

# **Design and Evaluation of Contextualized Video Interfaces**

Yi Wang

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy  
In  
Computer Science and Applications

Doug Bowman  
David Krum  
Chris North  
Francis Quek  
Tonya Smith-Jackson

August 5, 2010  
Blacksburg, Virginia

Keywords: Contextualized Videos, visualization, video surveillance, design space, 3D model, virtual environment, navigation, testbed, user study, testbed evaluation, task taxonomy, task performance, design guidelines, spatial mapping, mental rotation

Copyright ©2010, Yi Wang

# Design and Evaluation of Contextualized Video Interfaces

Yi Wang

## ABSTRACT

If “a picture is worth a thousand words,” then a video may be worth a thousand pictures. Videos have been increasingly used in multiple applications, including surveillance, teleconferencing, learning and experience sharing. Since a video captures a scene from a particular viewpoint, it can often be understood better if presented within a larger spatial context. We call such interactive visualizations that combine videos with their spatial context “Contextualized Videos”.

Over recent years, multiple innovative Contextualized Video interfaces have been proposed to taking advantage of the latest computer graphics and video processing technologies. These interfaces opened a huge design space with numerous design possibilities, each with its own benefits and limitations. To avoid piecemeal understanding of the design space, this dissertation systematically designs and evaluates Contextualized Video interfaces based on a taxonomy of tasks that can potentially benefit from Contextualized Videos.

This dissertation first formalizes a design space. New designs are created incrementally along the four major dimensions of the design space. These designs are then empirically compared through a series of controlled experiments using multiple tasks. The tasks are carefully selected from a task taxonomy, which helps to avoid piecemeal understanding of the effect of the designs. Our design practices and empirical evaluations result in a set of design guidelines on how to choose proper designs according to the characteristics of the tasks and the users. Finally, we demonstrate how to apply the design guidelines to prototype a complex interface for a specific video surveillance application.

## ACKNOWLEDGEMENT

I would like to sincerely thank:

- My advisor, Dr. Doug A. Bowman, for the guidance on my research and career
- The members of the thesis committee: Dr. David M. Krum, Dr. Chris North, Dr. Francis Quek, and Dr. Tonya Smith-Jackson, for their advice and discussions
- The 3DI group at Virginia Tech for all the helps and the inspirations
- The numerous experimental subjects who volunteered their time

I also want to thank my parents, Jianhua Wang and Xinrui Jia, for their great support of my career. Finally, I would like to thank my dearest friend and wife Yan Liang for sharing all the challenges and joys of life at Virginia Tech.

## TABLE OF CONTENTS

ABSTRACT.....	ii
ACKNOWLEDGEMENT .....	iii
1 Introduction.....	1
1.1 Problem Scenarios .....	1
1.1.1 Monitoring and Tracking .....	1
1.1.2 Procedural Learning.....	2
1.1.3 Virtual Tourism.....	3
1.2 The Challenges.....	3
1.3 Possible Solutions .....	4
1.4 Contextualized Videos .....	5
1.5 Research Questions .....	6
1.6 Contributions.....	8
1.7 Approach.....	10
2 Literature Review.....	12
2.1 Human Information Processing .....	12
2.1.1 Sensory Processing and Short Term Sensory Store .....	12
2.1.2 Perception .....	13
2.1.3 Working Memory.....	13
2.1.4 Attention .....	14
2.1.5 Long Term Memory.....	14
2.1.6 Conscious and Non-conscious Processing.....	15
2.2 Spatial Information Processing.....	15
2.2.1 Spatial Knowledge Representation.....	15
2.2.2 Viewpoint Specific Representations.....	16
2.2.3 Mental Registration of Multiple Views .....	17
2.2.4 Aids for Spatial Knowledge Acquisition and Navigation.....	17
2.3 Multiple Views of Spatial Data.....	18
2.3.1 Multiple Views Design .....	18
2.3.2 Combining 3D and 2D Views.....	19
2.4 Image-based rendering.....	20
2.5 Information Rich Virtual Environment.....	21
2.6 Non-Contextualized Abstraction and Visualization of Videos.....	22
2.7 Contextualized Video Techniques and Applications.....	23
2.7.1 Video Surveillance.....	23
2.7.2 Computer Supported Cooperative Work .....	24
2.7.3 Web Mapping Services and Geobrowsers .....	24
3 Current Practices of Building Surveillance.....	27
3.1 Building Surveillance Tasks .....	27
3.2 Video Surveillance Systems .....	28
3.3 Field Study .....	29
3.3.1 Social Background.....	29
3.3.2 Workplace Observations.....	30
3.4 Domain Characteristics.....	31
3.5 The Potential of Contextualized Videos .....	32
3.6 Summary .....	32

4	The Design Space of Contextualized Videos.....	33
4.1	Video-Model Layout Design .....	34
4.2	Model Visualization.....	39
4.3	Video Processing .....	40
4.4	Navigation.....	41
4.4.1	Navigation Context .....	41
4.4.2	Navigation Mode .....	42
4.5	Task Taxonomy .....	43
4.5.1	Existing Task Taxonomies.....	43
4.5.2	Contextualized Videos Task Taxonomy.....	44
4.5.3	Design and Evaluation Using the Task Taxonomy.....	47
4.6	Summary .....	47
5	Model Visualization and Video-Model Layout .....	48
5.1	Design Tradeoffs.....	48
5.1.1	Video-model Layout .....	48
5.1.2	Model Visualization.....	48
5.2	Exploratory Study .....	49
5.2.1	Usage Patterns.....	50
5.2.2	Result Summary.....	51
5.3	Formal Study.....	54
5.3.1	Tasks .....	56
5.3.2	Procedure .....	56
5.3.3	Results and Discussion .....	60
5.4	Guidelines .....	65
5.4.1	Design According to Task Characteristics.....	65
5.4.2	Design According to User Characteristics.....	68
5.4.3	Other Guidelines .....	69
5.5	Summary .....	70
6	Embedded Videos .....	71
6.1	Experiment.....	71
6.1.1	Designs.....	71
6.1.2	Tasks .....	72
6.1.3	Procedure .....	77
6.1.4	Results and Discussion .....	79
6.2	Guidelines .....	82
6.2.1	Design According to Task Characteristics.....	83
6.2.2	Design According to User Characteristics.....	83
6.3	Summary .....	84
7	Effect of Video Processing .....	85
7.1	Experiment.....	85
7.1.1	Designs.....	85
7.1.2	Tasks and Hypotheses.....	86
7.1.3	Procedure .....	87
7.1.4	Results and Discussion .....	87
7.2	Guidelines .....	88
7.3	Summary .....	89

8	Navigation.....	90
8.1	Multi-View Interface .....	90
8.2	Navigation Designs.....	92
8.3	Experiment.....	95
8.3.1	Experiment Design.....	95
8.3.2	Scene Generation .....	96
8.3.3	Tasks and Hypotheses.....	97
8.3.4	Procedure and Settings.....	100
8.3.5	Results and Discussion .....	101
8.4	Guidelines .....	108
8.4.1	Design According to Task Characteristics.....	108
8.4.2	Design According to User Characteristics .....	108
8.4.3	Other Guidelines .....	109
8.5	Conclusions.....	109
9	Application of Guidelines .....	111
9.1	Problem Scenarios .....	111
9.1.1	Monitoring and event handling .....	111
9.1.2	Creating a daily report .....	112
9.1.3	A security guard’s first week in the building.....	113
9.2	Interface Design.....	113
9.2.1	From Problem Scenarios to Tasks .....	113
9.2.2	From Tasks to Guidelines.....	114
9.2.3	From Guidelines to New Interface.....	115
9.2.4	Other possible features.....	121
9.3	Usage Scenarios.....	121
9.3.1	Monitoring and event handling.....	121
9.3.2	Creating a daily report .....	122
9.3.3	A security guard’s first week in the building.....	122
9.4	Summary .....	123
10	Summary and Future Work.....	124
10.1	Design Guidelines.....	125
10.1.1	Design According to Task Characteristics.....	125
10.1.2	Design According to User Characteristics.....	126
10.1.3	Other Guidelines .....	126
10.2	Future Work.....	127
	References.....	130
	Appendix A: Experiment 1 (Model Visualization and Video-Model Layout).....	138
	Appendix B: Experiment 2 and 3 (Video-Model Layout and Video Processing) .....	143
	Appendix C: Experiment 4 (Navigation).....	150

## LIST OF FIGURES

Figure 1-1: Video surveillance scenario: (a) Traditional interface; (b) Contextualized Videos. ....	1
Figure 1-2 Procedural learning: (a) a car brake repair tutorial video on YouTube. (b) a car brake system diagram [Sauers 2009]. ....	2
Figure 1-3: Virtual tourism scenario: (a) YouTube videos; (b) Contextualized Videos (Google Earth). ....	3
Figure 1-4: The incremental design and evaluation process.....	10
Figure 2-1: Wickens’ HIP Model [Wickens and Hollands 2000].....	12
Figure 2-2: (a) Video Flashlight [Sawhney, Arpa et al. 2002], (b) Dynamic Imagery [Sebe, Hu et al. 2003], (c) Moving Office [Schnädelbach, Penn et al. 2006], (d) Spatial Multi-Video player [Girgensohn, Shipman et al. 2007], (e) MERL forensic surveillance system [Ivanov, Wren et al. 2007]......	23
Figure 2-3: Google Street View in Google Maps .....	25
Figure 3-1: The Building Security Surveillance System at Building A.....	29
Figure 4-1: The Contextualized Video design framework .....	33
Figure 4-2: Mapping between transform functions and the resulting video-model layout designs.....	35
Figure 4-3: Associated Video. Callout lines are used to show association. ....	35
Figure 4-4: 2D billboard Video.....	35
Figure 4-5: Video on Fixed-planes. The video is hard to observe in (b). ....	36
Figure 4-6: Video Projection: (a) original video, (b) viewpoint approximately follows the video camera, (c) viewpoint far away from the video camera, revealing severe distortion and image fragmentation due to the missing door of the model.....	37
Figure 4-7: Video projection, video on fixed planes and dynamic imagery.....	38
Figure 4-8: The Drawer technique for visualizing a single floor.....	38
Figure 4-9: The Rotate-and-Shear technique for visualizing all the floors. ....	38
Figure 4-10: Landmark technique for visualizing the structure and reducing occlusion. ....	39
Figure 4-11: (a) Semitransparency and (b) Wireframe for visualizing internal structure of the building. ....	39
Figure 4-12: The Contextualized Videos task taxonomy. ....	44
Figure 5-1: The testbed for preliminary evaluation .....	49
Figure 5-2: Usage Patterns found in Experiment 1: (a) Pattern 2, 2D billboard + landmark view + explosion. (b) Pattern 2, Billboard + semi transparency. (c) Pattern 3, associated video + fixed plane video + landmark. (d) A view behind the camera used in Pattern 5. (e) Pattern 6, video projection + walls.....	53
Figure 5-3: Testbed Interface using the 2D Associated design. In the 2D Associated design, the camera glyphs indicated the camera’s location in the building model and the short line on the camera glyph indicated where the camera was facing. The testbed provided a “Replay” button allowing the user to replay the video multiple times, and a “Confirm” button allowing the user to confirm the path of the target actor and end the session. The elapsed time since the start of the task was shown in the lower-right corner.....	58
Figure 5-4: 2D Embedded Design. Left: By default videos were placed to face the	

camera on a 2D plane. In this camera-facing view, the video content was harder to observe, but mapping the actor’s location from the video to the model was easier. Right: When participants pressed the “ALT” key the videos were rotated to an upright position. .... 58

Figure 5-5: 2D Combined Design that integrated 2D Embedded and Associated Videos. .... 59

Figure 5-6: 3D Associated Design..... 59

Figure 5-7: 3D Embedded Design. The embedded videos were enlarged to make them easier to observe. Left: By default videos were placed to face the camera. Right: When participants pressed the “ALT” key the videos were rotated to face the user. .... 59

Figure 5-8: 3D Combined (Autorotation) Design, embedded videos were enlarged to make them easier to observe. Here, the user has clicked on the yellow video. 60

Figure 5-9: The two different observation cases in the 3D Embedded and 3D Autorotation Designs. Since the user is behind the yellow camera, the yellow video shows exactly what the camera sees. But since the user is on the opposite side of the red video from the red camera, the video canvas should be understood as a mirror reflecting what is happening in the model. .... 60

Figure 5-10: Comparing Task Time among designs..... 61

Figure 5-11: Comparing Task Precision among designs, higher score means higher precision. .... 61

Figure 5-12: Subjective Mental Workload Score of different designs ..... 61

Figure 5-13: Subjective rating of the designs, with 5 being the highest score. .... 62

Figure 5-14: A Task Characterization using 3 criteria..... 66

Figure 5-15: Cues of Embedded Video ..... 70

Figure 5-16: The video orientation and the object space proximity cue indicate Door 1 to be the one captured in the video, while camera icon orientation indicates Door 2 in the video. .... 70

Figure 6-1: The four Contextualized Video designs used in Task 1: (a) closely-related and camera-aligned video (CC); (b) closely-related and user-aligned video (CU); (c) remotely-related and camera-aligned video (RC); (d) remotely-related and user-aligned video (RU). .... 72

Figure 6-2: Task 1 screenshot. The 2D map was shown as a grey-scale image. The walls and doors were shown in black, while desks and cubicles, which could partly occlude the camera’s view, were shown in grey. The camera’s coverage area was shown in semi-transparent red on top of the 2D map, so that the landmarks captured by the camera can be easily identified..... 73

Figure 6-3: The virtual world path reconstruction activity. .... 77

Figure 6-4: Task 1 accuracy with controlled task time. (a) Main effect of video-camera distance; (b) Main effect off video-camera alignment; (c) Interaction effect of video-camera distance and video-camera alignment..... 79

Figure 6-5: Task 1 time with controlled task accuracy. (a) Main effect of video-camera distance; (b) Main effect off video-camera alignment; (c) Interaction effect of video-camera distance and video-camera alignment..... 80

Figure 6-6: The possible cognition processes of users when performing Task 1 and Task 2..... 82



Figure 7-1: Dynamic Path Visualization Prototypes used in the formal experiment. The path dynamically shows the target person’s trajectory from his first appearance in a video to the current moment. (a) Dynamic Path Visualization without videos. (b) Dynamic Path Visualization with videos.....	85
Figure 8-1: Prototype for multiple view interface. ....	92
Figure 8-2: Prototype for overview navigation interface. All the cameras and their coverage area should be shown in the 3D model. The red camera is selected and highlighted. ....	95
Figure 8-3: A screenshot of Task 2 with AO technique. ....	99
Figure 8-4: A sample map containing 6 choices. The path is drawn in color. Some choices differ in their overall shape and other choices differ in details.....	100
Figure 8-5: Task 1 result: the mean number of correctly selected videos in Task 1 (normalized among trials). (a) Main effect of navigation context; (b) Main effect of navigation mode; (c) Interaction effect of navigation context and navigation mode.....	102
Figure 8-6: Video and camera cluttering problem in AO condition. ....	104
Figure 8-7: Task 2 result: the mean number of correctly identified rooms (normalized among trials). (a) Main effect of navigation context; (b) Main effect of navigation mode; (c) Interaction effect of navigation context and navigation model.....	105
Figure 8-8: Task 3 result: the mean number of excluded choices (6 choices in total). (a) Main effect of navigation context; (b) Main effect of navigation mode; (c) Interaction effect of navigation context and navigation model. ....	106
Figure 9-1: The current video surveillance system used in Building A. ....	111
Figure 9-2: Interface for the new system, which contains a monitoring display and a working display. The working display contains four areas: The area on top is the Popup Videos area. At bottom is the Timeline. Of the two areas in the middle, the right is the Detailed View, which contains five tabbed panels. The left area is called Context Overview. ....	116
Figure 9-3: An enlarged view of the working display. The working display contains four areas: The area on top is the Popup Videos area. The area at bottom is the Timeline. Of the two areas in the middle, the right one is the Detailed View, which contains five tabbed panels. The left area is called Context overview. ....	117
Figure 9-4: Description of the five tabs in the Detail View panel. ....	117
Figure 9-5: Report generation panel. ....	118
Figure 9-6: 2D Context View panel. ....	118

## LIST OF TABLES

Table 1-1: The pros and cons of video and model data. Notice that they can complement each other. ....	5
Table 3-1: Working hours of the security guard at Building A. ....	30
Table 5-1: Six designs compared in the experiment.....	56
Table 6-1. The three difficulty levels of a Type 1 integration task.....	75
Table 8-1: Navigation technique categories with their pros and cons. ....	93
Table 8-2: The total number of left mouse clicks on three views of Contextualized Video interfaces recorded in Task 1. ....	103

# 1 Introduction

If “a picture is worth a thousand words,” then a video may be worth a thousand pictures. Videos record what happens around the world and contain huge amount of useful information that can be utilized to improve the way we work and live.

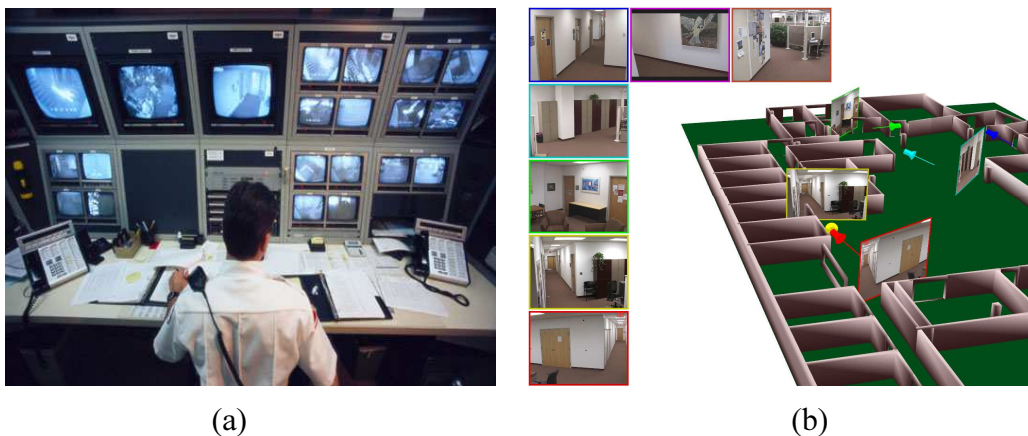
As the cost of video capture devices keeps dropping, videos have been increasingly used in multiple applications, including building and traffic surveillance, learning, experience sharing, teleconferencing, and tele-collaboration. Pocket camcorders and camera phones allow experience sharing at any time. ComScore Video Metrix service reported that a total of 31 billion videos were viewed online in the U.S. during the month of November 2009 [comScore 2010]. Wakefield, in 2002, reported that there were 25 million CCTV cameras in operation worldwide [Wakefield 2002]. With a huge amount of video data generated daily, it is often a challenge to efficiently present the events captured by the video to the viewers.

One strategy to utilize the video data is to support data mining and information retrieval on video data. For this purpose, researchers have been exploring automatic video content analysis technologies [Xu 2007] to extract information from the videos. Although significant progress has been made in this direction, it is still humans who need to analyze the extracted video data, synthesize a wide range of context information with the video content, and make decisions based on the data. Since humans have limited mental resources to do such high-level cognition tasks—and the situation is becoming much more complex as the number of videos increases—new video presentation interfaces should be invented to relieve this bottleneck.

## 1.1 Problem Scenarios

The following scenarios illustrate the importance of new video presentation interfaces with examples from multiple domains:

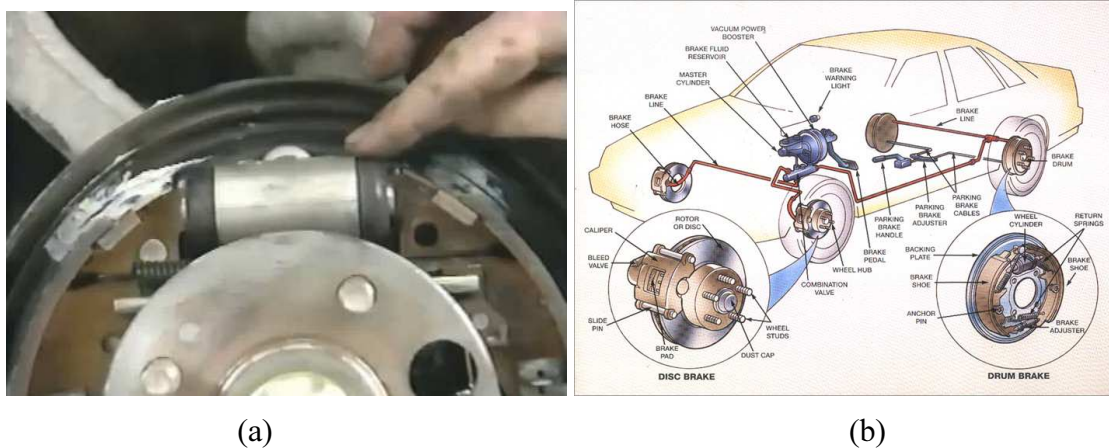
### 1.1.1 Monitoring and Tracking



**Figure 1-1: Video surveillance scenario: (a) Traditional interface; (b) Contextualized Videos.**

A large building security surveillance system often contains tens or even hundreds of videos monitoring different locations of a building (Figure 1-1 a). In a standard interface that displays the videos as thumbnails or cameos, the operator must maintain a detailed mental model of the building or site and perform numerous mental mappings in order to understand the activities shown in the videos. Previous research has shown that mental registration of multiple views is a challenging cognitive activity [Shepard and Metzler 1971; Pillay 1994]. Although some surveillance systems label the videos with text to describe the location of the videos, they are not very intuitive. For example, when a suspicious person is noticed walking through a video, the security guard often needs to determine the next video in which the suspect might appear, in order to observe closely. Because new security guards may not have a detailed mental model of the building and their response time might be too long, they may lose the opportunity to observe the suspicious person.

### 1.1.2 Procedural Learning



**Figure 1-2 Procedural learning: (a) a car brake repair tutorial video on YouTube. (b) a car brake system diagram [Sauers 2009].**

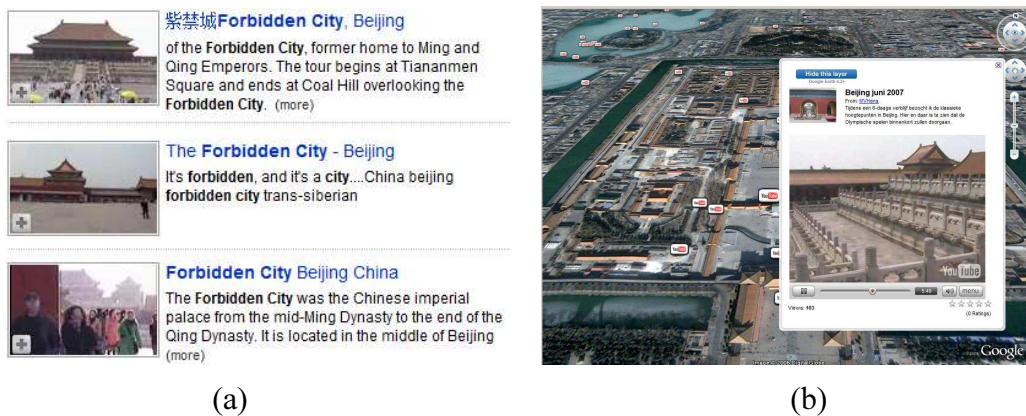
Video is efficient in conveying procedural knowledge. For example, to learn how to repair a car brake, one can look at a tutorial video on youtube.com (Figure 1-2 (a)). However, if the viewer wants to fix his own brake after watching the video, he also needs to have the structural knowledge about the brake system, which may not be clearly conveyed by the video. It might be helpful to augment the video with a visualization of the car brake system (Figure 1-2 (b)) when the teacher is operating on a car brake component, it will be highlighted in the diagram correspondingly.

Another example is offline route learning. Although navigation aid systems, typically portable GPS devices, are readily available, there are still situations where we need to learn a route beforehand. For example, when a new student is driving onto the Virginia Tech campus for the first time, he might stop at the map board located at the campus entrance and try to learn how to get to his dormitory in Room 310 of Slusher Hall. He may write down the street names where he needs to make turns into a route list. In this case, the knowledge he learned is mainly stored verbally. He may also draw the route and label the street names. However, he still has no idea of what the streets and buildings look like. Consequently, when he actually navigates through the environment, he has to

carefully look for street names and building names. He cannot utilize the rich visual cues provided by the actual environment, e.g., the structure and the material of the buildings when observed from the street, the actual size of the parking lot, and the appearance of the trees.

### 1.1.3 Virtual Tourism

In online video repositories like YouTube, people have uploaded many tourism videos that they captured while visiting famous sites (Figure 1-3 a). For example, dozens of videos can be found capturing the Forbidden City in Beijing. The current YouTube interface displays these videos in a list with simple text descriptions. After watching these videos one after another, people who have never visited the Forbidden City may not be able to construct a well-organized mental model of the site.



