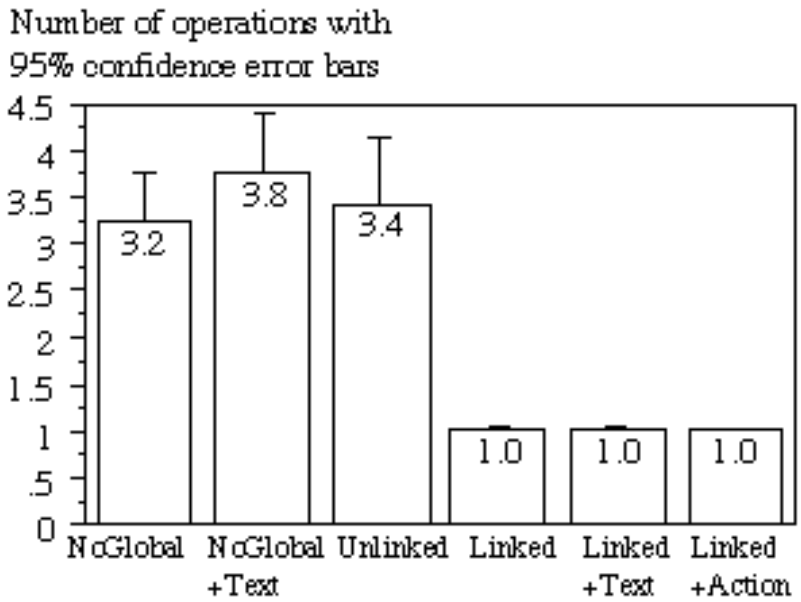


Figure 9. Means of number of operations in each experimental condition.

Conclusion 2: Sensory Information is Useful

- The time improvement from linked views to linked views with a text alert (p¡0.05, see Table A2) indicates the added value of sensory information.
- As most users explained, the alert allowed them to reduce the size of the visual search area.
- Users also appreciated the audio feedback of a beep, provided simultaneously with an alert message, indicating that a new letter was about to appear.

Conclusion 2: Sensory Information is Useful

- We note that sensory information may have compensated for the low update rate (approximately 1-2 frames/s in our present implementation) of the global view.
- In many instances, the indication of various alerts preceded the appearance of a new letter on the global view by one second or more.
- This enabled users to begin their navigation toward the desired monitor before the letter was actually visible.

Conclusion 2: Sensory Information is Useful

- Although the differences in time and number of operations between Linked+Text and Linked+Action were not statistically significant, users indicated that the graphic alerts were more useful than text messages.
- The graphic alerts completely specify the relevant visual regions, as opposed to text alerts, which require the user to read and then perform a search.

Table A2. Posthoc analysis of six experimental conditions: time (seconds).

	Vs.	Diff	Crit. diff	P-Value	
Linked +Action	Linked+Text	.440	1.872	.6350	
	Linked	2.554	1.872	.0092	S
	NoGlobal+Text	6.423	1.872	.0001	S
	NoGlobal	8.762	1.872	.0001	S
	Unlinked	9.172	1.872	.0001	S
Linked +Text	Linked	2.115	1.872	.0282	S
	NoGlobal+Text	5.983	1.872	.0001	S
	NoGlobal	8.323	1.872	.0001	S
	Unlinked	8.732	1.872	.0001	S
Linked	NoGlobal+Text	3.868	1.872	.0002	S
	NoGlobal	6.208	1.872	.0001	S
	Unlinked	6.617	1.872	.0001	S
NoGlobal +Text	NoGlobal	2.340	1.872	.0160	S
	Unlinked	2.749	1.872	.0054	S
NoGlobal	Unlinked	.409	1.872	.6585	

S = Significantly different at the level:0.05.

Further Issues

- Having described Extra Eyes and our preliminary evaluation of this system, we now turn to some other issues.
- The global view provided by our present system can not capture a view of the entire room.
- Other designers may prefer to use multiple cameras, or a very wide angle lens, possibly a fisheye, for this task.

Further Issues

- In the former case, some form of image processing will be required to combine the images, while in the latter, unwarping to compensate for image distortion will be necessary.
- Detractors may argue that transmitting video for the global view is too expensive.
- Either more bandwidth is required, or the frame rate of the detail view will suffer.

Further Issues

- We suggest that since the global view is only required to provide a sense of peripheral awareness, both its frame rate and resolution can be relatively low.
- In fact, we reduced our global view to a quarter size (160 x 120 pixels), and found that users were still very aware of activities occurring in the periphery.

Future Work

- While the sense of peripheral awareness offered by a fixed global view is a helpful navigation tool, it does not accurately replicate the mechanics of human vision, in which the periphery is dictated by the orientation of the fovea.

Future Work

- A future version of Extra Eyes should remedy this shortcoming, either by attaching the global camera to the motorized detail camera, or by using another motorized camera for the global view, synchronized with the detail camera.
- This improvement is presently being applied to our initial large-screen prototype, discussed earlier.
- To maximize effectiveness, we are locating the smaller display near the center of the large screen.

Future Work

- This way, the foveal and peripheral cones will maintain the correct geometric relationship at all times.
- We are presently combining such a system with the Virtual Window head-tracking mechanism, and look forward to reporting on its results in the near future.

Conclusions

- We have crossed the complexity barrier of current camera-monitor mediated telepresence applications.
- To beat the limitations imposed by this barrier, we propose a new design to support views of the foveal and peripheral cones simultaneously.
- To minimize the effects of spatial discontinuities, we also provide a seamless linkage between the two views.

Acknowledgments

- The authors would like to thank William Hunt and Shumin Zhai of the University of Toronto, Abigail Sellen of Rank Xerox EuroPARC and Masayuki Tani of Hitachi Research Laboratory, for their invaluable suggestions and contributions to this paper.
- We would also like to thank John Tsotsos of the University of Toronto for helping us sift through the relevant literature on biological vision.

Acknowledgments

- This research has been undertaken as part of the Ontario Telepresence Project.