**Figure 1-3: Virtual tourism scenario: (a) YouTube videos; (b) Contextualized Videos (Google Earth).**

## 1.2 The Challenges

Multiple challenges exist in the above scenarios:

- Videos capture periods of real world situations. The observer often needs a more comprehensive presentation of that information. In such cases, **contextual information** is needed to help people make sense of the video content. Taking the virtual tourism scenario as an example; if a model of the real site is provided and the videos are mapped to their shooting locations on the model, the user's mental model will be better organized.
- Multiple correlated videos are often observed together, but **relating and mapping between multiple videos** is not an easy task. In the building security scenario, the videos can be presented together with a map or a model of the building (Figure 1-1 b). In such a case, the spatial relationship between the nearby videos can be understood at a glance. The reason is that such a presentation allows a user to replace the spatial knowledge recall task with the visual perception and recognition task, which requires less mental effort than the recall task [Wickens and Hollands 2000]. Also, an external representation of the building is more reliable than a user's long-term memory.
- To reduce the workload of watching a large amount of videos, some information

can be extracted from the videos and presented abstractly, though such abstractions often **lose the rich subtle cues** in the original video. These cues often contain important information and can often be easily captured by human beings subconsciously using very little attentional resources. In the route learning scenario, videos and photos capturing different parts of the campus can be presented together with the map. Such a system would allow the users to learn the appearance of the landmarks and even the process by which to navigate through a complex road system.

Please notice that the above challenges are likely to appear in high-level spatial cognition tasks involving situational awareness, sense-making and decision making. Hence I mainly investigate high-level cognition tasks involving spatial knowledge in this dissertation.

### **1.3 Possible Solutions**

Researchers have been addressing these challenges for decades. The video processing and computer vision community has been seeking mathematical concepts and algorithmic approaches to automatically extract information from video, e.g., tracking human forms and detecting anomalous behaviors from video sequences. These algorithms work in constrained situations. When the scene is complex or contains unexpected events that cannot be recognized by the algorithms, the intervention of human operators is needed.

Machine learning techniques have been employed for “trying to get machines to do much of the work in place of the human eye-brain system” [Davies 2005]. Although significant progress has been made, current systems are still not comparable with human eye-brain system in terms of intelligence, flexibility and reliability. The complexity and diversity of the realworld environments makes it unlikely for a general computer vision algorithm to make decisions for humans in any general condition in the near future. For example, if the crowd is dense, the person being tracked can be easily occluded or partially occluded by other people. The user often needs to integrate history information, subtle motion cues, and even social background information in order to solve the ambiguity. Automatic tracking algorithms often meet ambiguous situations because of the limited knowledge base of the algorithms. Whenever automatic video processing algorithms can not handle a situation reliably, video processing results will need to be presented to humans operators to solve ambiguities and make final decisions.

Since video processing cannot fully replace human operators in the near future, another promising direction is to visualize the video processing result for human operators. For example, if the camera parameters are known, we can extract the people and their movements and display them in correct 3D position in the environment model, so that the whole scene is reconstructed in 3D, eliminating the need for viewing individual videos. Recently, multiple approaches have been proposed along this direction. Sebe et al. presented an initial prototype of an “Augmented Virtual Environment” system which detected moving objects inside the video and visualized them as textured dynamic rectangles moving around in the 3D model [Sebe, Hu et al. 2003]. Following a similar idea, Girgensohn et al., in 2007, introduced a working DOTS surveillance system, which used video processing to extract foreground feature segments and visualize them in 3D environments. As technologies for precise tracking, 3D scanning and image-based rendering become mature, designs that visualize video processing results in 3D models

can be potentially helpful for certain tasks, e.g., understanding the path of the target person.

“Sight is the sense of choice” [Davies 2005]. Video processing, in a way, makes the choice for users by predicting users’ information requirements and then extracting that information from videos. However, a user’s information requirement might be unclear until he sees the video. Therefore, extracting the target in the video from the background may result in the loss of important cues, especially when the target interacts with the environment. For example, if a person passes a potted office plant, the motion of the plant gives the observer an important cue about the motion of the person, but that cue is hard to reconstruct in the 3D scene. Therefore, presenting raw videos together with the 3D environment might be a simple and effective way that can put the videos in context as well as preserve the subtle perceptual cues of raw videos.

## 1.4 Contextualized Videos

I propose *Contextualized Videos* – that is, a combination of raw videos with a model of the 3D environment – as a potential solution to the challenges.

A model of a spatial environment abstracts the main features of the environment into graphics elements, including geometry, simplified lighting, colors, and textures. As summarized in Table 1-1, model data has multiple advantages that can compensate for the disadvantages of video data. Videos convey dynamic information and rich subtle cues that are easy for humans to perceive and interpret. However, videos only capture 2D views of the 3D scene; hence they do not fully convey the internal 3D structure of the scene. Even the captured 2D features can be hard to interpret by humans and computers because of visual noise and poor view angles. Model data inherently does not have such problems.

Video Data	Model Data
+/- Detailed data	+/- Simplified data
+ Dynamic information	- No dynamic information
+ Numerous subtle cues	- Lack subtle cues
- Large data size	+ Small data size
- View dependent	+ View independent
- Noisy	+ Clean
- Lack structure	+ Structured
- Lack context	+ Context

**Table 1-1: The pros and cons of video and model data. Notice that they can complement each other.**

By combining videos with the model of the environment in which the videos are captured, the contextual information as well as the spatial relations between videos are presented in the visualization, allowing some cognitive work to be offloaded onto the perceptual system [Thomas and Cook 2005]. Hence Contextualized Videos respond directly to the challenges of a video data presentation interface by visualizing the users’ information need:

- Contextualized Videos allow videos to be played and understood in a **larger spatial context**, e.g., a 2D map or a 3D environment model. For example, in Google Earth, the videos can be linked to a specific location on the 2D map and



the user can have a rough idea of where the video was captured (Figure 1-3 b). With more precise registration of the video to the model, the observers can even see the activities in the videos in their proper locations.

- Contextualized Videos can provide a **visualization of the spatial relationships** among multiple videos and can ease spatial mapping. Figure 1-1 b shows an example of a Contextualized Video design for video surveillance tasks. The video screen is put in the object space of the environment model and is rotated to align with the camera's orientation. The relative position and orientation of the videos can then be easily perceived.
- Contextualized Videos allow abstract information to be presented in the model while still **keeping the subtle cues** in the original video. Lacking the live videos to learn the process of traveling from one place to another, we are used to learning a route by looking at 2D maps and remembering street names. As the number of surveillance cameras and online experience-sharing videos increases, it will be possible to find a proper video, or combination of videos to learn how to navigate along a selected route. In the route learning scenario, we can integrate the relevant videos with a map or model of the environment to show how to navigate through the environment. In this case, the user can utilize other types of cues such as the dynamic change of landmarks (visual cue) or even a personal feeling of a site (affective cue).

With the performance increase in computer hardware and the maturity of 3D graphics and video analysis algorithms, various Contextualized Video approaches are now technically feasible. A huge design space exists on how to combine videos with models to support varied tasks. Up to now, most research and applications have focused on the design of Contextualized Video techniques for a particular application [Sawhney, Arpa et al. 2002; Sebe, Hu et al. 2003; Girgensohn, Shipman et al. 2007; Ivanov, Wren et al. 2007; Chen, Neubert et al. 2009; de Haan, Scheuer et al. 2009]. Some of those works provided examples of good designs; a further question is how to help the designers to create good designs for a broad range of applications. The designers need advice on how to design the display based on the tasks to be supported in the application. Design guidelines were shown to serve this purpose [Bowman 1999; Baldonado, Woodruff et al. 2000]. Therefore, the **research goal** of this dissertation is to **provide design guidelines based on systematic investigation of the effect of Contextualized Video designs on tasks**.

Contextualized Video can be viewed as a special case of an Information-Rich Virtual Environment (IRVE) [Bolter, Hodges et al. 1995; Bowman, Hodges et al. 1999; Bowman, North et al. 2003; Polys 2006]. The integrated information space of IRVEs allows users to have a comprehensive mental model of the data to better understand the relationships within each data type individually as well as the relationships among the data types. However, Contextualized Videos pose additional challenges and trade-offs over traditional IRVEs. For example, the problem of mental mapping between multiple videos and between videos and the model is not addressed in previous IRVE research. The usefulness of the rich visual cues provided by videos is also to be investigated.

## **1.5 Research Questions**

The research goal can be decomposed into five sub-goals, or research questions. We achieve the research goal by addressing these questions step by step:



**Q1: What is the ontology of the design space of Contextualized Videos? What are the major design dimensions?**

I address this question by setting up a framework to systematically analyze, design and evaluate Contextualized Video techniques. Such a framework was proposed based on multiple research practices, including field study, design practice and literature review. The major design dimensions are the important decisions that every designer has to make when creating Contextualized Video interfaces. These decisions and the combination of these decisions are likely to have a significant impact on the usability of the interfaces.

My contributions related to this question include not only theories, but also implementations of Contextualized Video prototypes on a common testbed. These prototypes are created by combining representative design choices along each major design dimension. These prototypes form a reasonable coverage of the design space. Therefore, the design practice verifies the framework.

**Q2: For a particular domain (surveillance) and activity (path reconstruction), what are the usable Contextualized Video designs and their limitations?**

Once the representative Contextualized Video designs are prototyped, we want to understand their advantages and limitations. To address Q2, I need to delve into a domain and evaluate the designs using a domain-specific activity instead of basic and generic tasks. I believe that these tasks can be more reliably extracted after a case study of a particular domain.

Since the building security surveillance domain makes intensive use of videos, a path reconstruction activity is selected from this domain to evaluate the designs. The evaluation result, as well as the observations, will improve our understanding on how and exactly when the designs are usable. Such an understanding can help us decompose the activity into basic tasks in a reasonable way. Furthermore, with hard performance data in hand, I can form promising hypotheses about the effect of the designs on the basic tasks.

**Q3: What are the distinctive tasks in Contextualized Video interfaces, and how can we classify them in a way that is useful to designers?**

In order to help future designers to select the appropriate Contextualized Video design according to the tasks in hand, I propose a taxonomy of the tasks that can be performed on such an interface.

The design space is proposed from a designer's point of view, while the task taxonomy is proposed from a user's point of view.

Based on the literature review and case study, I extract general tasks from domain level activities and organize them into a taxonomy. I also consider tasks of other domains to avoid holes in the taxonomy. The taxonomy clearly shows the distinctive tasks of Contextualized Video interfaces.

**Q4: What are the effects of various Contextualized Video designs on the performance of key Contextualized Video tasks?**

After evaluation of the Contextualized Video designs using building surveillance activities (Q2), I have generated multiple promising hypotheses about the advantages and disadvantages of the designs. Naturally, the next step is to formally test these hypotheses using the basic tasks selected from the task taxonomy (Q3). Multiple formal

experiments are performed to evaluate the designs created by combining the design choices along the four design dimensions. The evaluation results are summarized into the design guidelines.

### **Q5: How beneficial are Contextualized Videos in complex activities?**

This research question motivates me to test the external validity of my findings on Q4 with complex and realworld activities. The tasks used in addressing Q4 are basic tasks and may only take a small portion of the whole activity time. Moreover, users normally have a very flexible strategy to achieve a higher level task goal. For example, they might be able to utilize other information to avoid tasks that take a relatively long time.

Although hierarchical task analysis can be used to decompose a complex activity into basic tasks and the proportion of a particular task in the whole activity can be estimated, it is often difficult to fully capture the highly dynamic and flexible human strategy. Often, overemphasizing the steps can cause one to miss the forest for the trees [Rosson and Carroll 2002].

Application designers are often faced with the tradeoffs among the optimization of multiple different tasks of a complex activity. The findings related to Q5 give a reference case that quantitatively demonstrates how much a complex activity will be affected by different designs. In this way, application designers will be able to compare expected activity with the demonstrated activity, and make decisions accordingly.

Summarizing our findings from addressing Q1-Q5, we are able to propose a set of design guidelines that can help future designers to select proper Contextualized Video designs for a given task, or at least make reasonable design decisions. To demonstrate how to use the design guidelines, I apply the design guidelines to create a complex interface for a specific video surveillance application at the end of this dissertation.

## **1.6 Contributions**

This dissertation makes a number of contributions to the fields of visualization and HCI (Human Computer Interaction):

- A structured description of the design space of Contextualized Videos. Individual Contextualized Video designs existed before this dissertation. This dissertation describes a design space that allows us to analyze existing designs along multiple design dimensions. The design space also helps us systematically identify new design possibilities by combining choices along each design dimension.
- A set of new Contextualized Video techniques along each design dimension. This set of techniques forms a design palette from which the application designer can knowledgably select techniques to match the needs of a particular task and its users.
- A task taxonomy that highlights the distinctive tasks that can benefit from Contextualized Videos, and guides systematic evaluation of Contextualized Video designs.
- Design guidelines for Contextualized Video application design. These guidelines not only provide design recommendations for given tasks but also provide a framework that guides designers to systematically analyze the tasks and

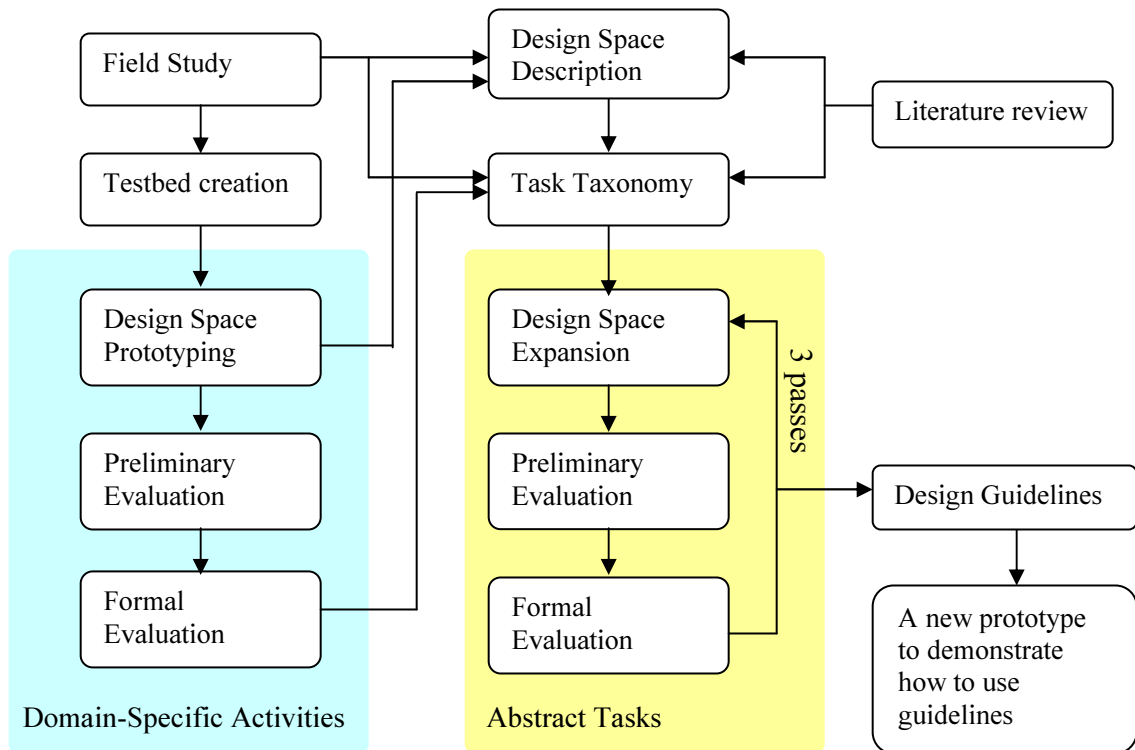
understand the tradeoffs between design choices. The application designers can find example designs by specifying the task characteristics, user characteristics and data characteristics of their application. These examples should be viewed as starting points of the design process instead of final designs. They should be viewed as visual components instead of a complete interface. Also, the design guidelines are by no means complete. Instead, they provide an initial version that other researchers can expand and improve.

- A Contextualized Video testbed as well as a library of Contextualized Video techniques. These can be reused to construct and evaluate new Contextualized Video designs.

## 1.7 Approach

The research in this dissertation follows an incremental design and evaluation approach.

I start with a survey of the building surveillance domain in order to understand the problem scenarios and how Contextualized Videos can help (Chapter 3). Then a testbed is created to support the prototyping and evaluation of multiple designs. The testbed allows me to form creative Contextualized Video designs by freely combining various techniques along multiple design dimensions [Bowman and Hodges 1999; Bowman, Johnson et al. 2001].



**Figure 1-4: The incremental design and evaluation process**

Based on the survey and the design practice, I characterize the design space into four major dimensions (Chapter 4). Since it is overly complex to explore all four dimensions in one experiment, I separate the process into four passes, one domain-specific and three abstract.

The first pass delves into the video surveillance domain (Chapter 5). In this first pass, I only prototype simple and static techniques along two design dimensions and use a domain-specific activity to evaluate them. The evaluation result generates promising hypotheses about the designs.

After the first pass, a general task taxonomy is proposed and the distinctive tasks of Contextualized Video interfaces are highlighted (Chapter 4.5). The taxonomy allows me to systematically investigate the effect of different designs on general tasks across domains.

In the second pass, I then use general tasks and activities to test the hypotheses (Chapter 6). The evaluation result forms an initial set of design guidelines. Also, a

baseline design is selected. After the two passes, a deep understanding of the two design dimensions is formed. I then create more complex designs during the third and fourth passes by utilizing the previous evaluation results and adding design choices along the other two dimensions (Chapter 7 and 8). The evaluation results are summarized into design guidelines. I demonstrate how to apply these design guidelines to create a relatively complex interface to support multiple tasks in a specific video surveillance application (Chapter 9).

Figure 1-4 summarizes the whole incremental design and evaluation process followed by this dissertation.

## 2 Literature Review

Contextualized Videos are designed to support complex cognition tasks that require the integration of spatial information with dynamic video information. A model of the human information processing (HIP) stages can provide a framework to analyze the psychological process used in performing such tasks and provide theoretical support for the creation of new designs. We review HIP literature in Chapter 2.1. Since the emphasis of my research is spatially related tasks, literature on spatial information processing will be covered in Chapter 2.2. Because Contextualized Videos can be viewed as a special case of Multiple View Visualization and an Information Rich Virtual Environment, related works will be discussed in Chapter 2.3 and 2.5 respectively. Chapter 2.4 discusses the similarity and difference between Contextualized Videos and Image-based rendering. Literature on non-contextualized visualization and abstraction of video data will be summarized in Chapter 2.6, and that of Contextualized Video techniques and applications will be summarized in Chapter 2.7.

### 2.1 Human Information Processing

A widely accepted human information processing model is summarized by Wickens [Wickens and Hollands 2000], as shown in Figure 2-1. HIP is represented as a series of stages. I assume the user is working on a purely visual interface when describing the stages.

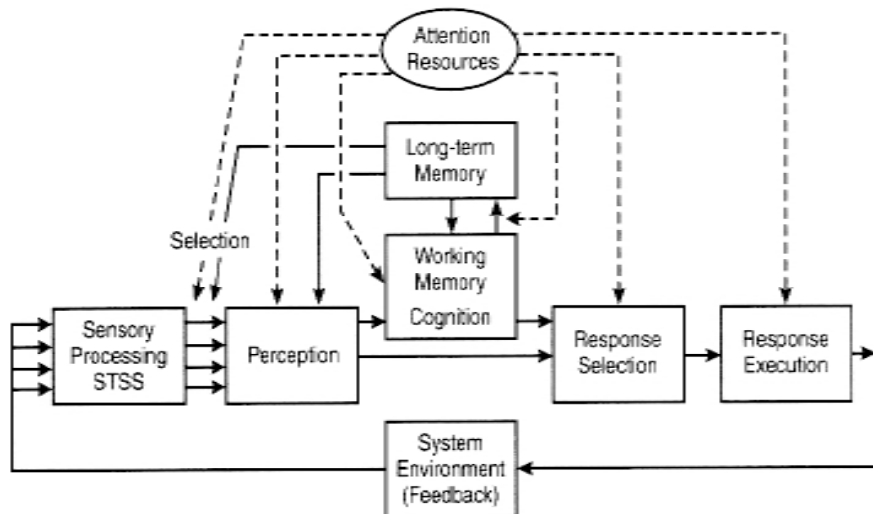


Figure 2-1: Wickens' HIP Model [Wickens and Hollands 2000]

#### 2.1.1 Sensory Processing and Short Term Sensory Store

During the first phase of Wickens' HIP model, retinal sensors in the eye receive visual stimuli from the interface and transmit the signal to occipital lobe, where the basic visual properties, including edges and shapes, are extracted. The visual sensory system is associated with a short-term sensory store (STSS), which buffers the raw stimuli for about 0.5 seconds to support information processing. Even such a short term and low-level information buffer can be utilized in interface designs. For example, in some

Contextualized Video designs, I allowed the users to fast shift between two different video layout designs by pressing the space key on the keyboard. The users reported this technique to be useful because they could use both views simultaneously.

### **2.1.2 Perception**

Through the stage of perception, raw sensory data were interpreted and given meaning. Perceptual processing has two important features that can be utilized in interface designs. First, it generally proceeds automatically and rapidly, requiring little attention (pre-attentive processing). Second, it is driven both by sensory input and by inputs from long-term memory about what events are expected. Most of the time, the two factors work harmoniously together. But under certain situations two typical kinds of processing can occur: bottom-up processing, where the sensory input plays a more important role, and top-down processing, where the internal context plays a more important role.

A good Contextualized Video design should allow some cognition work to be offloaded to the perceptual system [Thomas and Cook 2005]. For example, visualizing the floor plan around the video's coverage area allows users to perceive the spatial context, instead of having to recall the context from their long term memory into working memory.

### **2.1.3 Working Memory**

Cognition, such as rehearsal, reasoning, or image transformation, is a conscious activity that transforms or retains information [Norman and Bobrow 1975]. It differs from perception by requiring greater processing time, higher mental effort and more attention. Cognitive operations are carried out by using working memory [Baddeley 1986]. While there are many different working memory models [Miyake and Shah 1999], most include a central executive that controls the flow of information, a verbal component (comprised of a phonological store and an articulatory loop) and a spatial component called the visuospatial sketchpad. The visuospatial sketchpad stores information in an analog, spatial form, often typical of visual images [Logie 1995].

Since it was found that multi-tasking using only the visuospatial sketchpad or phonological store is susceptible to interference with the concurrent activities, Contextualized Video interfaces can possibly utilize both of the two components simultaneously to better support multi-tasking without introducing interference problems.

Working memory (WM) has a limited capacity, holding only the gist of meaning, instead of the exact stimuli. The information recorded in WM is in a simpler form and has to use information from long-term memory to explain it. For adults the information in WM can only be retained 10-25 seconds without rehearsal. Research on working memory [Phillips 1974; Jiang, Olson et al. 2000; Vogel, Woodman et al. 2001] suggests an item retention capacity of 5-9 for verbal items and 3-5 for visual items, where conjunctions of encodings can be considered a single item. The capacity limit of working memory imposes the challenges for relating multiple videos and their context. Contextualized Videos should help users to chunk the video information and reduce the mental effort to recall and transform spatial information.

## 2.1.4 Attention

One of the most formidable bottlenecks in HIP is the limitation of human attentional resources. Failure of attention can happen when the internal task goal and the external interface require too much information to be processed consciously. Three different types of failure of attention problems were identified by Ware [Ware 2000]:

- **Selective Attention.** In this case, we select inappropriate aspects of the interface or subtasks to process. For example, when multiple videos are playing simultaneously we may fail to select the correct video to watch at a given moment because no proper cue is provided on the interface. Previous research has summarized multiple design guidelines to prevent the problem of selective attention [Elkind, Card et al. 1990; Wickens, Vincow et al. 1997; Wickens and Hollands 2000]. For instance, it is stated that data that is often of interest should be placed centrally, and information that is typically viewed sequentially should be placed close together [Elkind, Card et al. 1990; Wickens and Hollands 2000].
- **Focused Attention.** When a focused attention problem occurs, the user is distracted by extraneous information and could not focus on one particular stimulus. Such a situation may happen when the user is monitoring multiple videos in which several distracting people walk through the videos. The user may find it hard to focus his attention on a particular video because the distracters' movements keep catching the observer's eye.
- **Divided Attention.** When a divided attention problem occurs, the user is unable to divide his attention among multiple stimuli or tasks, which need to be processed immediately. In a video surveillance situation, the user often needs to do multiple tasks simultaneously. For example, he may need to closely observe one suspicious person in a video while continuing to monitor the rest of the videos. Typically, the more attention that must be focused on a single task, the worse people will be with dividing their attention between tasks. Hence a proper interface should minimize the effort needed to closely observe the suspicious person.

A good Contextualized Video design should try to avoid such failures of attention. For example, if the video size is too small to observe, a divided attention problem may occur.

## 2.1.5 Long Term Memory

Long Term Memory (LTM) is memory, stored as meaning, that can last as little as a few days or as long as decades. It differs structurally and functionally from working memory. Biologically, short-term memory is a temporary potentiation of neural connections that can become long-term memory through the process of rehearsal and meaningful association. A user's spatial knowledge about a site is stored in LTM.

The brain does not store memories in one unified structure. Instead, different types of memory are stored in different regions of the brain. LTM is typically divided up into two major headings: declarative memory and procedural memory.

- **Declarative memory** refers to all memories that are consciously available. Declarative memory also has two major subdivisions:
  - **Episodic memory** refers to memory for specific events in time, such as a video showing how to make a sequence of turns in a complex road network.



- Semantic memory refers to knowledge about the external world, such as a map of the surrounding environment of the video.
- Procedural memory refers to the use of objects or movements of the body, such as how exactly to drive a car.

Contextualized Videos can potentially allow the user to utilize both episodic memory and semantic memory in performing complex tasks like route learning.

### **2.1.6 Conscious and Non-conscious Processing**

Perception and cognition lie on the two ends of a continuum of consciousness. A widely accepted theory is that our brain works at multiple conscious levels in parallel [Wang, Wang et al. 2006]. Humans can be trained to perform some tasks non-consciously or sub-consciously. When humans acquire a skill, some aspects of performance can be automatized to require less cognitive and attentional resources. Automatized processes reduce cognitive overhead, as they do not involve conscious control or attentional resources. As such, they can usually be performed in parallel with other tasks, and are usually obligatory [Eysenck 2001]. However, some aspects of complex task performance should not be automatized in order to guarantee sensitivity and flexibility to novel situations. These aspects of performance should remain controlled and receive proper attentional resources.

The levels-of-consciousness theory is actually the basis for embodied interaction theory [Dourish 2001]. We make sense of things at multiple levels simultaneously, and an effective interface should support higher-level sense-making by providing visualizations and affordances that matches those stored in the users' memory. If users have to consciously make sense of many lower level interaction steps or the connection between these steps, the interface is not as effective. One goal of my research is to provide Contextualized Video design examples that allow users to focus their attention on high level tasks by reliably offloading low-level processing onto sub-conscious interaction with the interface. The interface should know the user's information needs and provide them in an intuitive way.

## **2.2 Spatial Information Processing**

### **2.2.1 Spatial Knowledge Representation**

Spatial memory is the part of memory responsible for recording information about one's environment and its spatial orientation. Lynch's classic work attempted to determine characteristics that made it possible for individuals to perceive, develop, and maintain a mental image of their own city [Lynch 1960]. He postulated five urban elements that form a design palette for a city that is easy to mentally visualize. His work sparked a large number of studies on spatial cognition.

Later works found that mental representation about a geographical area is based on three kinds of knowledge: landmark knowledge, route knowledge and survey knowledge [Piget and Inhelder 1967; Thorndyke and Hayes-Roth 1978; Wickens and Hollands 2000]. Landmark knowledge and route knowledge are both egocentric, while survey knowledge is exocentric. Landmark knowledge is a visual representation and route knowledge is a proceduralized verbal knowledge. Survey knowledge is an abstract that allows the traveler to draw an accurate map of the environment. If we repeatedly visit an area, we

normally gain knowledge in the order: landmark, route and survey. But if we learn an environment exclusively through a map, we can have very good survey knowledge, but very little landmark and route knowledge [Thorndyke and Hayes-Roth 1978; Williams, Hutchinson et al. 1996].

Stevens [Stevens and Coupe 1978] determined that our survey knowledge (mental maps) appears to be hierarchical. The hierarchy can be based in part on political, cultural, or structural information. These hierarchies allow simplifications in encoding the mental map. For example, many individuals feel that Reno lies east of San Diego since the state of Nevada lies east of California. While this is an inaccurate notion of the relative locations of the cities, it reveals a common simplification in the mental model of the western US.

The three kinds of spatial knowledge can also be learned in a virtual environment. Multiple research studies have been performed to understand the transfer of spatial knowledge from the virtual environment to the real world [Waller, Hunt et al. 1998; Jones 1999; Péruch, Belingard et al. 2000; Foreman 2005]. For example, Waller found that given enough training time, the route and survey knowledge learned in a virtual environment can successfully transfer to real environment [Waller, Hunt et al. 1998]. However, Contextualized Videos augment the virtual and sometimes the abstract environment with videos. The usefulness and usability of such a system is not well studied. Videos provide a more realistic view of the landmarks, and show how to navigate along the route with more details. Therefore, users might be able to learn better landmark and route knowledge using Contextualized Videos. Moreover, we are likely to dynamically integrate the three types of knowledge to make decisions when navigating through the real world. Contextualized Videos might improve our knowledge integration performance.

Spatial knowledge may not be encoded only in analog representation. Medin reported that people used both analog and propositional representations to make spatial judgment [Burgess, Becker et al. 2001]. This research indicates that people may also use multiple visual cues, instead of a single cue, to link between a video and the model. For example, a numbered label can be used together with color coding and callout lines to provide redundant cues to help users remember the links.

### **2.2.2 Viewpoint Specific Representations**

It has long been noticed that we can use two different viewpoint specific representations when performing spatial tasks. Howard etc. gave a good summary of the two representations [Howard 1991]. The egocentric, or user-centered representation, is developed for close interaction with the realworld. Thus, it encodes the distance and orientation of the objects in the environment relative to our body. The information changes as we move around. It is useful for maintaining the perception of a stable world from moment to moment, but cannot ensure the alignment of spatial representation in large-scale navigation [Wang and Brockmole 2003]. It becomes unreliable after a small number of rotations [Etienne, Maurer et al. 1996]. Nonetheless, humans can become familiar with the layout of a large-scale environment by using exocentric representation, or a world-centered view, which is developed for route planning or understanding of the environment during a longer time span. It is a type of survey knowledge and is not affected by our current orientation or position.

When designing an interface to support spatial tasks, we can visualize the environment either in an egocentric view or an exocentric view. A task-display dependency was suggested by several works [Aretz 1991; McCormick, Wickens et al. 1998; Wickens and Hollands 2000]. Generally, tasks involving navigation or actual travel through the environment are best supported by an egocentric view, and tasks involving understanding of the structure of the space tend to favor more exocentric viewpoints.

Recent research in experimental psychology indicates that egocentric and exocentric representations exist in parallel and complement each other in spatial related tasks [Sholl and Nolin 1997; Burgess, Becker et al. 2001; Wang and Spelke 2002; Mou, McNamara et al. 2004; Burgess 2006; Waller 2006]. Although the question of how they interact is not yet fully answered, the experiments implied a process of translation between the two representations. Spatial updating can be seen as a continuous repetition of this translation between systems, if movement velocity is also taken into account [Byrne, Becker et al. 2007]. Contextualized Videos visualize the two representations in one interface and have the potential to support more efficient and reliable translation between the two representations.

### **2.2.3 Mental Registration of Multiple Views**

Previous research has shown that using mental rotation to register multiple views is a challenging cognitive activity [Shepard and Metzler 1971; Pillay 1994]. For instance, Shepard found that it takes approximately one second for every 60 degrees of mental rotation [Shepard and Metzler 1971]. Nevertheless, Tory's experiment suggested that when aligning a 2D and a 3D view with corresponding features, people can use feature mapping to avoid mental rotation [Tory 2003]. In Contextualized Video designs, it is also possible to provide common features between the video and the model to facilitate mental registration of the two views.

### **2.2.4 Aids for Spatial Knowledge Acquisition and Navigation**

Paper-based maps, wearable computers and GPS devices are widely used as online aids for navigation and acquisition of spatial knowledge. On the one hand, the lessons learned from the design and evaluation of such systems are helpful for Contextualized Video design. On the other hand, Contextualized Videos can potentially be used to augment these systems.

Multiple works have investigated the effect of map design on navigation tasks [Gailing, Lindberg et al. 1983; Moeser 1988; Lawton 1996; Darken and Cevik 1999; Hochmain and Frank 2002; Dalton 2003; Soh and Smith-Jackson 2004]. Darken and Cevik compared a standard north-up map to a forward-up map and found that forward-up map was preferable in egocentric search tasks while a north-up map led to better performance for exocentric search tasks [Darken and Cevik 1999]. Soh and Smith-Jackson investigated two map design features, i.e., contour representations and color representations, on a wayfinding task [Soh and Smith-Jackson 2004]. The task was quite different from video surveillance tasks. Also, the maps were paper-based 2D maps, but their findings revealed users' high dependence on natural contextual cues to perceive map cues.

Kehikoinen and Suomela introduced a wearable computer based navigation guide named Walkmap [Lehikoinen and Suomela 2002]. Walkmap used a two dimensional map

that could be rendered from a topdown viewpoint or a perspective viewpoint. The authors performed a study involving a navigation task to determine which viewpoint would lead to better performance. The authors found that the perspective map led to slower completion of their navigation course. Participants commented that the top-down map was easier to use than the perspective map.

Krum investigated the use of wearable computers to help individuals understand and learn the layout of the surrounding environment [Krum, Ribarsky et al. 2001; Krum 2004]. The wearable computer rendered the environmental data from a top-down perspective and rotated the map to maintain a forward-up alignment. The user studies showed that a wearable computer, when appropriately configured, can help an individual learn the structure of the environment better than an unassisted individual. The studies also suggested that a perspective viewpoint may be distracting for users. A spatial cognition aid should provide a top-down viewpoint, which better matches the survey knowledge that should be taught to the user. Finally, the spatial cognition aid should account for divided visual attention, because the user has to contend with information from the wearable computer display as well as from the environment.

GPS navigation systems have grown into a huge market in recent years. A significant deficiency of models of the environment used in current GPS navigation systems is that too few details are shown to support maneuvering within a local space. A video can show an egocentric view of the world with enough details for a particular site, but it does not show the large-scale context. A natural hypothesis is that a combination of the videos with the map or model, i.e. Contextualized Videos, allows free switching between egocentric and exocentric views, and hence can more fully support spatial navigation tasks. For example, when following a route in realtime using GPS, videos allow fast and easy matching between the navigation interface and the realworld view.

## ***2.3 Multiple Views of Spatial Data***

Since the video and the model show different views of a common environment, Contextualized Videos share some common characteristics with the interfaces that utilize multiple views to present spatial data.

### **2.3.1 Multiple Views Design**

Baldanodo et al. summarized eight guidelines for the design of multiple views for visualization [Baldanodo, Woodruff et al. 2000]. These design rules could be used to analytically evaluate different Contextualized Video designs, yet not always lead to a clear answer. For example, according to the Gestalt similarity principle [Wertheimer 1923; Ellis 1938; Chandler 1997; Ware 2000] and the Rule of Consistency in Baldanodo's guidelines, 3D models have the potential to facilitate mapping between the video and the model, because when observed from the camera's viewpoint (an egocentric view) the 3D model's geometry features match those captured by the video. The matching features can help communicate the location and orientation of the subject captured by the video. However, it has also been shown that tasks involving spatial understanding favor more exocentric viewpoints like a 2D map or a top-down overview of the 3D model, while tasks involving navigation and tracking favor more egocentric views [McCormick, Wickens et al. 1998; Wickens and Hollands 2000]. Since the path reconstruction task used in the formal evaluation of the first research cycle involves both

tracking and spatial understanding subtasks, the tradeoffs can be better understood through an empirical study.

The focus-plus-context principle for information visualization [Ware 2000] suggests a multiple view design to show both the local detail and the global context. Ben Shneiderman summarized a general guideline for interactive information visualization [Shneiderman 1996]. Among the seven top-level subtasks to be supported are overview, zoom, and detail-on-demand, which can be well supported on a focus-plus-context design. Ruddle et al. showed that a combination of local and global maps is better than a single map for navigation in a very-large scale virtual environment [Ruddle, Payne et al. 1998]. Contextualized Videos are inherently focus-plus-context designs. However, the design space is very large because of the dynamic nature of the videos, the possibly large number of videos and the flexibility of interactions to accomplish Shneiderman's subtasks.

Quek et al. proposed VisSTA -- a multiple-view interface for analysis and annotation of videos [Quek, Bryll et al. 2002; Quek, Shi et al. 2002; Shi, Rose et al. 2004]. To help finding the right video shots from a large collection of videos, videos were represented hierarchically in multiple views. To ease perception of the actions, the actions of the person in conversation were extracted and animated using an avatar. A timeline-based graph was further provided to assist in content searching. The system maintains temporal situatedness by keeping all components synchronized with the current time focus. Although spatial context is not the focus of their research, they revealed the importance of context to understanding and using video data.

Yost et al. investigated the visual scalability of integrated (space-centric) and multiple views (attribute centric) for large high resolution displays [Yost 2006a; Yost and North 2006b; Yost, Haciahmetoglu et al. 2007]. They specifically used geospatially referenced multidimensional time-series data in their experiments. In [Yost, Haciahmetoglu et al. 2007], they found that a space-centric view shows better display scalability on spatial and temporal overview tasks. Users also felt more comfortable using space-centric visualizations. The authors believed that spatially grouping information and visual aggregation are both important factors in the improved performance of space-centric visualizations on large displays. Visual aggregation resulted in less physical navigation and thereby allowed for less mental demand and effort. Spatially grouping information also appeared less busy to the users compared to the attribute-centric visualization. Space-centric visualizations also out-performed attribute-centric visualizations for some detailed searching tasks where the query contains spatial information. Although it is unclear whether their results directly apply to Contextualized Video application designs, their findings indicate that embedding video data, or an abstraction of video data, into their spatial context might be useful, especially for large high resolution display.

### **2.3.2 Combining 3D and 2D Views**

Additional models for coordinating multiple views for exploratory visualization that includes 2D and 3D views has been described by [Roberts 1999] and [Boukhelifa, Roberts et al. 2003]. In Roberts' Waltz system, multiform 2D and 3D visualizations of data are displayed and coordinated as users explore sets and subsets of the data. Tory et al. investigated how to combine 2D and 3D views for volume visualization and spatial relationship tasks [Tory 2004; Tory, Kirkpatrick et al. 2006]. In [Tory, Kirkpatrick et al.

2006], they found that the users may use pattern matching instead of mental rotation to link two views. This may also be true when people try to link a video and a model. Following Tory et al.'s definition for "3D views," both a video frame and a perspective view of the model are 3D views. However, they come from different sources: the video is captured while the model is pre-computed. Combining several 3D views from multiple data sources for visualization is not a well explored problem.

## **2.4 Image-based rendering**

At first glance, Contextualized Videos are similar to image-based rendering. I will explain image-based rendering and clarify the difference in this chapter.

Image-based rendering techniques render novel views of the 3D world directly from input images. It is different from the traditional 3D computer graphics in which the 3D geometry of the scene is known. However, there are some image-based rendering methods that utilize some geometry information [Shum and Kangues 2000]. The two rendering methods can be understood as two points on a continuum, depending on how much geometric information is used. Pure image-based rendering techniques, e.g., light field rendering [Levoy and Hanrahan 1996], does not need any geometry information. Some other methods render with implicit geometry or depth information of the image pixels, e.g., the correspondence between images. High computation cost limits the use of a pure image-based rendering technique for realtime scene navigation. The most practical approach is to combine image-based rendering with traditional 3D graphics in a joint image and geometry space. Some examples are image warping using depth information [Mark, McMillan et al. 1997], view morphing using correspondence between images [Seitz and Dyer 1996], and lumigraph using approximate shape [Gortler, Grzeszczuk et al. 1996]. The hybrid image-and-geometry-based rendering techniques look similar to Contextualized Videos in the sense that captured data and pre-modeled data are used together.

However, the goal and application of image-based rendering and Contextualized Videos are quite different. Image-based rendering is used to create a static three-dimensional virtual world for navigation, but Contextualized Videos seek to use the environment model to help understanding of the dynamic information from the videos. The environment model can be composed of either geometry data or the pixel data derived from images (image-based rendering). Thus, theoretically we can create Contextualized Videos using image-based models.

It is still not very practical to use image-based models in Contextualized Videos, nor will it be in the near future. First, computer vision technologies are not robust enough to recover accurate depth information and 3D models. Second, the videos are often captured from fixed locations and directions, and do not provide dense enough views for image-based view reconstruction. Image-base warping needs either a panoramic view or multiple different views with camera information. Light field methods need even denser camera coverage. It is virtually impossible to reconstruct a reasonably complete light field [Levoy and Hanrahan 1996] or plenoptic function [McMillan and Bishop 1995] for image-based rendering. Thirdly, in real world situations, we often do not have access to adequate camera parameters, which are necessary for automatic mapping between the static objects captured by the video and those in the model. Gaining such parameters

requires a tedious calibration process. Therefore, I will only investigate Contextualized Videos where the environment context is represented as geometry models.

## **2.5 Information Rich Virtual Environment**

Information Rich Virtual Environment (IRVE) refers to a virtual environment (one that contains mainly spatial and perceptual information) that integrates abstract and symbolic information like text, links, numbers, graphical plots, and audio/video annotations [Bolter, Hodges et al. 1995; Bowman, Hodges et al. 1999; Bowman, North et al. 2003]. IRVE is the combination of Information Visualization and Virtual Environment techniques. The integrated information space of IRVE allows the users to have a comprehensive mental model of the data, to better understand the relationships within each data type individually as well as the relationships between the data types.

In [Dykstra 1994], Dykstra demonstrated how X11 windows could be embedded and rendered within virtual environments; this is an enabling technique for IRVEs. Plumlee and Ware have used multiple embedded views and frames of reference for the navigation of large-scale virtual environments [Plumlee and Ware 2003].

Bowman et al. implemented and evaluated Information-Rich Virtual Environments through a Virtual Venue and a Virtual Habitat [Bowman, Hodges et al. 1998]. The Virtual Venue included a 3D model of the venue, text information, and audio “labels” for various components of the environment, spatial hyperlinks between text and locations, a slideshow of images related to the environment, and an animated diver that could perform several types of dives. An evaluation of this application showed that the most effective information presentation techniques were those that were “tightly coupled” to the environment. The Virtual Habitat application is an IRVE for environmental design education. The 3D model was enhanced with abstract information including textual signs, audio clips, and overview maps. An evaluation of student learning in this application showed a trend towards increased understanding of environmental design principles. These two early applications showed the enormous potential of IRVEs to allow users to form mental associations between perceptual and abstract information.

Chen et al. further innovated IRVE interfaces involving immersive technologies [Chen, Pyla et al. 2004] . In [Chen, Pyla et al. 2004], they examined exploration and search tasks in immersive IRVEs using Head-Mounted Displays (HMDs). They found that for naïve search, the Viewport Space layout was significantly better than Object space layout of annotations for both navigation types. In [Chen, Narayan et al. 2005], she compared the time and attention costs of context switching between a desktop display (Display Space) and a tablet PC + CAVE wall (User space).

Parallel Graphics’ Virtual Manuals solution demonstrates the integration of temporal information within the spatial world and with external windows for training applications in operation and maintenance. Temporal information is rendered through animated sequences of assembly and disassembly. This approach is consistent with HCI research in comprehension and user’s mental models. For example, users gained improved situational awareness, understanding, and recall through multimedia presentations integrating these features [Sutcliffe and Faraday 1994; Faraday and Sutcliffe 1996]. Their result supported the validity of Contextualized Videos research. However live videos were not used and tested in their system.

Polys' dissertation is the first work that systematically investigated the layout techniques in IRVEs that include complex visualizations of textual information [Polys 2006]. In his dissertation, Polys proposed an IRVE design space and an IRVE task taxonomy, evaluated the effectiveness of various design choices, and summarized his and other previous findings into design guidelines. He manipulated the design dimensions of Layout Space, Association Cues, and Display Size in four user studies in order to understand the two tradeoffs: (1) between visibility and occlusion, and (2) between legibility and relative size, when designing IRVEs. The tasks he used were search and comparison, both of which contain the subtask of relating annotations to their referent. Several of his recommendations are relevant to Contextualized Video designs:

- We should guarantee visibility of the annotations and their reference, and not introduce too much occlusion when increasing association cues, because “Just enough and no more” association can improve performance. This recommendation should be considered when designing Contextualized Videos, as the Visibility-Occlusion trade-off also exists in Contextualized Videos. When embedding videos as screens into the model, occlusion is often a severe problem.
- We should choose legibility over relative position, unless we want high precision in reference-based searching and comparison tasks. For Contextualized Video tasks that involve close observation of videos, legibility is very important.
- We should keep the display location of annotations stable. In Contextualized Video designs, the videos correspond to annotations in IRVE and should be kept stable when the user is observing the videos.

Moreover, Contextualized Videos pose additional challenges and trade-offs over IRVEs with pure text annotations. For example, mapping between multiple videos and between videos and the model are not addressed in previous IRVE research. The usefulness of the subtle cues provided by videos also needs to be investigated.

## ***2.6 Non-Contextualized Abstraction and Visualization of Videos***

One strategy to support rapid decision making is to reduce the amount of information to be processed by human operators. For this purpose, researchers have been exploring automatic video content analysis technologies [Viktor, Eide et al. 2003; Forsyth and Ponce 2004; Xu 2007] to extract information from the videos. Although significant progress has been made in this direction, the high computational cost of these techniques will limit their usage in realtime situations in the near future. Also, these techniques normally require a relatively stable environment and proper lighting to work reliably. Not all potential applications can provide such a working condition.

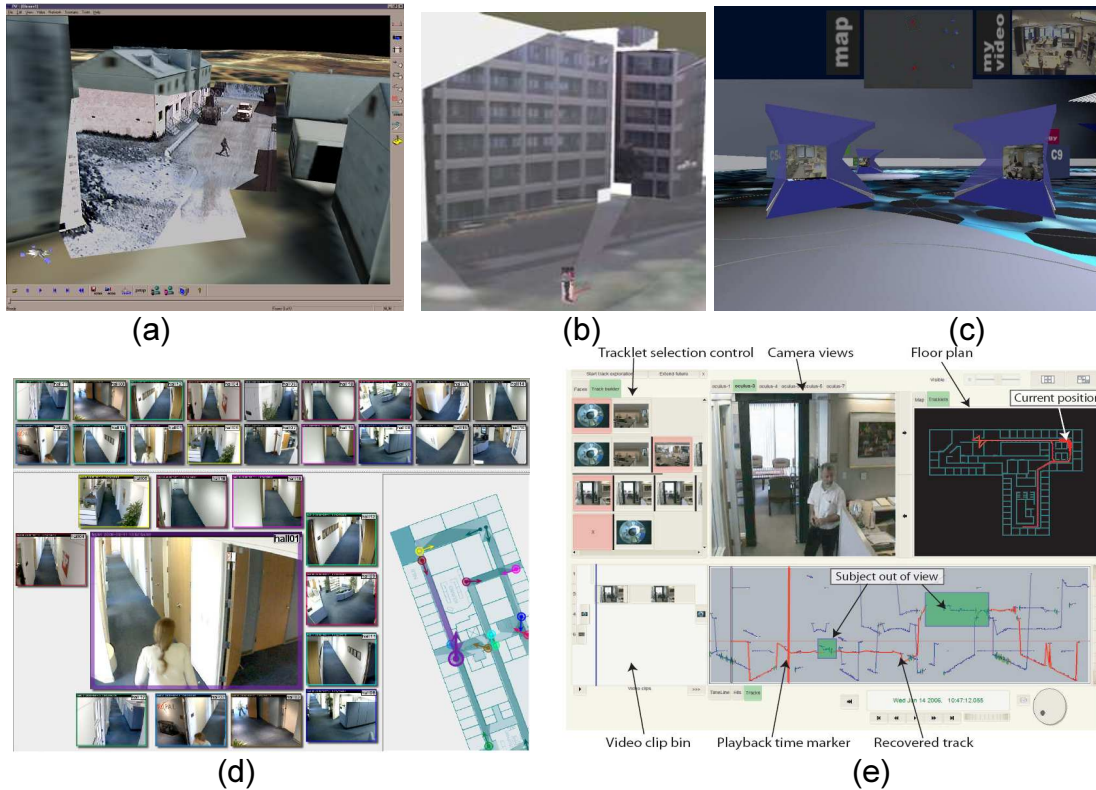
While the automatic video analysis and abstraction communities have developed techniques to track human forms and detect anomalous behaviors from video sequences, these techniques will not soon replace human operators in many application areas, for example, surveillance systems. There is still a need to present the results of these algorithms to human operators. Chen et al. proposed the concept of video visualization [Daniel and Chen 2003; Chen, Botchen et al. 2006]. They treated a video as 3D volume data and adopted a variety of volume and flow visualization techniques to summarize the activities captured by a video. They showed that people can identify the patterns in the visualization after a short period of training. Assuming the video data were already



processed by such non-contextualized abstraction and visualization algorithms, Contextualized Videos research focuses on how to present the video data in such a way that users can offload the difficulty of spatial relationship reconstruction onto the display.

## 2.7 Contextualized Video Techniques and Applications

Parts of the Contextualized Video story exist in the literature from multiple research communities.



**Figure 2-2: (a) Video Flashlight [Sawhney, Arpa et al. 2002], (b) Dynamic Imagery [Sebe, Hu et al. 2003], (c) Moving Office [Schnädelbach, Penn et al. 2006], (d) Spatial Multi-Video player [Girgensohn, Shipman et al. 2007], (e) MERL forensic surveillance system [Ivanov, Wren et al. 2007].**

### 2.7.1 Video Surveillance

Multiple Contextualized Video designs have been proposed by researchers from different domains. Sawhney et al. demonstrated the feasibility of projecting multiple videos onto a 3D environment model in their Video Flashlight work [Sawhney, Arpa et al. 2002]. Sebe et al. presented an “Augmented Virtual Environment” system which integrated multiple videos into a 3D context model [Sebe, Hu et al. 2003]. They detected moving objects inside the video and visualized them as textured dynamic rectangles moving around in the 3D model.

Girgensohn et al. used the 2D layout of videos (Spatial Multi-Video player) to show the spatial proximity of the cameras’ field of view [Girgensohn, Shipman et al. 2007]. Their user study showed that this design can improve user performance. They also found that a 2D map with camera glyphs led to a similar performance improvement. We took

advantage of Girgensohn's result and investigated whether embedding videos inside a map can further improve path reconstruction performance. We also investigated the effect of spatial context presentation. The path reconstruction task we used requires the participants to reconstruct the target individual's path.

Recently Ivanov et al. proposed an interactive prototype that allows users to browse and search the video and sensor data recorded by a surveillance system [Ivanov, Wren et al. 2007]. They used multiple views to represent both the spatial and temporal context of the videos. Since the target individual's path is mainly captured by motion sensors, human operators need to manually resolve the ambiguity in the data using a tracklet selection control. This is the most sophisticated Contextualized Video interface we have seen, but it was mainly designed for non-realtime search and data mining tasks.

Haan et al. blends videos captured by nearby cameras according to users' viewpoint to support smooth and egocentric navigation between behind-camera views [de Haan, Scheuer et al. 2009]. However, it is not clear whether users can get disoriented when rotating large angles without seeing the spatial relationship between the two cameras. To achieve a good intermediate view, the objects have to be of proper depth in the view of the two cameras. In a real world situation, the camera settings are often not ideal. Therefore, we assume less ideal camera settings when creating and evaluating techniques in this dissertation.

### **2.7.2 Computer Supported Cooperative Work**

Several teleconferencing and CSCW systems have placed live videos into collaborative virtual environments [Han and Smith 1996; Schnädelbach, Penn et al. 2006]. In [Schnädelbach, Penn et al. 2006], the realtime videos were used as windows through which users could communicate with their colleagues. A user could freely configure his spatial relationship with others by moving the video that represented him in the virtual environment. Spatial understanding was not reported as an issue, probably because the virtual space was very simple and the number of videos was small.

### **2.7.3 Web Mapping Services and Geobrowsers**

Since the early 2000s, multiple geobrowsers were released to the public. The most well-known include Google Earth (previously Keyhole Earth Viewer), NASA World Wind [NASA 2006] and Microsoft Virtual Earth [Microsoft 2008]. These geobrowsers normally provide a model of the earth that is mapped by the superimposition of images obtained from satellite imagery, aerial photography and GIS 3D globe.

Google Earth is probably the most popular one to date. The release of Google Earth caused a more than tenfold increase in media coverage on virtual globes between 2005 and 2006 [Scharl and Tochtermann 2007], driving public interest in geospatial technologies and applications. Google Earth users can use Keyhole Markup Language to integrate marks, text, 3D models, images and videos into Google Earth. Keyhole Markup Language (KML) [Google 2008c] is a file format used to display geographic data in an Earth browser such as Google Earth and Google Maps. As an XML-based schema, KML uses a tag-based structure with nested elements and attributes. Currently KML allows users to embed photos and 3D Models into Earth browsers, and put videos and photos into the descriptor Balloons associated with the Placemarks. Using Network Link mechanism, dynamic information can be obtained from the internet and updated in

Google Earth every second. For example, the latest release Google 4.3 allows users to monitor traffic speeds at loops located every 200 yards in realtime. Although embedding or mapping realtime videos directly into the object space of Google Earth and Google Maps is not yet supported, it is likely to occur in the near future with the advance of computer graphics technologies.

Within the year 2007, thousands of videos were uploaded onto Google Earth by people from all over the world. Many of them were about a famous site. The videos can be put on the 2D map and the user can have a rough idea of where the video was captured (Figure 2-3). With more precise registration of the video to the model, the observers can even see the activities in the videos in their proper locations. I will prototype and evaluate various video-model registration techniques in this dissertation.

Google Maps, Yahoo! Map and MapQuest are popular online web mapping applications for route planning and route learning. Older versions of these applications could output a route list and a line to illustrate the route on a 2D map. Although this function was shown to be very useful, it could be made better. For example, an egocentric view of the real site to better support local navigation is not yet supported. In May 2007, Google Maps and Google Earth integrated such a function, which is called Google Street View [Wikipedia]. It provides 360° panoramic egocentric views of some cities and their surrounding metropolitan areas. The users can navigate using either the arrow keys on the keyboard or by using the mouse to click on arrows displayed along the street. The user can also link from the route lists to the Street View to learn how to make a turn in a complex situation. The photos complement a verbal description by providing rich context cues, e.g., realworld landmarks and egocentric views of the road layout, for egomotion at the site. Taking it one step further, Chen et al. proposed an interface that utilizes videos to illustrate the procedure of making turns by providing an adaptive and continuous visual flow of landmarks and configuration of the space [Chen, Neubert et al. 2009]. Furthermore, the videos can be shown in the object space of Google Earth so that the user can perceive smooth transition between an overview of the route shown by the model and details on how to maneuver through complex situations shown by the videos. I prototype and evaluate these designs in the second research cycle.



Figure 2-3: Google Street View in Google Maps

All the research projects (summarized in Chapter 2.7.1, 2.7.2 and 2.7.3) plotted interesting points in a multi-dimensional design space [Wang, Krum et al. 2007], and numerous new designs may be invented in the future. None of them applied their research in more than one domain or investigated the fundamental design factors of Contextualized Videos in order to provide general design guidance. A deep and cross-domain understanding of these factors can lead to general guidelines valuable to a wide variety of applications involving Contextualized Videos, e.g., video surveillance interfaces, firefighting control consoles and teleconferencing interfaces.

Also, all of the systems mentioned so far used either a 2D map or a 3D model of the spatial environment, but none of them compared the two. Some designs, e.g., Video Flashlight [Sawhney, Arpa et al. 2002], required a 3D model to work properly and many others did not. Theoretically, both 2D maps and 3D models have their advantages and disadvantages. We directly compared them in our user study.

## 3 Current Practices of Building Surveillance

The video surveillance domain can potentially benefit greatly from Contextualized Video techniques. Therefore, this chapter describes a survey of video surveillance and a field study into the building security industry, in order to discover the potential of Contextualized Videos in the real world.

### 3.1 Building Surveillance Tasks

Security monitoring work is said to be “hours of boredom and seconds of pure terror” [State of California Employment Development Department 2002]. We summarized the typical video surveillance tasks described in previous literature. These tasks are further verified by a field study described in Chapter 3.3.

**Monitoring:** The main responsibility of a guard is to observe. They sit in front of a control panel and watch one, or up to tens of closed-circuit television screens, and monitor the key areas in the work place. Once they notice a suspicious behavior, they often zoom in the camera or enlarge the video to examine more closely. They also often follow the suspicious person from one camera to another, until they are sure there is no security issue. Based on the above observation, the monitoring task can be further divided into two sub-tasks: overview and details-on-demand. In the overview task, the security guards continuously, or very frequently, monitor the whole situation. When they notice some suspicious behavior, they want to observe it more closely and track the suspect in a detailed view.

**Patrolling:** Security guards are often required to regularly patrol around the building. They check that windows, doors, electrical systems, alarms, and sprinkler systems are working properly. A security guard often collaborates with others when performing some tasks. For example, when doing a routine check or searching for something suspicious, they often want to keep in contact with other guards. Phones are often used to communicate location/situation information for cooperation.

When security guards are alerted by the occurrence of some events, they then shift their concentration to the following tasks:

**Normal Response:** door access control; answering regular phone calls; receiving visitors, etc. Most normal alarms can wait for response or be dismissed without further action.

**Emergency Response:** fire; power; vandalism and illegal entry; answering emergency calls, etc. I want a security guard to be quickly notified about the alarm type, position, situation, etc. When an emergency happens, guards may need to report the problem to police, and perhaps inform and guide people inside the building to safe places. They may also go and physically search an area thoroughly when anything unusual is noticed.

**Maintaining security record:** Security guards may need to keep a record of security information, including names of visitors, suspicious events, a written summary at the end of each shift, etc. Either computers or pen and paper may be used for this task.

**Training:** There is a high turnover rate in the security guard job market. New employees need to get familiar with the building and the work process. It takes time to construct a mental map at the configurational knowledge level [Wickens and Hollands].

### **3.2 Video Surveillance Systems**

A typical building security & car park control system normally includes the following subsystems [BAPS 2002]:

- (1) Video surveillance system. The most widely used are Closed Circuit Video Surveillance Systems (CCTV). Some are integrated into Building Access Systems.
- (2) Intruder Alarm and Smoke Detection Systems.
- (3) Building-entry access system. It may include car park access control for entry and exit.
- (4) Elevator Access Control, restricting tenants within floors.
- (5) Secure Alarmed Areas within office complexes.
- (6) Energy Management & Building Service Control Systems (lighting and air conditioning).

More advanced video surveillance systems use Pan-Tilt-Zoom (PTZ) cameras to cover larger areas surrounding the building, especially the outside grounds. These cameras allow security officers to manually aim and zoom-in on areas of interest. The cameras can also be programmed to automatically sweep predetermined areas of interest [ITI Technologies]. Where facial recognition is vital in order to pursue legal action, special high resolution cameras with remote positioning and aiming, as well as their quantity, placement and lighting become an important cost factor. Where only observation of events is required, moderate cost systems are readily available. All camera video outputs are delivered to video cassette recorders (VCRs) or digital video recorders (DVRs) which can store video data for a period of time ranging from a few weeks to many months, for later viewing.

Currently, mainstream CCTV systems record data to VCRs (video cassette recorders), and because of its perceived ease of use and manageable price point, analog was probably the right choice when this dissertation is written. However, the rise of digital has laid bare analog's many shortcomings. Analog CCTV systems are generally maintenance intensive, offer no remote accessibility, and are notoriously difficult to integrate with other systems [Axis].

Within the past five years, IP-Surveillance solutions have emerged as an attractive alternative to the DVR, as it provides a bridge to enter the digital world with the ultimate solution of high-performance, low-cost digital video surveillance and monitoring. IP cameras or network IP cameras stream live video via digital packets across an internet protocol (IP) network such as a LAN (Local Area Network) or the Internet. The benefit of network cameras is that they reside on IP networks; the video streams can then be accessed and stored remotely. This enables users to view and/or manage the IP cam using a standard web browser or video management software from different locations, giving businesses increased flexibility [ipcamerasupply].

In summary, the future trends of video surveillance system are:

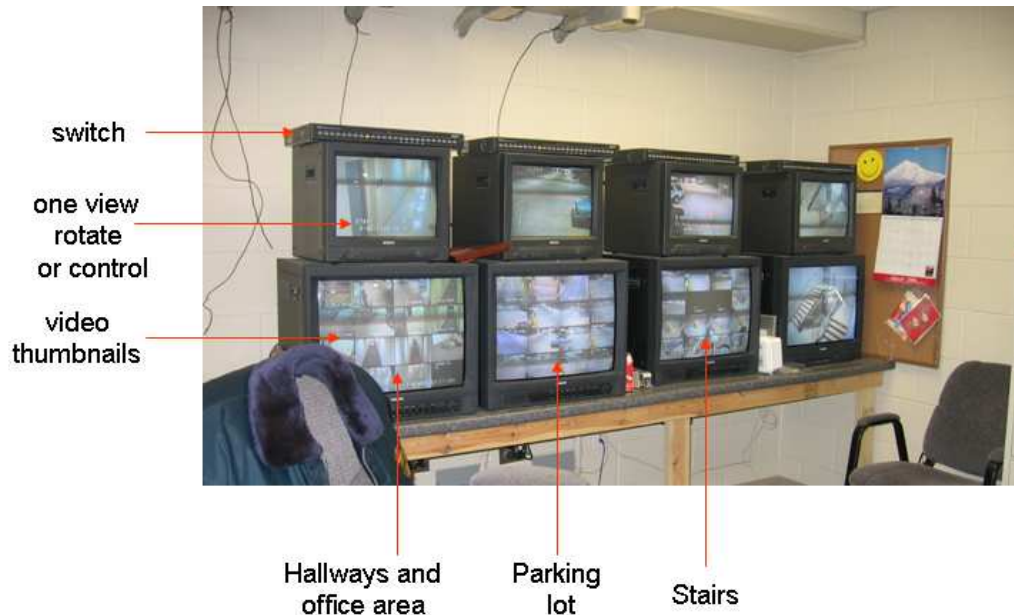
- (1) IP surveillance systems will replace analog surveillance systems.
- (2) Automatic video analysis will be more intensively used.
- (3) Video surveillance systems will be integrated with other building control and management systems.

The control interface of surveillance systems varies. Small buildings may need a single display system. Larger buildings may need multiple displays and multiple security guards.



### 3.3 Field Study

I formed a basic notion of the building security working process during the literature review phase. The field study provides some facts to support the hypothesis, so we can form a more concrete understanding of the problem domain, and identify the tasks that are likely to benefit from Contextualized Videos.



**Figure 3-1: The Building Security Surveillance System at Building A.**

During the field study, I visited a mixed-use retail, office and parking building with five floors. The whole building was monitored by one security guard through about 50 CCTV cameras. The videos were arranged as 4×4 arrays on three 23 inch monitors. Each video could be selected by a hardware switch and enlarged on a single monitor.

I arranged the field study in two sessions: a preliminary fact-gathering interview and an activity analysis session. The first session lasted for about one hour. During that time, I asked the security guard to describe his general working process and how he used the monitoring system. The second session lasted for two hours. In that session, I joined the security guard both while he was patrolling the building and monitoring the console in the office. He also provided us with further details such as how to determine the blind areas of the cameras and how to decide whether a person is suspicious or not.

The security guard reported that he took several weeks to become familiar with all the cameras to a degree that he could identify the blind areas that were not covered by any camera. This process might be even longer for novice users and low spatial ability users [Velez, Silver et al. 2005].

#### 3.3.1 Social Background

Only one full-time security guard, JB, is hired to work at night hours. JB works forty hours each week. Table 3-1 shows his working hours.

JB has been doing building security work for four years. Before that, he had served in the military for ten years. He said some security guards have military background. Before

coming to Blacksburg, JB had been working security at a large building in California. The security system there contained seventy monitors covering an area of 6 acres. Three security guards worked on three monitoring consoles.

Security guards need to attend professional classes in order to know what they can do and what they cannot do. They are not police, so they cannot put their hands on people. When they find someone behaving improperly in the building, they try to banish them from the building using non-physical methods, e.g., talking to them politely or intimidating them.

Monday 8:00pm-2:00am
Tuesday 8:00pm-2:00am
Thursday 8:00pm-2:00am
Friday 8:00pm-4:00am
Saturday 8:00pm-4:00am
Sunday 8:00pm-2:00am

**Table 3-1: Working hours of the security guard at Building A.**

### 3.3.2 Workplace Observations

The two major tasks of security guards are monitoring and patrolling. JB told me he patrols every two hours. Each patrol lasts about half an hour. He monitors the building from the office the rest of time.

#### A. Security System Overview

The monitoring system is located in a glass-walled office at the entrance of the parking lot on the ground floor. There are a total of forty-five cameras in the building. The cameras are fixed and do not have pan and zoom capabilities.

The system contains eight monitors as shown in Figure 3-1. Each of the four larger monitors at the bottom display an array of 4x4 video overviews. Each small monitor on the top displays, in sequence, the enlarged views for the small videos on the larger display underneath it. Each video is enlarged for about 5 seconds on the small monitor.

Figure 3-1 shows the security monitoring system at Building A. Any visitor to the parking lot can see the monitoring system through a glass wall. The security guard can choose to enlarge any of the 4x4 videos by pushing the corresponding button of the switch sitting on top of the small monitor.

I noticed that the videos are organized in an interesting way: one set of cameras monitor the hallways, one set monitors the parking lot and the other set monitors the stairs. The text displayed at the top of each video shows the position of the camera and the time of the video. The position information is shown in two or three words, e.g., “L1 Main Entry”, “LL Elev”, “P3 Down Ramp”, etc.

The security systems also contain a building alert system. The security guard uses it to inform the residents to leave the building. JB said the alert is very loud.

#### B. Monitoring Task

The monitoring task is less formal than I had imagined. JB glances at the monitors occasionally instead of staring at them. He usually sits at the desk and reads newspapers. He also does other things during this task, e.g., preparing some food using the microwave,



reading other guards' paper records, etc. He glances at the monitors every a few minutes. When he notices somebody or some car in the video, he switches it onto the higher resolution display and watches it closely. If a person disappears from one video, he knows exactly in which video he will appear again. JB follows the suspicious person until he enters an office.

After hours of watching, I still could not form a building model in my mind to link most of the videos. I asked JB how he managed to do it. JB felt he could get familiar with the buildings very quickly. He also knew all the blind areas in the building that the cameras cannot see. He said he figured them out in the first several days of his work at the building. In the second session, after I patrolled with JB, I found that I could link the videos together better. I believe I could form a spatial model in my mind to link the videos within one week.

I noticed that when a person was far away from the camera, or located at the boundary or the corner of the camera view frustum, I could hardly notice him/her in the thumbnail view.

### **C. Patrol Task**

JB patrols around the building every two hours. Each patrol lasts about half an hour. He patrols all the areas, whether with or without camera coverage. I observed that he showed more emphasis on areas without cameras. He also checked whether the doors were locked and whether the lights were turned off.

## **3.4 Domain Characteristics**

This section summarizes some important implications of the survey and the field study observations. These implications provide domain specific design guidelines. These domain characteristics, together with the general guidelines summarized in Chapter 5-8, should be considered during the design of Contextualized Video systems in the future.

**Domain Characteristics 1:** Security monitoring tasks are boring and lonely. Security guards prefer some peripheral relaxation, e.g., music and radio.

**Domain Characteristics 2:** When monitoring a large area, security guards prefer to have an overview of the whole situation at a glance and then choose details to observe. This strategy is consistent with the widely known information exploration strategy summarized by Shneiderman and Plaisant [Shneiderman and Plaisant 2005].

**Domain Characteristics 3:** Security guards do not stare at the videos all the time. They are likely to miss very fast actions in the video. It is better to have the cameras cover a large area and long distance, so the security guards have a greater chance of observing improper behaviors.

**Domain Characteristics 4:** Security guards are prepared for timely reactions to emergency situations. They are trained on how to properly respond.

**Domain Characteristics 5:** Security guards need to have a good situational awareness of monitored content. They should be very familiar with the spatial content, the configuration of the cameras, the social background of the residents and the companies, and the usage pattern of the monitored area. For a five-story building with forty-five cameras, experienced security guards can form a precise understanding of the cameras' placement. They learn clearly which areas are covered by cameras and which areas are

not. They learn the spatial relationship between cameras. The time needed to become familiar with such a system is about one week.

**Domain Characteristics 6:** Security guards prefer to eliminate blind areas between cameras; especially those accessible from outside of the building. They need to patrol blind areas frequently.

### **3.5 The Potential of Contextualized Videos**

Based on the above analysis, a Contextualized Video interface can potentially support the following tasks:

**Monitoring:** The new interface can provide an interactive overview of the situation in the entire building. With 3D visualization, security guards can actively look for correlations between video content and its context (location, relationship with other monitors/sensors) within the building. It helps exploration according to what is observed in the 2D video. For example, it provides detail on demand for video examination and filters for specific sensor data.

**Normal Response:** Some alarms can be visualized in the 3D model, e.g., a door is opened. The 3D model can further be used as a control panel to replace current physical button panels.

**Emergency Response:** An emergency alarm, with its location and event context can be immediately visualized in the 3D model. The security guard can easily be aware of the setting of nearby areas, e.g., which people in which offices need to be informed of the problem.

**Training:** Use the tool to help a new security guard to become familiar with the building and the relationship between the 3D model and the real building.

**Collaboration:** The 3D visualization tool can also support collaboration. The security guard sitting before the console can collaborate with other security guards that work on site, e.g., visualizing a fire area for firefighters, guiding on-site workers to an accident site.

**System configuration:** The 3D visualization tool can be used to visualize the coverage areas of cameras and sensors.

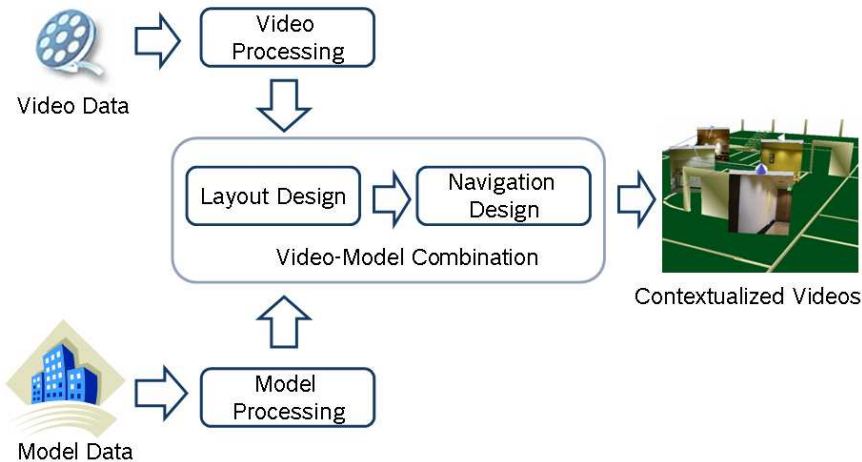
**Advertising:** The tool gives buyers of the security systems an integrated view of the whole system.

In summary, the rationale to introduce Contextualized Video interfaces is to create a unified and contiguous view, by combining live sensor data with a 3D model of the building. The model visualization can potentially (1) reduce the cognition load, the reaction time and the learning time for the monitoring task by helping security guards form a comprehensive understanding of the videos and their context. (2) Help the security guard work in a more active and comfortable way. (3) Provide an integrated interface for: video, alarm, intrusion, access control, and integration of online/offline data.

### **3.6 Summary**

In this chapter, we surveyed the video surveillance domain through literature review and a field study. This chapter, together with Chapter 5, addresses research question Q2 (For a particular domain and activity, what are the usable Contextualized Video designs and their limitations?).

## 4 The Design Space of Contextualized Videos



**Figure 4-1: The Contextualized Video design framework**

The purpose of our research is not only to find individual effective visualization designs, but also to develop general guidelines or theory that can help visualization designers understand Contextualized Videos. For the latter purpose, exploring the structure of the design space is an important step, after which we will be able to identify the primary choices for testing in the follow-up controlled experiments. This is consistent with House et al.'s argument: "controlled experiments are quite limited in their ability to uncover interrelationships among visualization parameters, and thus may not be the most useful way to develop rules-of-thumb or theory to guide the production of high-quality visualizations" [Aretz 1991].

Contextualized Videos combine video and model data to help people understand complex situations. In a Contextualized Video interface, the video provides physical, detailed and dynamic information, while the model provides structural and abstract information. Naturally, layout of the videos and the model is a key design dimension. Nonetheless, several other issues are also relevant and need to be addressed. The functional module diagram (Figure 4-1) shows the major design dimensions. From the designer's point of view, the Contextualized Video design space contains the following primary dimensions:

- Video Processing Method: how video data are processed before being combined with the model.
- Model Visualization Method: how the environment is modeled and rendered.
- Video-Model Layout Design: how to lay out videos and models together in one display, from which the observer can infer some relationship between the videos and the model, as well as between multiple videos.
- Navigation Design: how to navigate between different views of a video, between multiple videos, between a video and a view of the model, and between different views of a model.

## 4.1 Video-Model Layout Design

The video-model layout problem is a special feature of Contextualized Video visualization. In principle, either the model or the video can be in the center of the display. We focus on how to organize videos around models in this paper.

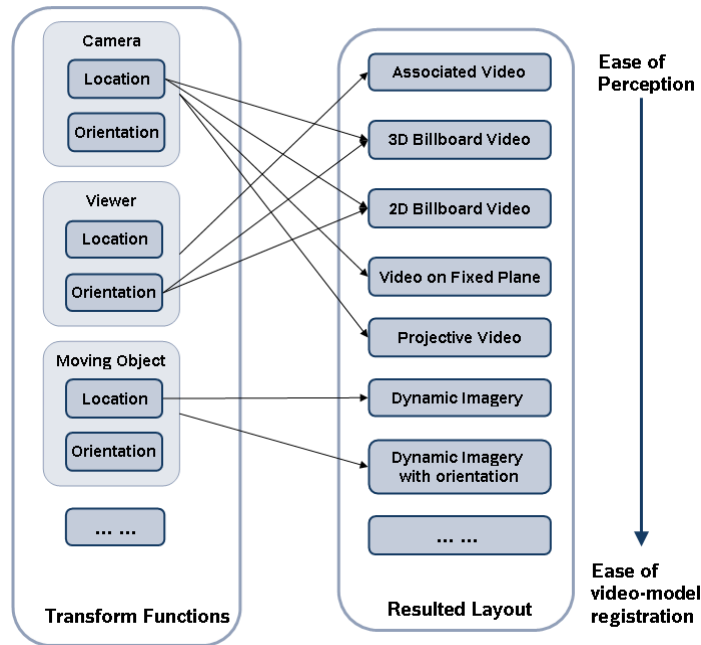
When videos and the environment model are both present in an interface, a decision must be made on how to share the display space between them. Videos and the model can be displayed in separate views, as in the Spatial Multi-Video Player interface [Girgensohn, Shipman et al. 2007]. In this case, some association cues, such as color coding and callout lines, can be used to link the video to its location in the model. We call this type of design Associated Videos. Videos and the model can also be integrated into a single view, as in Video Flashlight [Sawhney, Arpa et al. 2002]. In this case, the video content is put in the object space of the environment model. We call this type of design Embedded Videos.

The video-model layout design module can be characterized primarily as a layout matrix  $M_{layout}$  that defines how the post processed video data  $V_{post}$  are transformed and projected to form the combined visualization  $V_{aug}$ .

$$\bullet V_{aug} = M_{layout} (V_{post}) \quad (1)$$

$M_{layout}$  can be decomposed into multiple simple matrices, which determine  $V_{post}$ 's location, orientation, size and projection distortion respectively. For example, the location transformation matrix  $M_{layout}$  can be defined to follow the viewer's location, the physical video camera's location, or the location of a moving object segmented from the video. Furthermore, different simple matrices can be defined to follow different objects. Figure 4-2 illustrates some typical layout matrices and the resulting layout designs, which will be described later in this section. While it is not possible to describe all the layout designs in this paper, we analyze the ones that were prototyped in our testbed. I believe there are other promising designs not discovered in this design space.

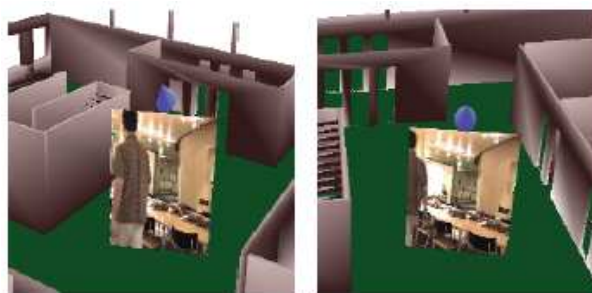
According to the spatial relationship of the video and the 3D model, we can classify video placement methods into two categories: associated videos and embedded videos.



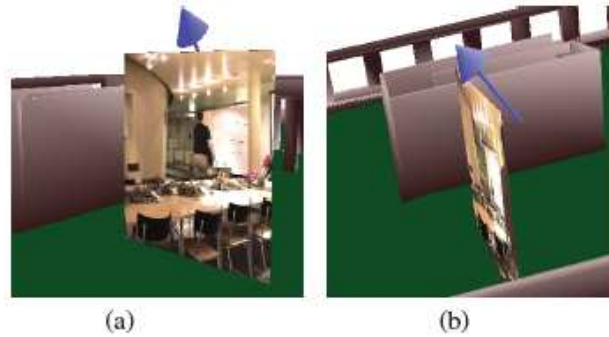
**Figure 4-2: Mapping between transform functions and the resulting video-model layout designs.**



**Figure 4-3: Associated Video. Callout lines are used to show association.**

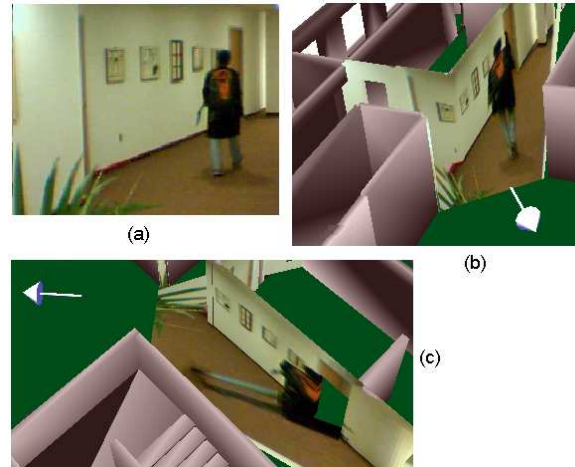


**Figure 4-4: 2D billboard Video**



**Figure 4-5: Video on Fixed-planes. The video is hard to observe in (b).**

- *Associated Videos* - As in traditional video surveillance systems, the videos are displayed as an array of thumbnails in the viewer's viewport. Its  $M_{layout}$  is defined to follow the viewer's location and orientation. Hence this layout provides excellent visibility of the video content. A major issue in associated video is how to help user relate videos to their corresponding locations. Some visual cues such as callout lines (as shown in Figure 4-3) or color coding can be used. But scalability is a major limitation. For example, as the number of callout lines increases, it gets harder for users to follow the links.
- *Embedded video* designs put the video content in the object space of the environment model. Its  $M_{layout}$  is defined to follow the physical cameras' location. Hence, embedded videos give an approximate location cue of the video. Associated video and embedded video can be used together to complement each other. Depending on how we define the orientation and projection matrix, there are a variety of designs for embedded videos.
- *Video Billboards* - This type of embedded video maps the video onto a rectangle that always orients itself to face the user (Figure 4-4). The billboard's location approximates the location of the video content. Since the orientation depends on the observer's view point, camera orientation is not apparent and video content location cannot be precisely determined. Compared with video projection and video on fixed planes, videos are easier to perceive in video billboards. The billboard can either rotate about a point in space (*3D billboard*), or about an axis (*2D billboard*).
- *Video on Fixed Planes* - This embedded video design maps the video onto a fixed rectangle (Figure 4-5). The rectangle is oriented to align with the camera's axis of projection, so it approximates the location of the content and reflects the orientation of the video camera. This technique avoids or minimizes the video distortions possible with video projection; however, it can be difficult to perceive the video information from vantage points that are far off the projection axis.



**Figure 4-6: Video Projection: (a) original video, (b) viewpoint approximately follows the video camera, (c) viewpoint far away from the video camera, revealing severe distortion and image fragmentation due to the missing door of the model.**

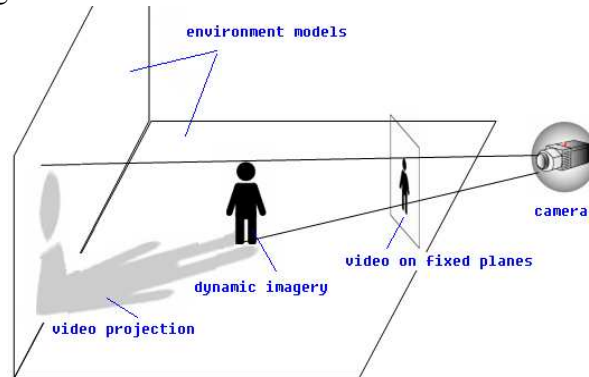
- *Video Projection* – This embedded video design projects videos onto the 3D model in the same way as a projector would (Figure 4-6). Video projection manifests the camera coverage area and camera direction on the model. If the video is texture projected onto the model from the actual camera location with the correct camera parameters, the walls and floors in the video can seamlessly match the model. However some objects can appear to be distorted if they are captured by the video but not modeled as 3D objects. Figure 4-6 b shows such a case: the human figure is distorted because the video is projected onto the wall and the floor instead of a corresponding 3D human model in the 3D space. When the projected model area contains broken walls, e.g., open doors, the projected video may be even harder to perceive and interpret because the video image is broken into multiple parts. Sawhney et al. showed how to implement video projection in [Sawhney, Arpa et al. 2002].
- *Dynamic imagery* – This embedded design maps the video or the extracted moving objects from the video onto a polygon whose movement follows the detected dynamic object’s movement in 3D space. In this design, the location and the height of the moving object would be shown precisely. However, if the whole video is mapped onto this polygon, the background of the video will be distorted. Dynamic imagery will be hard to perceive, because it often moves around when the user is observing it. Sebe et al. demonstrated an implementation of dynamic imagery in [Sebe, Hu et al. 2003].

It is interesting to note that video on fixed planes and video projections were created by the same layout matrix  $M_{layout}$ , even though they don’t look similar in appearance. They differ only in terms of what projection surface is used. Video on a fixed plane is projected onto a plane facing the camera, while video projection is projected onto the environment model. Figure 4-7 illustrates the difference between video projection, video on fixed planes and dynamic imagery.

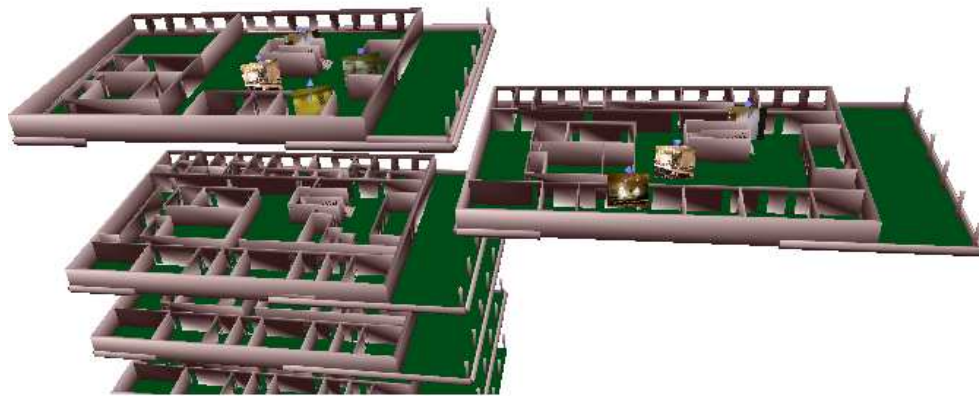
From the user’s point of view, the various video placement methods can be thought of as a continuum, balancing between ease of video perception and ease of video-model spatial alignment. On one end there are associated videos, which are very easy to



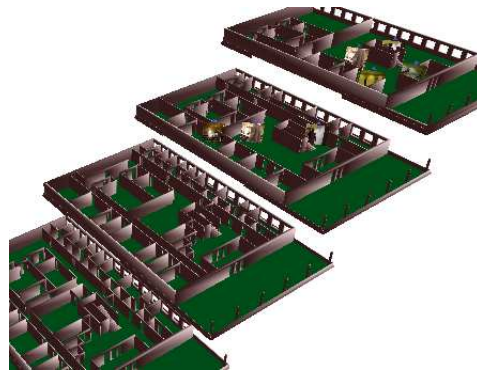
examine but need the most user effort to align with the model. On the other end, there are dynamic imagery and video projections, which are harder to examine but easier for the user to spatially register with the model. Video on fixed planes and video billboards lie between video projections and associated videos. Video on fixed planes eliminates the broken image and projection distortion problem of video projection, at the cost of more difficulty in matching the features between the video and those of the model. Video billboards further eliminate the vantage point distortion problem at the cost of more difficult orientation alignment.



**Figure 4-7: Video projection, video on fixed planes and dynamic imagery.**

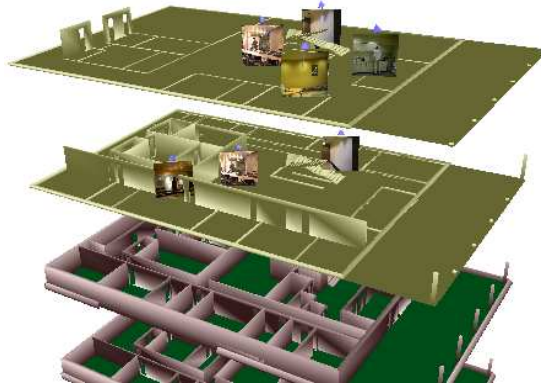


**Figure 4-8: The Drawer technique for visualizing a single floor.**

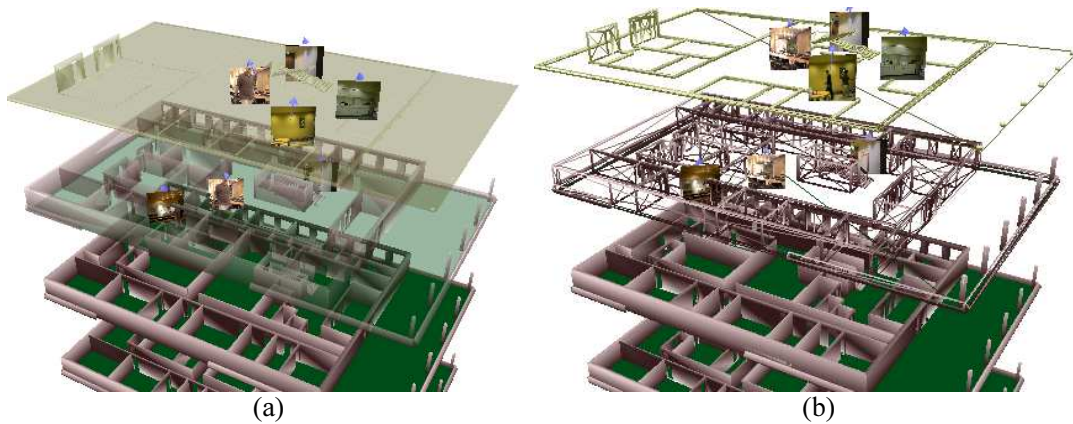


**Figure 4-9: The Rotate-and-Shear technique for visualizing all the floors.**





**Figure 4-10: Landmark technique for visualizing the structure and reducing occlusion.**



**Figure 4-11: (a) Semitransparency and (b) Wireframe for visualizing internal structure of the building.**

## 4.2 Model Visualization

In our work, an environment model describes the 3D spatial context of the videos. Complex 3D scenes present several known problems, e.g., occlusion and display clutter, which are particularly severe for embedded videos. To explore the usefulness of embedded videos, it is important to address these problems using some model visualization techniques.

Various techniques to deal with occlusion and clutter have been investigated in areas such as scientific visualization [McGuffin, Tancau et al. 2003; Elmqvist and Tsigas 2007] and engineering illustration [Martin 1989]. These techniques can be generally categorized into three strategies:

- *Explosion and deformation* – This rendering style separates subassemblies and components from the main object so that details can be seen. We briefly describe three explosion techniques that were prototyped in our testbed. Our basic implementation expands the floors vertically. We also implemented two variations on this basic implementation: *drawers* (Figure 4-8) and *rotate-and-shear* (Figure 4-9). With *drawers*, the building can be thought of as a bureau or dresser, where each floor acts as a drawer. The user can draw out a floor by selecting it. With *rotate-and-shear*, each floor can rotate along its own axis or shear out from

its neighbors like a pile of cards. The drawers variant was effective for exploring a single floor, while rotate and shear view reduced occlusion between multiple floors in one operation.

- *Cutaway* – In a cutaway, part of the 3D model is removed to show significant interior features. We implemented a simple *cutting plane* in our prototype; however, more complex geometric shapes such as spheres, ellipsoids, or arbitrary curved surfaces could be used to define the cutting boundary. We provided 4DOF (four degree of freedom) control of the cutting plane, shifting vertically or rotating along the three axis, in order to provide vertical cutaway views that can be used to reveal inter-floor features like stairways and elevators.
- *Ghosting* – Ghosting reveals the internal components by fading out less significant regions of the 3D model, such as occluding sections of the exterior skin. The distinction between cutaway and ghosting is that ghosting fades out, but does not entirely remove, the occluding parts. We implemented three ghosting techniques: landmark (Figure 4-10), wireframe and semitransparency (Figure 4-11 (a)). Each technique can be applied on a floor-by-floor basis or applied in combination on a single floor. The goal of the *landmark* view is to eliminate unimportant components while keeping the structure as a context. In the *semitransparent* view, all the components of the object are rendered in a translucent fashion, but additional depth cues, such as color, are employed to help the user perceive the 3D structure of the object. In the *wireframe* technique, only edges and vertices are displayed for the 3D model (Figure 4-11 (b)).

Among these ghosting techniques, landmark view not only reduces occlusion, but also reduces display clutter; hence the structure of one floor can be easily perceived. However, videos underneath the top floor may still be hidden. Wireframe and semitransparency are able to reveal more videos; hence the user can see an overview of all the videos in a single view. But wireframe and semitransparency may also lead to misjudgment of the video position, because they often fail to provide enough depth cues.

The above methods are mainly used to visualize the physical environment. We can visualize the cameras as well. For example, we visualized the camera's location and orientation using a very simple 3D camera model in our testbed. We could further visualize the camera's 3D coverage space using a semitransparent pyramid at some expense in clutter and occlusion.

### **4.3 Video Processing**

Video processing and computer vision are both well-developed research areas with numerous research results, many of which can be adapted to create innovative Contextualized Video designs. Sebe et al.'s "Augmented Virtual Environment" system [Sebe, Hu et al. 2003] is such an example. They detected moving objects inside the video and visualized them as textured dynamic rectangles moving around in the 3D model.

The simplest case is no video processing, as in our current implementation. The next possibility is to do video content analysis on the video streams and highlight the changes, as well as recognize objects like humans inside the 3D model. For instance, visualizing the video signatures [Chen, Botchen et al. 2006] inside the 3D model may be a promising idea. Furthermore, when the models do not provide enough details, we can derive additional 3D details from the videos to refine the model.

While the video processing and computer vision communities have developed techniques to track human forms and detect anomalous behaviors from video sequences, these techniques will not soon fully replace human operators in all application areas because of the reliability and the high computational cost of these techniques. They can be used to extract certain information, e.g., motion, from the videos and reduce the amount of information to be processed by human operators. There is still a need to present the results of these algorithms to human operators. Chen et al. proposed the concept of video visualization [Daniel and Chen 2003; Chen, Botchen et al. 2006]. They treated a video as 3D volume data and adopted a variety of volume and flow visualization techniques to summarize the activities captured by a video. They showed that people can identify the patterns in the visualization with a short period of training.

Contextualized videos focus on how to present multiple videos in such a way that users can offload the difficulty of spatial relationship reconstruction onto the display in realtime. Reliable and realtime video processing algorithms have the potential to greatly improve the usability of Contextualized Video interfaces. Such algorithms do not need to be perfect. The intermediate result can be presented to human operators through Contextualized Video interfaces. Some subtle cues that are hard to identify by computers can be easily captured by humans, who can provide their feedback to the video processing algorithm through the interface. In this way, humans and computers can collaborate more closely to perform the real tasks through Contextualized Video interfaces.

## **4.4 Navigation**

Interaction, particularly navigation, is a primary component of Contextualized Video interfaces. Although the high level goal of navigation can be searching, exploration or maneuvering [Bowman 1999], the low-level goal of navigation is always to change one's view of the space in order to do some tasks in a better view. For example, users navigate to find uncluttered and unobstructed views to examine multiple videos, to easily map objects captured by the video with their representation in the spatial context, and to gain an understanding of the building structure.

The realtime and opportunistic feature of video surveillance tasks pose special challenges for navigation in Contextualized Videos:

1. **Realtime:** Users often navigate to follow the progress of a realtime event, which is often unpredictable and happens quickly. Navigation must be done in realtime, which requires Contextualized Video designers to reduce the time and cognition cost of their navigation techniques.

2. **Multi-tasking:** Navigation is often performed in parallel with other tasks. Users may want to keep track of a video while navigating to look at the model or another video. Therefore, we should be careful to maintain the working context of the users when navigating to a new view.

### **4.4.1 Navigation Context**

Navigation can be performed in overview or in detailed view. The well known Visual Information Seeking Mantra of “overview first, zoom and filter, then drills-on-demand” [North and Shneiderman 1997] is an example of navigation in overview. Moreover,

navigation in overview can be used to get a detailed view, and vice-versa; the world-in-miniature technique for virtual world travel is an example [Stoakley, Conway et al. 1995].

The difference between **detailed view** and **overview** is the amount of data that is contained in the view. An overview shows more data with less detail. Detailed view and overview are relative.

For tasks that require detailed information from video and structural information from a model (see Category 3 tasks in the task taxonomy), we need to support both views in the interface. An unclear question is which view supports “systematically better” navigation, which is not necessarily faster or easier by itself, but can lead to higher task performance and lower mental workload when integrated into the working flow.

"Overview first, then drill down" is a well known strategy to acquire information from a data space. But in the video surveillance domain, the task often starts from a detailed view, where a target is detected in a video. Also the two views (video/context) are often very different; the switching cost between the two views is expensive. In which view should navigation take place? In other words, should we use "navigate in detail, then zoom out to overview" or "navigate in overview, then zoom in to detail"? When is each strategy helpful? These questions motivated me to compare the two types of navigation strategies.

It is necessary to explain two other concepts, *egocentric view* and *exocentric view*, which are often confused with detailed view and overview. Egocentric view and exocentric view are our mental representation of the physical space. Unlike detailed view and overview, egocentric view and exocentric view should not be used to describe external displays. They are not distinguished by the amount of data they contain, but by the relevance to the viewer's body. Egocentric view, or first-person view, is based on our self-centered needs. Compared with exocentric view, egocentric view usually contains more details, but can be overview as well. They are often used in virtual environments where a large part of the human body is involved in the interaction (including travel, selection and manipulation). I choose to use the term “detailed view” and “overview” in this experiment because of two reasons. First, our experiment mainly uses mouse and keyboard for interaction, and does not include embodied interaction techniques. Secondly, the user mainly focuses on information discovery and problem solving from a third-person view, not on a low-level interaction with the environment.

#### **4.4.2 Navigation Mode**

Many Contextualized Video tasks are highly time-critical – they must be done while the video are playing. This poses a challenge to navigation technique, which usually comes with certain time and cognition cost. Semi-automatic navigation techniques were often proposed to reduce the cognition cost [Tan, Robertson et al. 2001; Wijk and Nuij 2003; Elmqvist and Tsigas 2007; de Haan, Scheuer et al. 2009]. However, it is unclear whether they are really better than traditional manual navigation for Contextualized Video because of multiple reasons: (1) the time cost of automatic navigation is not necessarily low, (2) the disorientation problem of automatic navigation, and (3) limited destination. Therefore, it would be helpful for future designers if we can compare semi-automatic navigation with manual navigation techniques for Contextualized Video tasks and provide design guidelines based on the findings.

**Semi-automatic navigation:** The user specifies the destination view but has no control on the transition between the current view and the destination view. The transition is computed according to a predefined algorithm. Semi-automatic navigation relieves users from route planning, but it is still not easy to specify a destination of arbitrary location and orientation.

**Manual navigation:** The user controls the process of navigating from one view to another. Compared with semi-automatic navigation, manual navigation requires more attention resources, but it allows more flexible control and improvisational navigation.

The key difference between semi-automatic and manual navigation is whether the user has control of intermediate views and whether the user knows the destination beforehand. (Semi-automatic navigation can also reach any location in the space as in the World-In-Miniature technique [Stoakley, Conway et al. 1995], so location constraint is not the major difference). For video surveillance applications, both navigation methods have their pros and cons. Manual navigation allows the user to choose views improvisationally and work in a smaller action-evaluation cycle. Users may do better in keeping themselves oriented. They may also be faster when the goal is highly dynamic and can change before a predefined path finishes. But when users have a clear designation in mind, semi-automatic navigation allows the user to concentrate on higher-level tasks, and hence save the user's time and effort. However, in some situations, specifying the destination might be hard.

## **4.5 Task Taxonomy**

With the design space, we can describe various Contextualized Video techniques from multiple viewpoints. Techniques are designed for tasks. In order to analyze and evaluate these techniques in a systematic way, we also need a taxonomy of tasks that is abstract enough to cover most realworld usage cases, and at the same time have enough detail and specificity to be useful for both designers who want to improve their application interfaces, and evaluators who want to compare different visualizations.

The evaluation result, either analytical or empirical, described with this task taxonomy should be able to help future designers to select the appropriate Contextualized Video design according to the tasks at hand.

### **4.5.1 Existing Task Taxonomies**

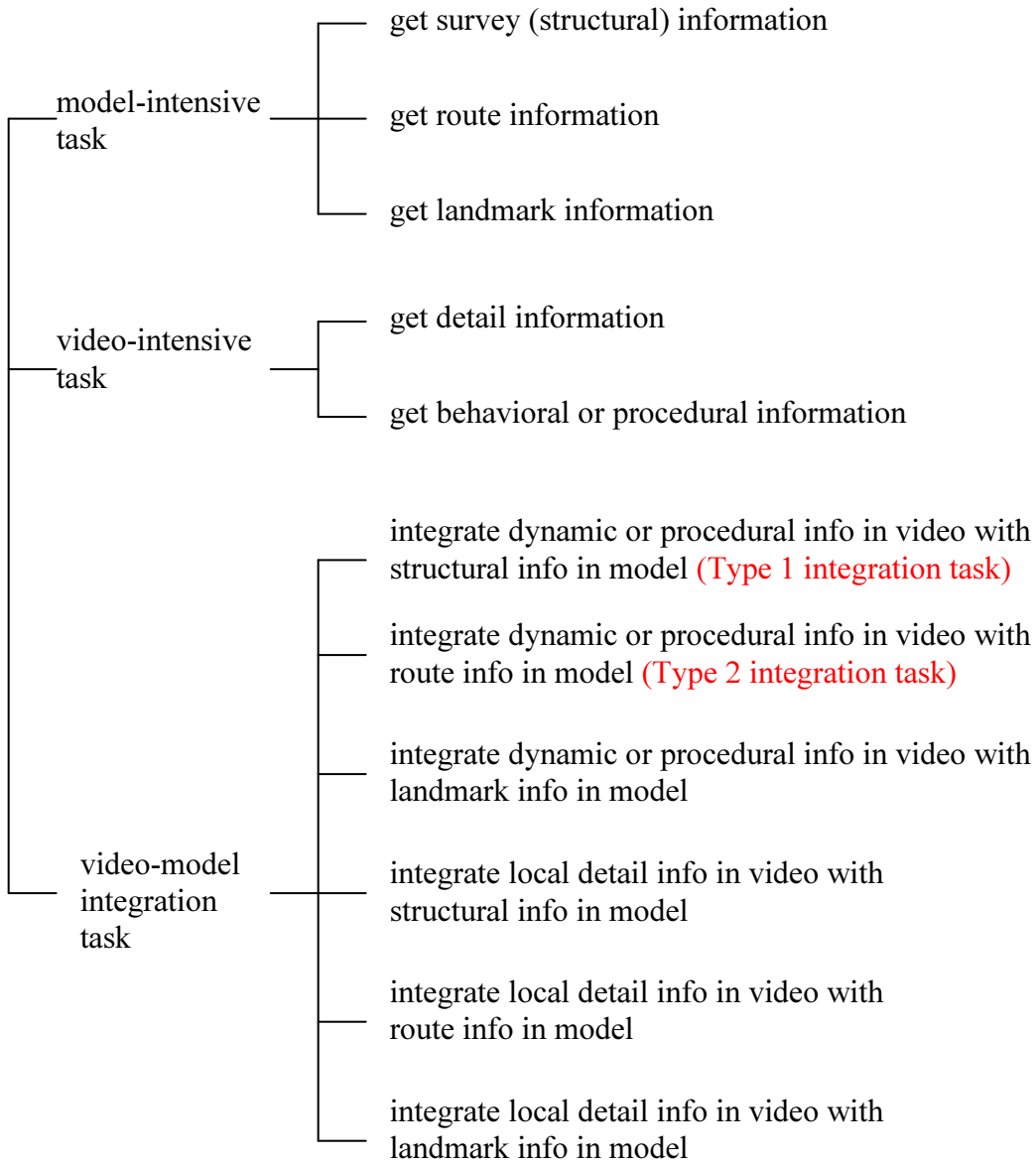
As we have noted, task taxonomies have been proposed in multiple related fields. For Virtual Environments, Bowman categorized the low-level interaction tasks into selection, manipulation, navigation and system control [Bowman, Kruijff et al. 2004]. Munro et al. gave a taxonomy of knowledge types that can be conveyed by Virtual Environment (VE) presentations, including location knowledge, structural knowledge, behavioral knowledge and procedural knowledge [Munro, Breaux et al. 2002]. We can thereby categorize the cognitive processing tasks according to what type of knowledge or information the user wants to extract from VEs.

Historically, visualization tasks were categorized into two major types: scientific visualization and information visualization, based on whether the data is physically based or abstract. For information visualization, Shneiderman enumerated a set of generic information visualization tasks based on a user's information requirements when performing a complex activity [Shneiderman 1996]. Recently, Tory and Moller proposed

a visualization taxonomy considering more of the user's conceptual model of the visualization [Tory and Moller 2003].

However, these task taxonomies are too general to capture the distinctive tasks that the users can perform on Contextualized Videos.

#### 4.5.2 Contextualized Videos Task Taxonomy



**Figure 4-12: The Contextualized Videos task taxonomy.**

The taxonomies must come from a deep and thorough understanding of the realworld task and the techniques that have been proposed for it. Therefore, some initial informal evaluation of techniques and/or design of new techniques for the task is almost always required before a useful taxonomy can be constructed. This Contextualized Videos task taxonomy is based on the literature survey (Chapter 2 and 3.1), the field study of a video

surveillance application (Section 3.3), and my initial design and evaluation experiences (Chapter 5). I also considered tasks from other domains to avoid holes in the taxonomy.

While the design space should be described from the designer's point of view, the task taxonomy should be described from the user's point of view. I decomposed the realworld activities into basic tasks and classified them according to two usage criteria of the users:

### **1. Which part of the visualization does the user focus on?**

Depending on their task goal, users can either focus on the video content, the environment context (i.e. the model), or the relationship between the video and the model.

### **2. What kind of information does the user want to extract from the data source?**

When focusing on a particular part of the visualization, users may want to extract different kinds of information to support different task goals.

The taxonomy is shown in Figure 4-12. The taxonomy clearly shows the distinctive tasks of Contextualized Video interfaces. Each task category is described with an example:

Category 1 – Model intensive tasks (requiring contextual information, which is mainly available from the 3D model):

Model intensive tasks can be further categorized into three types according to the kind of information that the user wants to extract. The user can extract structural information for the purpose of spatial understanding. The structural information can range from a local area to the whole model. The user can also extract route information for the purpose of navigation, or extract landmark information in order to use it as a reference.

- Get structural information: learn the structure of a site.  
Example: understand the high level design of the Forbidden City in Beijing.
- Get route information: plan or learn a route.  
Example: learn a route to go to a store from the entrance of the mall.
- Get landmark information: learn the approximate shape of an object or a local site.  
Example: learn the shape of a particular building on a campus map.

Category 2 – Video intensive tasks (requiring information presented in videos):

From the video, the user can extract either behavioral or procedural information, or the detailed appearance of an object or local site.

- Get behavioral or procedural information: learn a detailed procedure or observe the dynamic information that normally does not appear on the model.  
Example: learn how to maneuver through a complex situation according to the videos captured by a fire-fighter in a fire-fighting situation.
- Get detailed information: learn the detailed appearance of a site from the video camera's viewpoint.  
Example: closely observe the material used to decorate the walls of a building.

Category 3 – Video-model integration tasks:

This category of tasks requires information from both the model and the videos. The user has to relate information from one source to the other. Sometimes they need to do precise mapping between the corresponding information from the video and the model. Because three types of information can be extracted from the model and two types of information can be extracted from the video, there are in total six possible combinations, each of which defines an integration task. The most interesting tasks are described below in detail.

- Integrate dynamic information from the video with structural information from the model: mapping the location and orientation of the dynamic object from the video to the model. I call this task *Type 1 integration task* for convenience.  
Example: in the video surveillance application, the user often needs to tell the target person's location and orientation on the model when he walks outside the video camera's range.
- Integrate dynamic information from the video with route information from the model: registering the dynamic information to a particular point along the route. I call this task *Type 2 integration task* for convenience.  
Example: when learning a route, the user may want to integrate the egocentric knowledge on how to maneuver through a complex road system with the exocentric knowledge of the whole route.
- Integrate detailed information from the video with structural information from the model:  
Example: if the user wants to know the appearance of a building at a particular location on the map, he needs to find the proper video that captures such a feature according to the structural information.  
Category 3 tasks are rooted in the relationship of the videos to the spatial context. This category of tasks does not appear in pure video or pure model based interfaces. Therefore, the distinctive task of Contextualized Videos is information integration between the video and the model.

The above tasks are general tasks across domains. For example, Type 1 integration tasks not only appear in the video surveillance domain, but also appear in tele-collaboration and virtual tourism applications. In a tele-collaboration situation where one person in the control room is guiding another who is navigating through a wild area, the person in control may need to judge the exact location and orientation of the other person according to the map and the video captured by the person in the wild.

A complex activity is often composed of multiple tasks. For example, the path reconstruction activity contains the following tasks:

- Scan multiple videos to detect the target person. This task is video intensive and is about extracting dynamic information from the videos.
- Mentally map the target's location in the video to his/her location in the spatial context. This is a Type 1 video-model integration task.
- Connect the target's appearance in multiple sequential videos into a continuous path. This task is model intensive and is about getting route knowledge from the model.

Even the same type of tasks can have different time, precision and workload requirement in different situations. Taking the Type 1 integration task as an example, the



precision of the mapped orientation is important when tracking a person in an open area where the target can move in any direction. But when the target is travelling along a hallway, only two directions (forward and backward) are important for the observer.

### **4.5.3 Design and Evaluation Using the Task Taxonomy**

This task taxonomy highlights the distinctive tasks that can benefit from Contextualized Videos, and guides systematic evaluation of Contextualized Video designs.

Researchers can compare multiple designs using tasks selected from the taxonomy. The findings can be stated in terms of tasks. These guidelines not only provide design recommendations for given tasks but also provide a framework that guides the designers to systematically analyze the tasks and understand the tradeoffs between design choices.

Application designers can find example designs by specifying the task characteristics, user characteristics and data characteristics of their application. These examples should be viewed as starting points of the design process instead of final designs. They should be viewed as visual components instead of a complete interface. Also, the design guidelines are by no means complete. Instead, they provide an initial version that other researchers can expand and improve.

Application designers can also analyze their tasks according to the taxonomy, and then look for the appropriate design guidelines according to the tasks. Therefore, the task taxonomy works as a communication protocol for the transfer of findings between researchers and designers.

## **4.6 Summary**

This Chapter described the four major design dimensions within the design space of Contextualized Videos, which addressed research question Q1 (What is the ontology of the design space of Contextualized Videos? What are the major design dimensions?). This chapter also formalized a subspace composed of the model visualization and the video-model layout dimension.

Based on the literature review in Chapter 2, the domain survey in Chapter 3, and the design and evaluation experience described in Chapter 5, Section 4.5 further proposed a Contextualized Video task taxonomy, which addressed research question Q3 (What are the distinctive tasks in Contextualized Video interfaces, and how can we classify them in a way that is useful to designers?).

## **5 Model Visualization and Video-Model Layout**

Since the design space is very large, instead of trying to cover the whole design space, during the first research cycle, I focused on an important subspace which is composed of two major design dimensions: the video-model layout dimension and the model visualization dimension.

### **5.1 Design Tradeoffs**

#### **5.1.1 Video-model Layout**

Embedding videos inside the model reduces the distance between the videos and their context. According to the Proximity Compatibility Principle [Wickens 1995], complex tasks that require integration of video and model information are likely to benefit from this design, because the attentional demands are reduced.

However, Embedded Videos have two major issues. One is the view distortion problem: the video may be distorted when projected onto the model or observed from a viewpoint far from the central axis of the camera's view direction. The other is the occlusion problem: the video and the model may occlude each other.

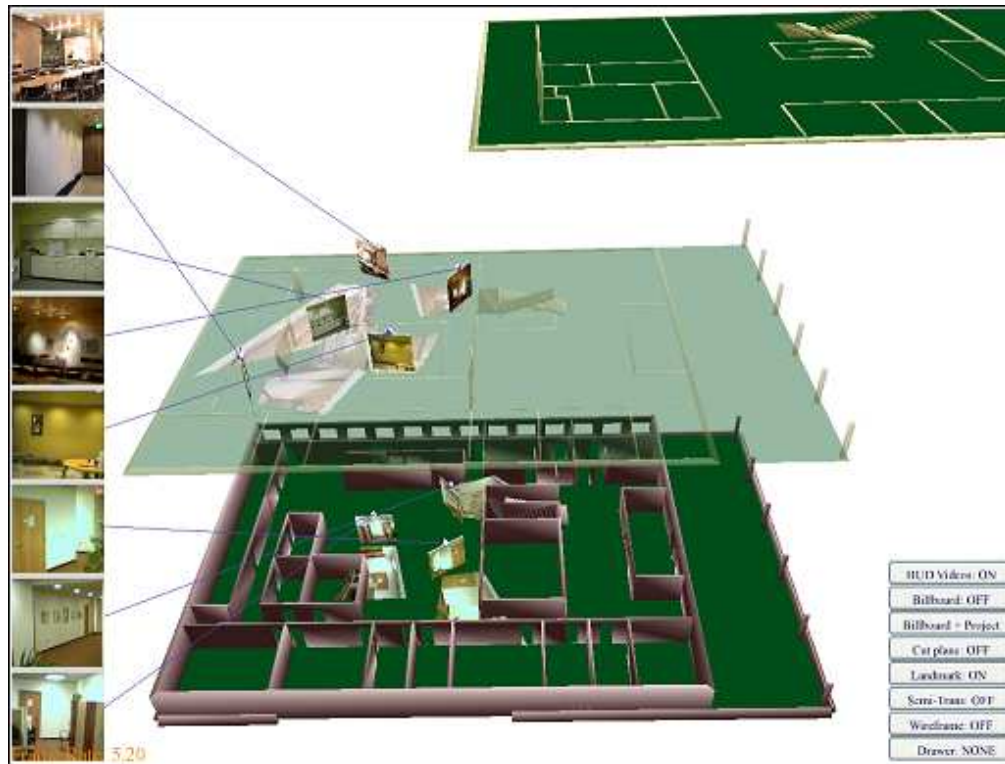
Thus, a controlled experiment is needed to understand the tradeoffs: are Embedded Videos still beneficial given the disadvantages? Can users understand a complex situation in real time? Can divided attention errors [Wickens and Hollands 2000] be reduced in a multi-tasking situation? These questions are interesting, because attentional resources are in high demand during time-critical situations. Security guards often need to perform multiple tasks in real time. For example, when a security guard notices a suspicious person, he may need to report the target's location to his colleagues and keep monitoring the situation simultaneously.

#### **5.1.2 Model Visualization**

The environment model is an abstraction of the real world and can be presented in different ways. A 2D floor plan as shown in Figure 5-3 provides a clear overview of the environment, while a 3D model of the floor as shown in Figure 5-6 allows dynamic observation from different viewpoints and both exocentric and egocentric views. With an angled view, it provides more depth cues. The disadvantages of the 3D representation are occlusion and display clutter, which can be minimized by existing visualization techniques [McGuffin, Tancau et al. 2003; Elmqvist and Tsigas 2007].

Previous research on task-view dependency did not lead to a clear answer on which visualization to choose. According to the Gestalt similarity principle [Wertheimer 1923; Ellis 1938; Chandler 1997; Ware 2000] and Rule of Consistency [Aretz 1991], 3D models have the potential to facilitate mapping between the video and the model, because when observed from the camera's viewpoint (an egocentric view) the 3D model's geometry features match those captured by the video. The matching features can help communicate the location and orientation of the person captured by the video. However, it has also been shown that tasks involving spatial understanding favor more exocentric viewpoints like a 2D map or a top-down overview of the 3D model, while tasks involving navigation and tracking favor more egocentric views [Wickens and Hollands 2000].

Since the path reconstruction task involves both tracking and spatial understanding subtasks, the tradeoffs can be better understood through an empirical study.



**Figure 5-1: The testbed for preliminary evaluation**

## **5.2 Exploratory Study**

We used a testbed method to explore this design space. A testbed allows rapid composition of solutions for different design dimensions into specific configurations that can be tested and compared. Figure 5-1 shows an overview of the testbed. The testbed is mainly implemented using OpenSceneGraph [Baldonado, Woodruff et al.]. The user can enable and disable each technique by menu selection and hot keys. Some techniques, e.g., landmark view and drawers, are applied on a floor by floor basis via mouse selection. The rest, like explosion and rotate-and-shear, are applied on the whole model.

Since the design space is huge and the interaction between multiple design concerns is complex, it would be very complex to run a fully-controlled experiment to compare all of these designs initially. Rather, we chose to run a loosely controlled exploratory study, allowing users to explore the design space and make comments on various combinations of techniques, with the goal of identifying specific hypotheses that we could later test more formally.

To allow users to freely select viewpoints in the testbed, I employed a trackball-like navigation technique (similar to [Shoemake 1992]) allowing users to rotate the model and zoom in or out. With proper model visualization and proper viewpoint, the occlusion effect can be reduced and the advantages of different video-model layout methods can be demonstrated. Besides these techniques, we also implemented several visualization

features that the user may choose to utilize, e.g., the 3D camera models that represent the cameras' location and orientation in the 3D building model.

Eleven users completed the study. For each task, the users created a variety of interesting and reasonable visualizations. Analyzing these visualizations, we discovered some common usage patterns, some of which involved two or more design dimensions, indicating that in some cases a video-model layout method needs proper model visualization support to show its advantages. These usage patterns helped us identify a limited number of promising designs to evaluate in the formal study.

## 5.2.1 Usage Patterns

### Video Monitoring Task

This task is a video intensive overview task. To support this task, users would create a visualization that put all the videos in one display so that they could monitor the whole situation of the building.

The following patterns were found:

**Pattern 1: Associated videos only**

**Pattern 2: 2D Billboard + semitransparency or landmark**

Not surprisingly, associated videos received higher preference than embedded videos among most users. However, two users preferred embedded videos to associated videos (Figure 5-2 a and b). Both users used 2D billboard. The common reason they gave was that the associated videos were arranged in a vertical line and the users had to move their eyes up and down frequently to scan all the videos. By manipulating the models, the users could arrange the videos in a smaller screen space while keeping similar resolution as associated videos, even though the videos were not neatly aligned.

No users selected fixed plane video or video projection for this task because the videos' orientations were fixed in object space and the users could not find a single view to see all the videos clearly.

### Tracking Task

This task requires the user to match the video and its nearby environment in the model. In the designed scenario, the users were asked to tell us the suspicious person's location and orientation. For this task the users would create a visualization that showed the details of a particular video, as well as the environment near this video. It is interesting to see the diverse strategies people used to figure out the orientation of the suspicious person in the model:

**Pattern 3: Associated Video + Fixed Plane Video + semitransparency or landmark**

**Pattern 4: Dynamically switching between Billboard video and Fixed Plane Video + semitransparency or landmark**

In Patterns 3 and 4, people used fixed plane videos to judge the suspicious person's position and orientation in the model. Some of these people turned to associated videos to closely observe the suspicious person (Figure 5-2 c) while others dynamically switched between billboard video and fixed plane video.

### **Pattern 5: Billboard videos + navigate to look behind the camera + 3D walls on (no landmark or wireframe)**

Pattern 5 was used because the video content's orientation matches the model's view when looking behind the camera (Figure 5-2 d). Users preferred to see the 3D walls, which were utilized to perform feature matching between the video and the model.

One user turned on video projection to judge the camera orientation and used billboard video to view the video details.

### **Route Planning Task**

This task is a model intensive task. It requires the user to have an overview of the model and a remote view of a particular video. The resolution of the video was less important. In the designed scenario, the users would plan a route starting from the 1st floor to catch the suspicious person in a particular video on the 2nd floor. The following usage patterns were found:

### **Pattern 6: Video on Fixed Plane or Video Projection + 3D walls with higher view angle**

### **Pattern 7: Video on Fixed Plane + landmark**

Video on fixed plane and video projection were selected because the approximate orientation information could be easily observed from some distance. Half of the users felt more comfortable seeing the 3D floors with walls on (Figure 5-2 e); while others preferred to do the tasks with a landmark view. This difference might be related to the user's mental model of the environment. Higher pitch angles were often selected when the walls were shown; because the users wanted to reduce the walls' occlusion in order to quickly see the route.

When creating designs, we try to understand them in terms of their primary features. After evaluating them, we try to understand the result in terms of design features. For example, embedded video is a design feature of video billboard and projective video. It provides proximity between video and its location and eases linking between the two. A primary feature comes inherently with the technique. A technique with a good primary feature has the potential to be effective in real use when the usability issues are solved.

## **5.2.2 Result Summary**

Summarizing the users' designs and rationales, we found that:

- Embedded video was preferred for video-model relation tasks while associated video was preferred for suspect detection tasks. Even for video intensive tasks, if occlusion can be effectively reduced, embedded video can still be a reasonable choice.
- All five video-model layout methods were employed by some users. This fact indicates that each method has its advantages and disadvantages, confirming our analysis in section 5.1. While video projection [Sawhney, Arpa et al.] and dynamic imagery [Gailing, Lindberg et al.] were useful for some tasks, they may not be an ideal solution for all tasks.
- Many people used a strategy that either combined multiple video-model layout methods or dynamically switched between them. This highlighted the requirement

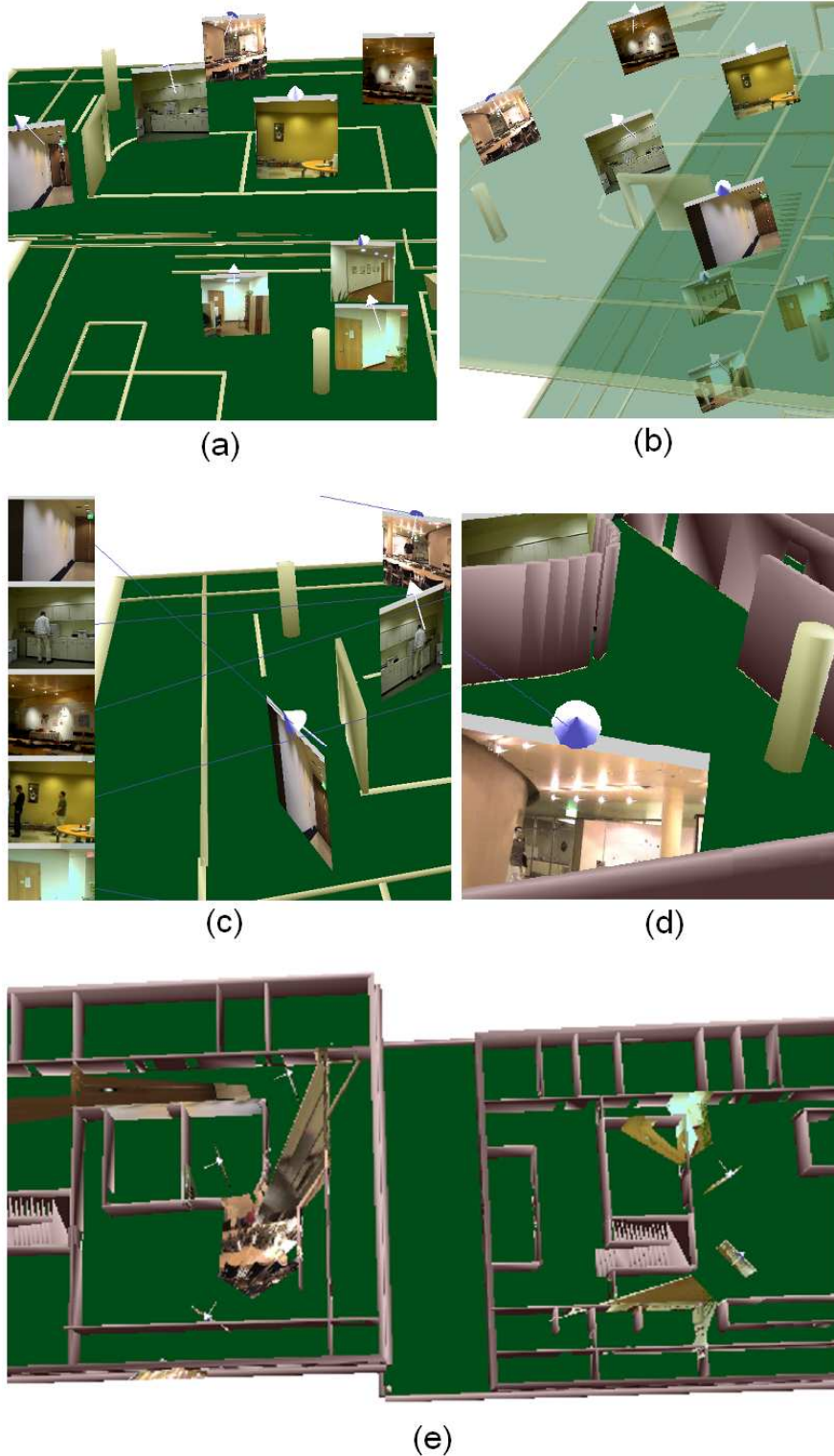
for effective interaction support.

Based on our experience, an often-effective design is to combine associated videos with embedded videos. In this way, the user can choose to use the proper representation when performing different tasks. When monitoring the whole building, the user relies more on the associated videos. When he detects a suspicious person, the user can track the suspicious person using the embedded videos.

This preliminary evaluation suggests that, despite some occlusion problems, embedded video is especially helpful for tasks where users need to consider a larger spatial context around the videos. Based on the usage patterns we identified, the following hypotheses were tested in our follow-on formal evaluation:

- Compared with associated video, embedded video can improve task performance for video-model relation tasks.
- Combining embedded and associated video will result in a balanced design that performs well for all types of tasks.
- Billboard videos, together with good 3D visualization support and an appropriate viewpoint, can achieve similar performance as associated videos for video intensive tasks.

The promising designs and design patterns were evaluated through a formal experiment.



**Figure 5-2: Usage Patterns found in Experiment 1: (a) Pattern 2, 2D billboard + landmark view + explosion. (b) Pattern 2, Billboard + semi transparency. (c) Pattern 3, associated video + fixed plane video + landmark. (d) A view behind the camera used in Pattern 5. (e) Pattern 6, video projection + walls.**

### 5.3 Formal Study

In the design space exploration phase, I have characterized several design factors common among all the designs [Wang, Krum et al. 2007]. A deep understanding of these factors can lead to general guidelines valuable to a wide variety of applications involving Contextualized Videos, e.g., video surveillance interfaces, firefighting control consoles and teleconferencing interfaces.

Through a formal evaluation, I investigated two important design factors: video placement method and spatial context presentation method. Existing visualization guidelines do not fully address the tradeoffs involved in these designs, so a user study was needed to understand the tradeoffs. A path reconstruction task was used in the study. I chose this complex task because it requires the participants to synthesize information from multiple videos, as well as the context information, in a limited time. Also, decisions have to be made when ambiguity or uncertainty exists.

Effective design should be compatible with the needs and capabilities of users. To have a comprehensive understanding of the effect of our technology on different populations, my study involved two groups of participants with differing levels of knowledge of the real environment visualized in our system.

In summary, I investigated the following research questions in this experiment:

- What type of video placement (embedded within an environment model or externally associated with a model) leads to better performance with path reconstruction tasks?
- To what extent do 2D and 3D models affect user path reconstruction performance?
- Does the spatial information help participants with little spatial knowledge to achieve a level of performance that is comparable to that of participants who are familiar with the environment?

I compared a total of six Contextualized Video designs and provide general design suggestions based on the study results.

The experiment included three independent variables: video placement, spatial context presentation, and users' level of spatial knowledge of the environment.

For video placement, three levels were selected: pure associated design, pure embedded design, and combined design. The associated design used callout lines as well as color coding to link videos to their contexts, as shown in Figure 5-3 (2D Associated Design) and Figure 5-6 (3D Associated Design). As mentioned in the related work section, there are multiple ways to embed videos into the 3D model. Since the purpose of this study was to understand the tradeoffs between basic factors, we chose to use a simple design as shown in Figure 5-4 and Figure 5-7. In the embedded design, the video planes' size was larger than the actual projection plane in order to make them easier to observe. For the same reason, the 3D camera glyphs were placed higher than their actual height. Embedded designs also allowed users to see the spatial context occluded by the videos by pressing the space key, and switch between camera-aligned video and user-aligned video by pressing the "ALT" key. A *camera-aligned* video is oriented to face the camera so that the orientation of the person and objects in the video can be easily judged. A *user-aligned* video (or billboard) is oriented toward the user to facilitate viewing. Figure 5-4 and Figure 5-7 compare the two orientation methods. The combined design showed both associated and embedded videos (Figure 5-5 and Figure 5-8).



The second independent variable was model presentation method, with two levels: 2D model representation and 3D model representation. The 2D design was a simple floor plan of the building. The 3D design used an angled view of the model, whose walls and doors were rendered in 3D with shading effects so that the floor and walls could be easily differentiated.

Table 5-1 lists the six conditions created by combining the first two independent variables. We varied video placement and model presentation between subjects because it was difficult to create multiple path reconstruction tasks with the same level of difficulty.

The 3D Embedded design needed some time to learn because the user had to differentiate two cases (as shown in Figure 5-9) before mentally mapping the scene from the video to the model. In the first case, the observer and the camera are on the same side of the video plane. Mental mapping required very little effort in this case. In the second case, the observer and the camera are on opposite sides of the video plane. Here, mental mapping was difficult because a mental rotation larger than 90 degrees was needed. We used a “mirrored video” metaphor to help reduce the mental effort. The video could be observed on both sides of the video plane. When viewed from the opposite side as the camera, the video screen should be understood as a large mirror. The scene observed on the video screen is the reflection of what is happening in the model.

Generally, the video size in the embedded condition was smaller than the associated condition, because two factors constrained the video size: (1) all seven videos and their nearby context had to be shown in one display, and (2) videos should not overlap. Moreover, the videos in the 2D Embedded condition were smaller than those in 3D Embedded condition, because the angled view in the 3D condition allowed a more compact layout of the spatial context. Thus, small video size is an inherent disadvantage of the 2D Embedded design.

To overcome the occlusion and distortion problems of 3D Embedded Videos, an autorotation function was added to the 3D Combined condition, resulting in a *3D Autorotation design*, where the system animates the model to align the viewpoint with the camera if the user clicks a video. In this way, the user can see the structure of the nearby environment and nearby videos with one click. The user could click either the associated videos or the videos embedded in the spatial context. We tuned the length of the animation to 3 seconds, which was an acceptable speed for most participants. The 3D Autorotation design took advantage of the fact that the 3D model allows more flexible views.

The third independent variable was users’ familiarity with the real environment simulated in our system. The *resident* participants had a high level of spatial knowledge, acquired through tacit and informal learning, by working in the building for at least 12 months. *Non-resident* participants only acquired their knowledge of the building by looking at the floor plan during a guided tour of the building just prior to the experiment. The non-resident participants were used to simulate new security guards who have just started to work in the building. The boredom of this job leads to a high turn-over rate, and new workers may not be familiar with the environment. We hypothesized that Contextualized Videos could help.

		Spatial Context Presentation	
		2D Floor Plan	3D Model
Video Placement	associated	2D Associated Design	3D Associated Design
	embedded	2D Embedded Design	3D Embedded Design
	combined	2D Combined Design	3D Autorotation Design

**Table 5-1: Six designs compared in the experiment**

### 5.3.1 Tasks

The path reconstruction task is designed to simulate a real world case involving situation understanding and decision making under time pressure. We simulated a suspicious person tracking situation, where the user not only needs to follow the suspicious person through multiple videos, but also needs to mark on the model the suspicious person's path by clicking a series of dots with a mouse (as shown in Figure 5-3).

Each user did four path reconstruction trials using one of the six visualization designs. In each trial, the target person walked across the building and appeared in 2-4 videos. A total of seven videos were used for each trial and the rest were distractions, where only distracting actors walked in and out. A real surveillance system often contains more videos. But automatic video analysis technologies will be able to significantly reduce the number of ambiguous videos that need to be closely observed by human operators.

The video footage was shot beforehand at different locations in the real building. To eliminate learning effects, the camera configurations were totally different in each of the four trials.

The length of each scene used in the trials varied from 45 seconds to 63 seconds. Each scene had 4-6 actors walking around, but only one of them was the target to be tracked.

The path reconstruction task contained the following subtasks:

- Scan multiple videos to detect the target person.
- Discriminate the target from distracters.
- Mentally map the target's location in the video to his/her location in the spatial context.
- Predict the next video(s) in which the target is likely to appear and observe those videos.
- Connect the target's appearance in multiple videos into a continuous path.

Multi-tasking between subtasks was possible.

### 5.3.2 Procedure

A total of 36 participants, ages 18 to 50, participated in the experiment. Half were classified as residents and half as non-residents. Color-blind participants were excluded from the subject pool and no preference was given to gender, ethnicity, or national origin. The 18 resident participants had worked inside the building for at least one year. The other 18 non-resident subjects had never been inside the building before. Each group was further divided into six subgroups of 3 participants each to evaluate one of the six designs

in Table 3-1. The resident group contained 8 females and the non-resident group contained 5 females. We controlled the gender distribution between subgroups so that each subgroup of 3 participants contained at most 2 females.

Prior to the experiment, each non-resident user was given a printout of the building's floor plan marked with several waypoints and then led on a short tour of the actual site, which was a floor of our research building. During the tour, we did not direct participants' attention. The only thing we pointed out for the participants were the physical locations of the waypoints labelled on the floor plan. Next, each participant was given a pre-questionnaire followed by the ETS standard Cube Comparisons Test to evaluate the participants' mental rotation ability, which was likely to have a significant impact on overall performance.

The participant was then seated at a desktop computer running the testbed software. During the training session, we first demonstrated how to use the visualization to track people. Then the participants were given a familiarization trial to learn to use it effectively. We showed participants the correct answer as well as how much precision was required for indicating the path. We asked the participants questions to make sure they correctly understood the designs. The entire training session was controlled to last no longer than 15 minutes.

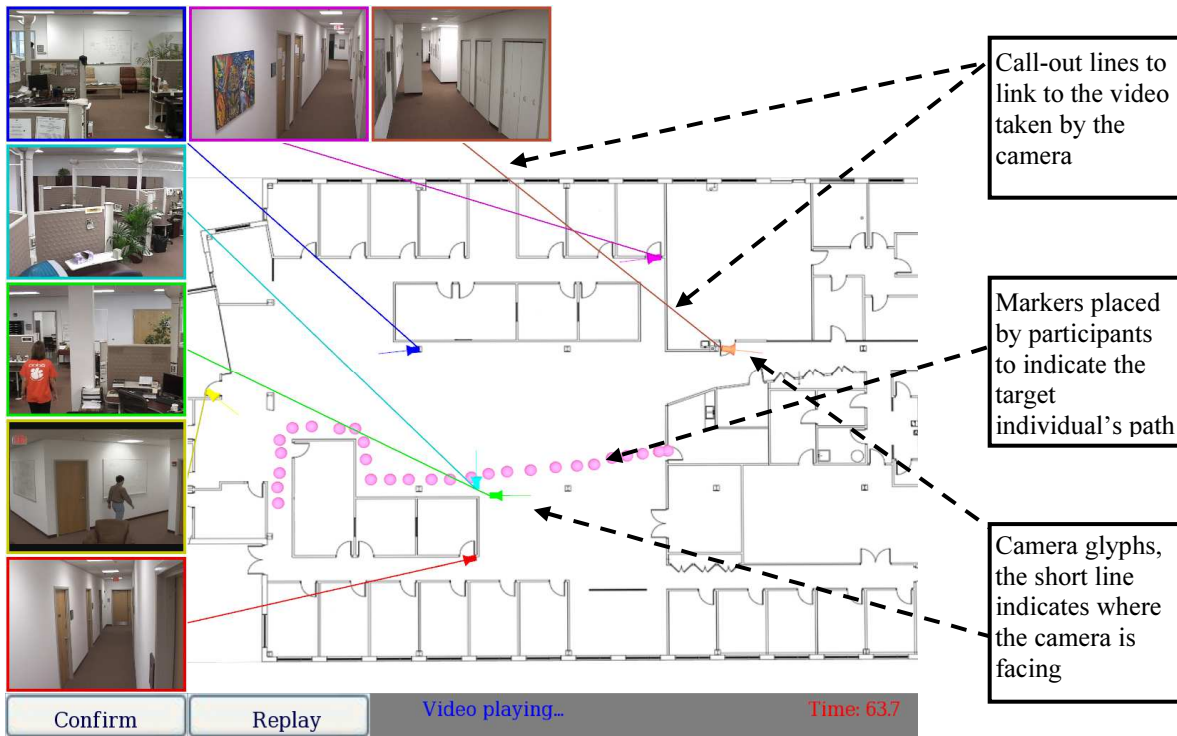
Following the training session, the participants completed four experimental trials. We showed the participant a picture of the target individual before each trial began. The participant then observed the videos, followed the target individual, and indicated the path taken by the target by making a series of mouse clicks on the model. The participant could complete this task either while the videos were playing or after they ended. The task time and path was automatically recorded by the testbed. The investigators manually noted the participants' behavior and strategy.

During each trial, the participants were able to replay the video clips once they ended, but were not allowed to pause the videos. However, they could click the model and specify the location and orientation of the target at any time. During the pilot study, we found that some scenes were hard to understand in a single viewing. Thus, we allowed replay to give the users another opportunity to finish the task, but we did not want the user to pause and replay at any time, because this would lead to too much variation in user strategy.

The participants were instructed to complete the task as quickly as possible as long as they felt the result was correct. A maximum of three minutes was allowed to perform the task. In most cases, participants finished the tasks before the time ran out.

On completion of their experimental trials, the participants were interviewed using two oral questions. The first question asked the participants to explain the path of the target actor in the final trial they performed. The second question asked the participants to give a brief account of any landmarks or cues that assisted them in performing the tasks and to what extent these landmarks or cues assisted them.

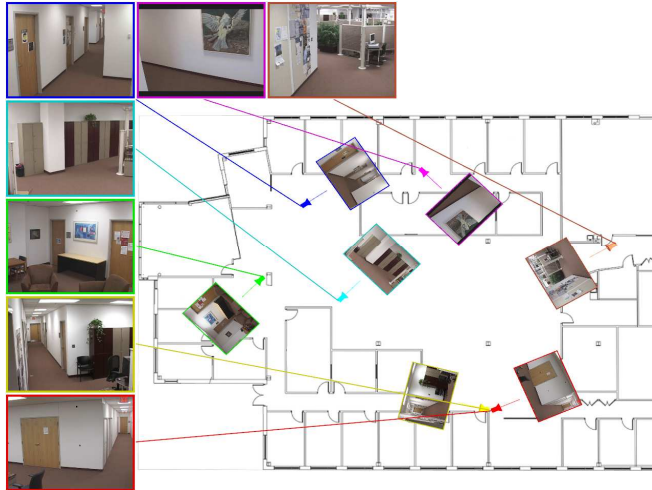
Finally, we asked participants' to rate (on a five-point scale) the design they used with regard to ease of learning, ease of use, enjoyment of use and the usefulness of the visualization in a real video surveillance system. A self-reported mental workload evaluation was administered last using simplified SWAT [Baddeley 2000]. The participants also commented as to which parts of the tasks seemed to cause the highest mental effort.



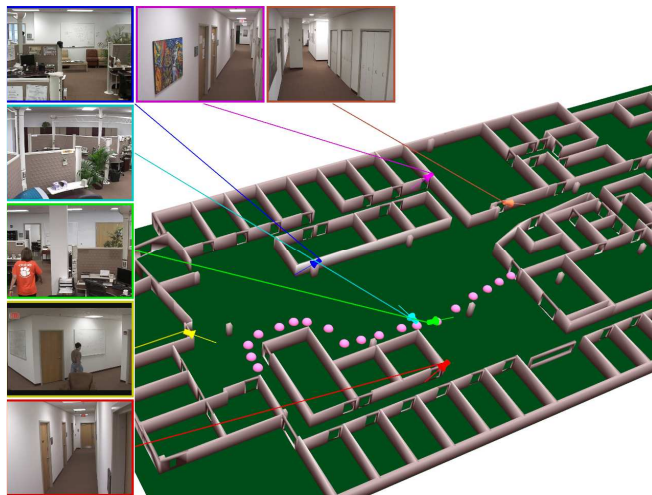
**Figure 5-3: Testbed Interface using the 2D Associated design.** In the 2D Associated design, the camera glyphs indicated the camera’s location in the building model and the short line on the camera glyph indicated where the camera was facing. The testbed provided a “Replay” button allowing the user to replay the video multiple times, and a “Confirm” button allowing the user to confirm the path of the target actor and end the session. The elapsed time since the start of the task was shown in the lower-right corner.



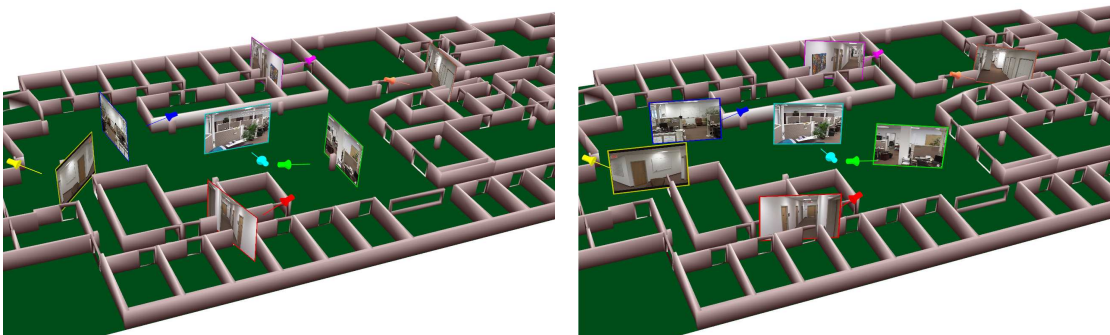
**Figure 5-4: 2D Embedded Design.** Left: By default videos were placed to face the camera on a 2D plane. In this camera-facing view, the video content was harder to observe, but mapping the actor’s location from the video to the model was easier. Right: When participants pressed the “ALT” key the videos were rotated to an upright position.



**Figure 5-5: 2D Combined Design that integrated 2D Embedded and Associated Videos.**



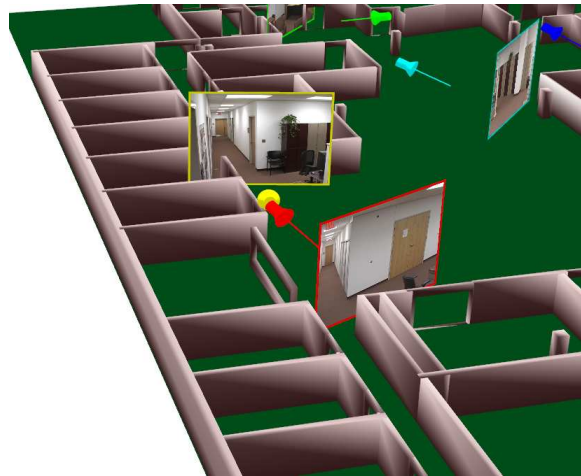
**Figure 5-6: 3D Associated Design**



**Figure 5-7: 3D Embedded Design. The embedded videos were enlarged to make them easier to observe. Left: By default videos were placed to face the camera. Right: When participants pressed the “ALT” key the videos were rotated to face the user.**



**Figure 5-8: 3D Combined (Autorotation) Design, embedded videos were enlarged to make them easier to observe. Here, the user has clicked on the yellow video.**



**Figure 5-9: The two different observation cases in the 3D Embedded and 3D Autorotation Designs. Since the user is behind the yellow camera, the yellow video shows exactly what the camera sees. But since the user is on the opposite side of the red video from the red camera, the video canvas should be understood as a mirror reflecting what is happening in the model.**

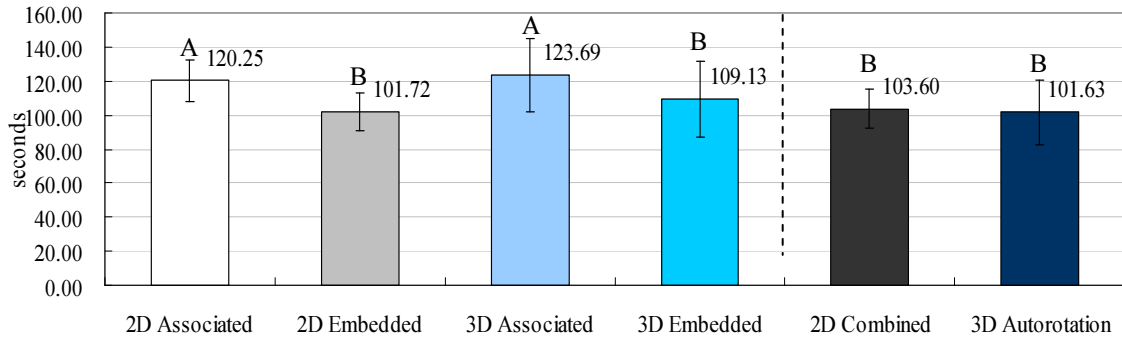
### 5.3.3 Results and Discussion

We measured user task time and the precision of the path as performance criteria. The task performance reflected how precisely and quickly people can map the position and orientation of the target person seen in the videos onto the real environment, as well as how well they can understand the scene as a whole across multiple videos.

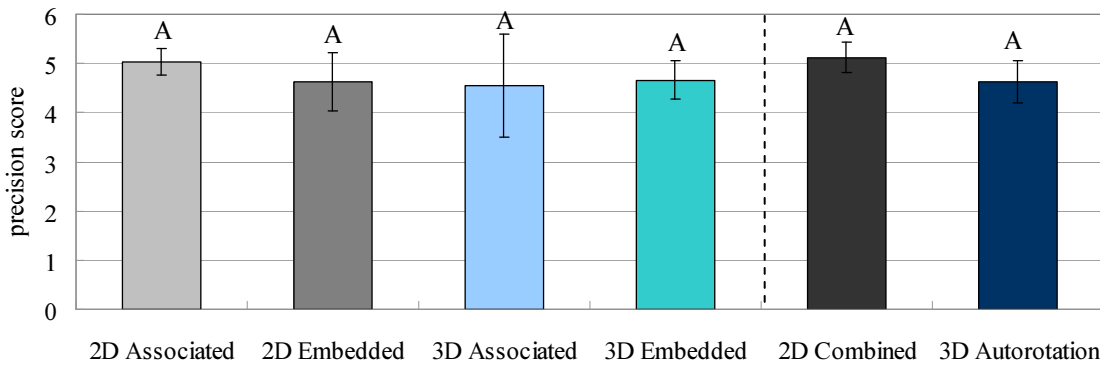
We graded the precision of the path on a six-point scale, with six indicating the most precise path. Since we could not guarantee the four tasks were of equal difficulty, we normalized the task time before taking the average as that participant's task time. To normalize the task time, we first calculate a normalization factor for each of the four tasks by dividing a constant with the mean task time across all the participants. Then we multiply each individual's task time with these factors. After the normalization, the average task time was equal for all the four tasks. We performed a three-way ANOVA on



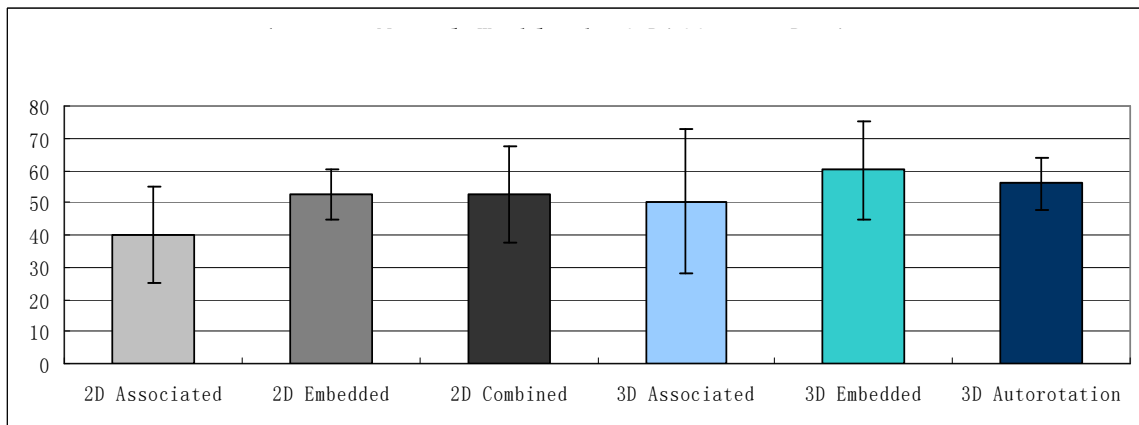
the participants' performance data. Chi-Square tests were used to analyze the subjective rating of the designs.



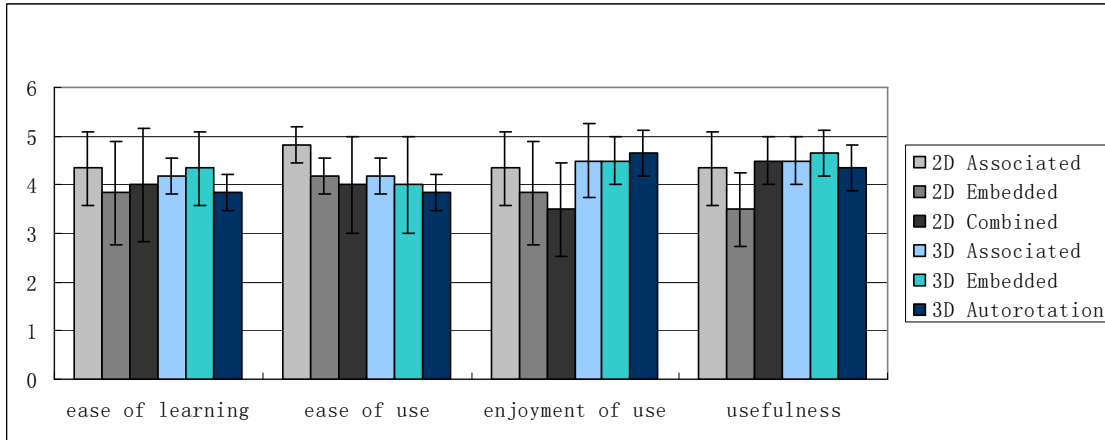
**Figure 5-10: Comparing Task Time among designs.**



**Figure 5-11: Comparing Task Precision among designs, higher score means higher precision.**



**Figure 5-12: Subjective Mental Workload Score of different designs**



**Figure 5-13: Subjective rating of the designs, with 5 being the highest score.**

### Effect of Video Placement

Video Placement had a significant effect on task time ( $p < .05$ ) as shown in Figure 5-10, but no significant effect on task precision as shown in Figure 5-11. A *post hoc* analysis using Tukey's HSD test showed that the embedded and combined conditions were significantly faster than the associated condition, but there was no significant difference between the embedded and combined conditions.

This difference can be explained by the participants' behavior. A correlation analysis showed that the task time was significantly correlated with the user strategy when performing the tasks ( $p < .05$ ). A Chi-Square test showed that the user strategy was significantly different between the associated and embedded videos ( $p < .05$ ). Nine out of 12 participants using embedded designs, and ten out of 12 using combined designs, indicated the path in real time while the video was playing the first time, but only four out of 12 participants using associated designs did so. The rest of the participants observed the videos once and indicated the path on the model after the videos ended. We believe embedded videos reduced the time needed to link a video to its model context, and therefore better supported realtime tracking.

The advantage of the realtime strategy is the user's more precise recall of the target's location when marking it on the model. Otherwise, users have to rely on memory or watch the videos again. However, the realtime strategy also requires the participants to map the target's position to the model in real time. We observed several cases when a user using embedded designs missed an important video because he/she spent too much time on mapping the other video to the model. Nevertheless, precision with the embedded designs was not statistically different than precision with the associated designs.

The self-reported mental workload with embedded designs was significantly higher than that with associated designs as shown in Figure 5-12. This was true for all three components of mental workload: mental effort load, time load and psychological stress. We think this is because the participants using embedded designs performed multiple subtasks simultaneously. The small video size and non-orthogonal view angles in the embedded design may have also increased the mental effort of discriminating a person from a distracter, as was pointed out by several participants. None of the participants using associated video reported video observation to be a problem. A third contribution to increased mental workload might be the more complex interfaces of the embedded



designs. We observed that a few participants forgot to use some of the functions (e.g., pressing the space key to see the 3D model occluded by the video) while doing the tasks.

Chi-Square tests did not show significant differences between the participants' subjective ratings of the video placement designs (Figure 5-13). This might be due to the small sample size in our study.

### **Effect of Spatial Context Presentation**

Spatial Context Presentation did not significantly affect task time and precision, although 3D designs trended toward lower precision ( $p=.13$ ).

According to participants' subjective ratings, 3D designs were slightly harder to use. 3D designs also had a relatively high mental workload, because the interface seemed more complex than 2D designs. 3D presentation of the environment might have helped in landmark identification and mapping. However, the walls in the 3D view sometimes occluded part of the path, making it more difficult to estimate and mark the path. Also, the video planes in 3D Embedded designs were not always perpendicular to the user's view.

3D Embedded and 3D Autorotation designs resulted in the highest mental workload and were rated the hardest to use. According to user feedback and our observation, the hardest part of using the 3D Embedded design was understanding the actors' real direction in the model when the observer was looking from the opposite direction of the camera's view. In this case, the scene in the video was mirrored. Although we explained this to the participants and let them practice before the real task, this still led to confusion for some participants. The hardest part of 3D Autorotation was to keep track of the person in the video while the model was rotating. Overall, it seems that the 3D designs we used have more negative effects than positive.

Despite the disadvantages, 3D visualizations received a significantly higher score than 2D visualization ( $p<.05$  for non-zero correlation) on the "enjoyment of use" question. This indicates that although the 3D Embedded design and 3D Autorotation design were hard to use, people were still happy to use them, at least for the first hour or so. Participants also rated 3D visualization more "useful" than 2D visualization ( $p=.10$  for non-zero correlation).

### **Interactions of Design Factors**

No significant interaction was found between the video placement and spatial context presentation factors. Thus, the two factors were relatively orthogonal.

However, *post hoc* analysis revealed one interesting local trend. 2D Associated Videos generally outscored 2D Embedded Videos in all four subjective ratings. This means that although participants performed better in the 2D Embedded condition, they were not as satisfied with the interface. Participants were not comfortable with the unaligned video orientation in the 2D Embedded condition. Also, the 2D Embedded condition had the smallest videos among all the six designs. Several participants reported that the small videos in 2D Embedded condition were hard to observe.

### **Effect of Participants' Spatial Knowledge**

On average, resident participants did a little better on both time and precision metrics, but the difference was not significant ( $p=.12$  for time and  $p=.27$  for precision). This indicates that non-resident participants benefit greatly from Contextualized Videos,

because they would not even be able to perform the task without a visualization of the spatial context.

With 3D visualizations, the difference between resident and non-resident participants was greater, in terms of both task time and precision. The advantage for residents was less obvious in 2D visualizations. Resident participants also showed more advantage, in terms of both task time and precision, in combined designs. This implies that resident participants can utilize a more abundant set of landmarks about the building. Thus, we should provide more cues for resident users, instead of oversimplifying the designs.

This result is consistent with the participants' subjective ratings of "ease of learning" and "ease of use." Non-resident participants clearly preferred 2D visualizations, while resident participants showed a weaker preference. This might be caused by the different forms of spatial knowledge in the minds of the two groups. Non-resident participants held a 2D map in hand when they toured around the site, thus the spatial information was likely to be stored in 2D in their minds. Resident participants were not trained with the 2D map before doing the tasks. They already had a comprehensive representation of the site in their minds due to their experience of working in the building. Thus, they could easily map any external form of spatial information into their mental model of the building.

It is interesting to see that the self-reported workload between the two user groups was very similar. Since the resident participants were more familiar with the building, their mental workload should have been lower. While a difference might be observed with a larger number of participants, nevertheless we think that the mental workload similarity between the two groups indicates that both groups were mainly relying on the external context presentation instead of the one in their minds. We observed several resident participants who wanted to look at each video for some time before each task began. They said "I am trying to understand where is where." In other words, they tried to register the camera location and orientation with their mental map. On the contrary, none of the non-resident participants tried to do so.

## **Interview Results**

In order to understand the difference between the two user groups' spatial memory, we interviewed them after the experiment. The post-task interviews were then analyzed to elicit their spatial knowledge.

The interview consisted of two questions for both types of participants. The first question asked the user to orally describe the path that the target individual took in the last task. The second question asked the user to enumerate what kinds of cues were used in reconstructing the path.

The interviews were recorded and transcribed. Following the approaches described in [Psathas 1995], we counted the number of words in answering the questions. The word-count analysis found that the resident participants had more to say in answering both questions. Residents used 24.6% more words than non-residents for question 1 and 18.6% more for question 2. It seemed that resident participants would take time and give a detailed account of the path, as it would require information retrieval from long-term memory and processing at a conscious level, which otherwise would be done sub-consciously in their daily routine.

Following the word-count analysis, we then analyzed the key words in the answers to determine the participants' frames of reference and landmark references when describing

the path. The frame of reference can be either the virtual model or the real world. The landmark references can be either physical attributes, e.g., “the large open area” or “the red camera”, or socio-functional attributes, e.g., “David’s office” or “the kitchen”. While answering the question regarding the actor’s path, resident participants referenced the real world more often than non-residents. Residents also referred to more landmarks. While the number of physical landmarks referenced by the two groups was similar, resident participants referred to many more socio-functional landmarks.

### **Other Findings**

Although the designs were early prototypes, all the designs were rated useful for the path reconstruction tasks (a usefulness rating of five means the most useful, while one means not useful at all).

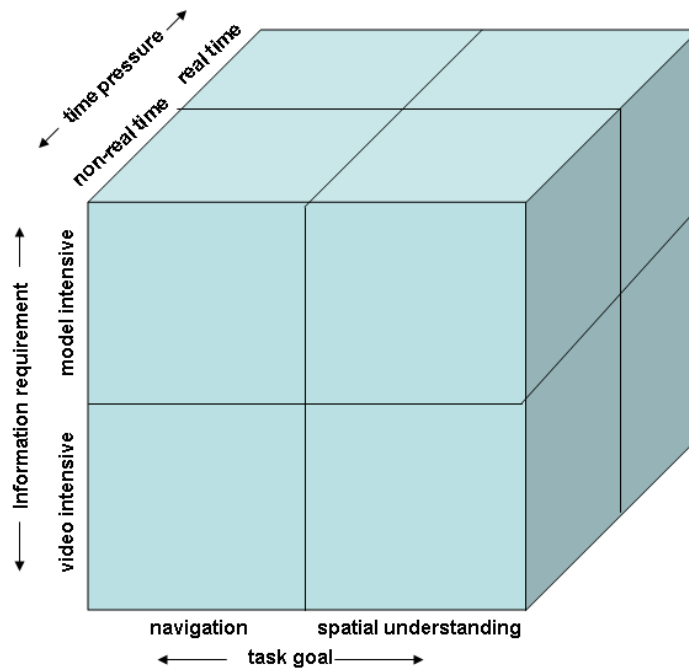
We also found that participants’ spatial rotation ability correlated with task time ( $p=.08$ ), but there was no obvious correlation with precision. The precision of the path reconstruction task does not solely rely on participants’ spatial rotation ability to map individual videos to the model. Even if participants failed to map correctly, they could still estimate a reasonable path between the videos, according to the target’s sequence of appearance in the videos. Unfortunately, we could not tell when each strategy was used in the experiment. Gender did not show a significant correlation with task performance.

## **5.4 Guidelines**

Summarizing our first cycle of investigation of the Contextualized Video design space, I propose multiple initial design guidelines in the following sections. These design guidelines propose an initial answer to research question Q4. To apply these guidelines effectively, designers should have a detailed understanding of the task features and the user’s knowledge of the environment.

### **5.4.1 Design According to Task Characteristics**

The goal of Contextualized Videos is to help high-level spatial cognition tasks. Apparently, not all of these tasks can benefit from all the designs. I characterize the tasks according to multiple criteria. The designers can find proper designs according to the character of the tasks to be supported. Figure 5-14 shows a 3-dimensional task characterization.



**Figure 5-14: A Task Characterization using 3 criteria**

### 5.4.1.1 Select Design According to Task Information Requirement

When performing different low level tasks, the user may require different information from Contextualized Video interfaces: the video content, the spatial context, or the relationship between multiple videos and between a video and its context.

For high level tasks, e.g., the path reconstruction task used in the formal evaluation, the situation is more complex. Users have to actively navigate through the interface in order to acquire the information they need. In such cases, a single design may not work well. The designers have to analyze the task before constructing their own interface, by combining a set of designs through proper interactions.

Some typical low level tasks are listed under the guidelines:

#### **Guideline 1.1: Emphasize videos for tasks that mainly require video information**

Typical tasks include:

- Overview monitoring - glance at videos without focusing on any specific one, as in the monitoring activity.
- Close observation - observe a person or activity in detail.
- Content-based search - the user knows part of the content, e.g., a landmark in the video, and wants to find its context, so she will scan the videos for the landmark.
- Content-based travel - shift from one video to another, without thinking of the two videos in a global reference frame.

The videos can be emphasized in multiple ways, including increasing physical display space of the videos, putting videos at the center of the display, highlighting the border of videos, hiding the spatial context visualization to make videos more salient, etc.

#### **Guideline 1.2: Emphasize model for tasks that mainly require contextual information**

Typical tasks include:

- Travel - travel from one place to another in the global reference frame.
- Route Planning - look for a route from one place to the other.
- Location based Search - the user knows the location in the model and wants to find the corresponding video.

The model can be emphasized in multiple ways, including increasing physical display space of the model, putting the model at the center of the display, highlighting the model with brighter colors, hiding the videos to make the model more salient, etc.

**Guideline 1.3: Maximize the videos and their nearby context for integrative tasks (involving information from both the model and the videos)**

Typical tasks include:

- Orientation-based prediction - predict where the person in the video will go, by mentally registering the video's orientation into the 3D model's reference frame. This task appears in tracking activities.
- Landmark-based prediction - predict the future location of a person outside the video camera's range, based purely on landmarks. This task appears in tracking activities.
- Multi-video registration - judge the spatial relationship between objects in two or more videos. This task may appear in teleconferencing or tele-collaboration activities.

### **5.4.1.2 Select Design According to Time Pressure**

**Guideline 1.4: Use embedded designs for time-critical Tasks**

I found that embedding videos into the model supports and encourages a realtime strategy. Thus, embedded designs should be used in time-critical situations where replay of the video is not possible or practical. The reason is that embedded videos have higher display proximity than associated videos. The display proximity reduces the spatial cognition time, therefore allows realtime strategy.

Because the users are multi-tasking in realtime, their mental workload is higher than non-realtime usages. Associated designs encourage the user to focusing on observing the videos before constructing a comprehensive understanding of the situation. Therefore, they can be used for non-time-critical tasks. Up to now, no significant precision difference was found between the two designs.

**Guideline 1.5: Use combined designs to allow flexible user strategy**

According to the preliminary and the formal evaluation result, combining embedded and associated video normally results in a balanced design that performs well for all types of tasks. Combined designs also provide more visual cues to link the videos and the model. Even though the embedded videos of combined designs are inherently smaller than pure embedded videos and not appropriate for close observation, the participants can use other strategies to overcome this problem. Unless the tasks supported are predominantly video intensive or model intensive, combined designs should be used.

### **5.4.1.3 Select Design According to Task Goal**

**Guideline 1.6: Use simple 2D maps and associated designs, or minimize the embedded videos, for spatial understanding tasks**

**Guideline 1.7: Use 3D and embedded designs for local navigation tasks**

I found that users would like to look behind the camera in order to understand the exact location and orientation of the video content. The navigation task often involves such an information requirement. The users want to see a full map in one view when trying to reconstruct a path based on the observed videos. In such a situation, a clear map without occlusion or distortion will be the best. Previous research also showed that tasks involving spatial understanding favor more exocentric viewpoints like a 2D map or a top-down overview of the 3D model, while tasks involving local navigation favor more egocentric views [Aretz 1991; McCormick, Wickens et al. 1998; Wickens and Hollands 2000].

### **5.4.2 Design According to User Characteristics**

**Guideline 1.8: Use simple designs for short-term use and non-expert users**

From the formal evaluation, I found that 2D Associated videos are relatively simple and easier to learn. Thus, they are more appropriate for short-term use and non-expert users. 3D designs and embedded designs take a longer time to learn.

In this formal study, the Contextualized Video visualizations helped non-resident participants in the path reconstruction task, as their performance was not significantly worse than resident participants.

**Guideline 1.9: Provide rich context cues for users who are familiar with the environment**

More cues can be presented in the Contextualized Video visualizations for expert users who are familiar with the environment, because I found that users with a high level of spatial knowledge of the real environment can utilize a more abundant set of landmarks and can think of the situation in the real environment. This is consistent with the engineering psychology finding that expert users can make sense of new information faster because they have skilled memory and more a flexible mental model of the system [Wickens and Hollands 2000].

The user integrates multiple cues to understand the spatial relationship and make decisions. Generally, how cues affect people's thinking is task and user dependent. The following is a summary of the visual cues we should try to provide in Contextualized Video designs:

- Temporal-Location Cues: quickly switching between two views, utilizing short term sensory buffer and short term memory.
- Gestalt Perception Cues: Gestalt laws of perception summarized many important cues that manifest association relationship [Wertheimer 1923; Ellis 1938; Chandler 1997; Ware 2000]. The users can see a big picture from the salient global features. Examples of Gestalt cues are often given in 2D display [Chandler 1997; Ware 2000]. It's not obvious how to utilize them to improve interactive 3D applications. Also the ranks between these cues are not clear.

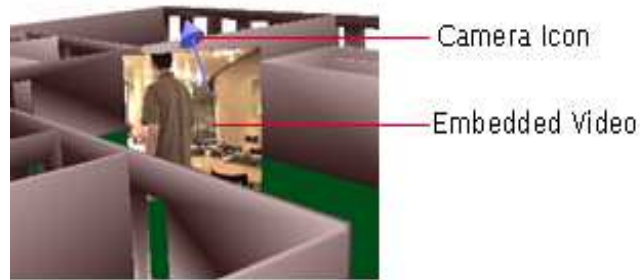
- Connectedness: e.g., callout lines between an object and its label.
- Proximity (see also Proximity Compatibility Principle [Aretz 1991]) . E.g. embedding video into its context.
- Common fate: same rendering style, consistent lighting... e.g., render the object as similar to its appearance in the video as possible.
- Common region: e.g., projective videos are rendered on exactly the same region as the model.
- Similarity cue: e.g., the video and the model views share some common features.
- Symmetry: e.g., show the video on the back of the video screen as a mirror of the original video for embedded.
- Kinetic cues: e.g., users can understand the shape and spatial relationship of the video and the model when he/she can interact with it.
- Single-representation spatial cues: These cues are processed in a higher level human information processing stage, e.g., in the visuospatial sketchpad in working memory. The user needs to infer depth information from the 2D display:
  - Spatial position cue: e.g., embedding a video into its 3D environmental model (Figure 5-15). This cue can be provided purely by the video or by the model.
  - Spatial orientation cue: e.g., the 3D camera icon in the 3D model shows both the video's position and orientation in the 3D model (Figure 5-16).
  - Occlusion cue: e.g., the occlusion between the 3D walls and the video in Figure 5-15 indicates their spatial relationship.
- Multi-representation (semantic) cues: These cues need cross-model processing in working memory. It can be understood as attention shift and information exchange between the three components of the working memory: visuospatial sketchpad, phonological loop and episodic buffer.
  - Text cue: e.g., symbolic labels of videos. Text cues may take advantage of non-spatial processing components in the working memory (e.g., phonological loop) to allow better parallel processing and larger short-term information storage.

### 5.4.3 Other Guidelines

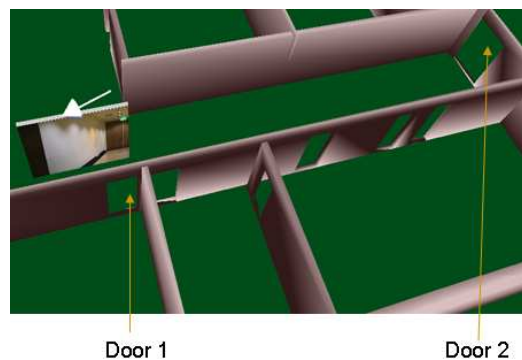
#### **Guideline 1.10: Avoid conflicting and misleading visual cues**

Since multiple visual cues exist together in Contextualized Video designs, we need to control these cues carefully to prevent misleading cues and conflicting cues. We have to carefully control the visibility of the cues in different conditions, because the visibility determines whether they can affect our thinking and action. For example, a partially occluded cue may be harder to notice. Also, too many cues in one display make each individual cue less salient (display clutter problem). The conflict between cues is another problem. We observed such a case in the preliminary evaluation. When a user saw a door in the video, she immediately assumed this door corresponds to the door that directly faces the video in the model. This is actually a wrong cue, because the video is presented as a billboard that automatically rotates to face the user. The 3D camera icon indicates that the camera is not shooting toward another door (Figure 5-16). When there's a

conflict, the user's attention is drawn to resolve it. Hence cue conflict may lead to high cognition load.



**Figure 5-15: Cues of Embedded Video**



**Figure 5-16: The video orientation and the object space proximity cue indicate Door 1 to be the one captured in the video, while camera icon orientation indicates Door 2 in the video.**

## **5.5 Summary**

In this chapter, we designed and evaluated Contextualized Videos designs for video surveillance domain activities along two design dimensions: model visualization and video-model layout. We finished the first of the four research cycles planned in our research approach (Section 1.7). This chapter, together with Chapter 3, addressed research question Q2 (For a particular domain and activity, what are the usable Contextualized Video designs and their limitations?).



## 6 Embedded Videos

The first experiment showed that embedded videos, a subclass of Contextualized Videos, led to better performance for the path reconstruction tasks, where the participants followed a target through simulated surveillance videos and marked the target paths on the environment model. However, the embedded video designs differ from the associated design in three aspects: spatial proximity to the camera, orientation alignment with the camera, and the Gestalt cues that help the users to link the video and camera [Wertheimer 1923; Ellis 1938; Chandler 1997; Ware 2000].

Hypothesizing that the orientation cue is the main factor that gives embedded videos an edge on tracking performance, the experiment reported in this chapter investigated the interaction between the orientation cue and the spatial proximity cue presented in embedded videos.

### 6.1 Experiment

#### 6.1.1 Designs

This experiment contained two independent variables. The first one was the distance between the video and the camera's focal point, or the *distance variable* in short. The second independent variable was the directional alignment of the video to the camera, or the *orientation variable* in short.

The distance variable had two value levels: the videos were either placed close to the camera's focal point (*closely-related video*, Figure 6-1 (a) and (b)) or faraway from the camera's focal point (*remotely-related video*, Figure 6-1 (c) and (d)). The model was presented in 2D, because the previous study showed that 2D designs are easier to learn. We used callout lines as well as color-coding to link videos to their contexts.

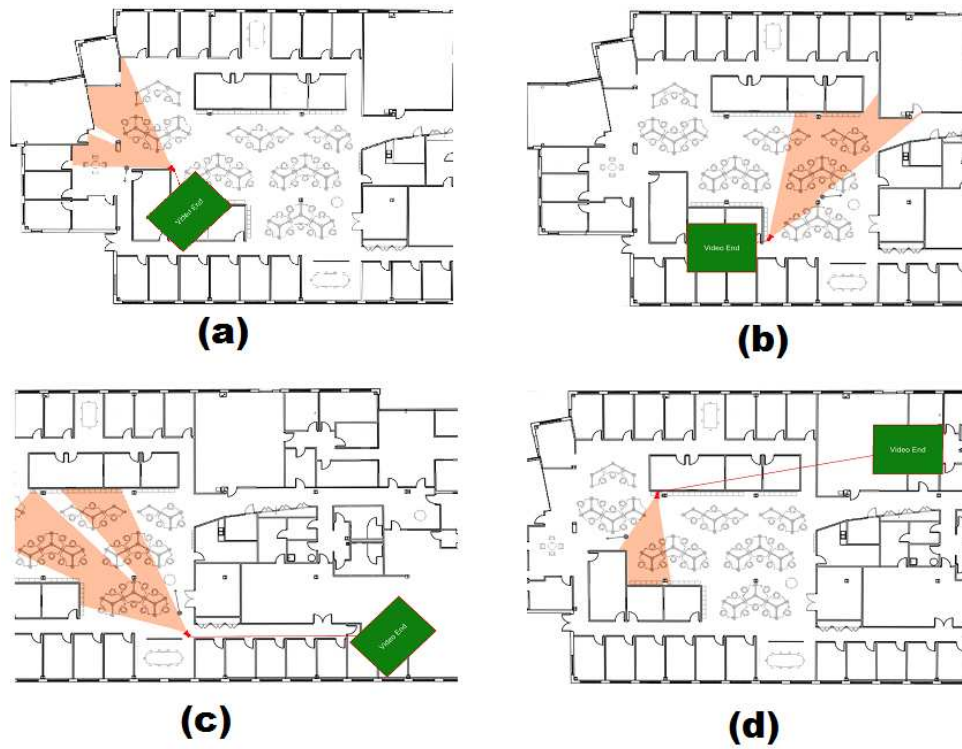
The orientation variable also had two value levels: *camera-aligned video* or *user-aligned video*. A camera-aligned video was oriented to face the camera so that the orientation of the person and objects in the video can be easily judged (camera-aligned video, Figure 6-1 (a) and (c)). A user-aligned video was oriented toward the user to facilitate viewing (user-aligned video, Figure 6-1 (b) and (d)).

Combining the two independent variables, four designs were created: closely-related and camera-aligned video (CC), closely-related and user-aligned video (CU), remotely-related and camera-aligned video (RC), and remotely-related and user-aligned video (RU).

In the previous experiment, the embedded designs allowed users to switch between camera-aligned video and user-aligned video by pressing the "ALT" key. To minimize user strategy effect, users were not allowed to dynamically shift between the two views in this experiment.

The coverage area of the camera showed a precise and strong orientation cue of the camera. The display size, the video size and the coverage area of the map was exactly the same between conditions, so that the information presented by different designs was equal. Given the same display size, closely-related designs often lead to smaller video size than remotely-related designs for two reasons: (1) The arrangement of the videos on the display was irregular in closely-related designs, since the video position was determined by the camera's location on the map. (2) All the videos and their nearby

context had to be shown in one display, yet none of the videos could overlap. Therefore, the video size in closely-related designs was reduced to match that of remotely-related designs.



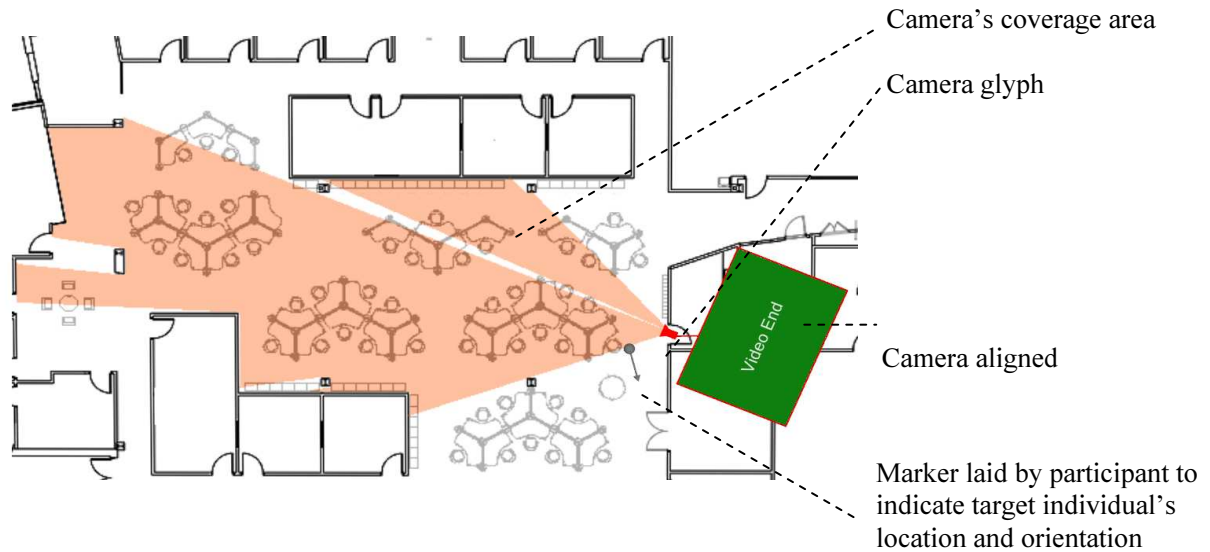
**Figure 6-1: The four Contextualized Video designs used in Task 1: (a) closely-related and camera-aligned video (CC); (b) closely-related and user-aligned video (CU); (c) remotely-related and camera-aligned video (RC); (d) remotely-related and user-aligned video (RU).**

### 6.1.2 Tasks

Since the previous empirical study focused only on the video surveillance domain, the findings may have their limitations. One way to generalize the findings is to evaluate the designs using more abstract tasks that are important for many Contextualized Video applications.

This experiment used the Type 1 integration task selected from the abstract task taxonomy summarized in Chapter 4.5. In this task, users map an object's location and direction from the video to the model. This task intended to measure the mental mapping operation. Therefore, we tried to minimize user action and other mental operations involved in this task. Multiple trials of Type 1 integration task were used to compare the associated design and the embedded design.

Since Type 1 integration task is a basic task, a realworld activity often contains multiple Type 1 integration tasks as well as many other tasks. To understand how well the findings using Task 1 apply in a real usage situation, a more complex virtual world path reconstruction activity (Type A activity for convenience) was used to check the external validity of the findings. The virtual path reconstruction activity was similar to the one used in the first user study, but it more strictly controlled user strategy.



**Figure 6-2: Task 1 screenshot. The 2D map was shown as a grey-scale image. The walls and doors were shown in black, while desks and cubicles, which could partly occlude the camera's view, were shown in grey. The camera's coverage area was shown in semi-transparent red on top of the 2D map, so that the landmarks captured by the camera can be easily identified.**

### 6.1.2.1 Mapping from video to model

#### Rationale:

The Type 1 integration task is a basic video-model mapping task in my task taxonomy. I hypothesize that the main difference of embedded vs. associated Contextualized Video designs is the display distance between the video and its context. This difference may lead to performance differences in this basic video-model integration task.

#### Task Description:

The Type 1 integration task asked the user to specify the location and orientation of the target on the model according to a single video and its nearby context. The video stopped playing when the moving target walked out of the video. The subject was then asked to specify the location and the orientation of the target within a limited precision. Time and the number of tries before success were measured. I chose to control the task precision and measure the task time, because users may have different understandings of the acceptable levels of accuracy when specifying the location and orientation of the target. Some users may feel that an error of 10 pixels is acceptable while others may not. We eliminated this subjective factor by controlling the precision of the answer.

The participants first specified the location by left clicking the model using the mouse. A blue marker appeared at the indicated location. The system then calculated the error of the participant's answer. The marker would turn green if the answer was within a precision range. Otherwise, the marker would turn red. In this case, the participant had to cancel the marker by a right click and then lay a new marker. In this way, we prevented

the participant from dragging the marker around to look for a sweet spot. Also, imprecise decisions were penalized by longer task time.

Some users may prefer to use trying-and-cancelling strategy to look for a sweet spot instead of pure mental mapping. One way to prevent the trying-and-canceling strategy is to make the penalty for an error action much longer than the time to think. We used a countdown counter to let the user notice the penalty time.

Once the location of the target was successfully specified, the participant then indicated the orientation of the target person. When the marker passed the precision test, a blue arrow appeared, starting from the center of the marker and ending at the current mouse position. When the participant moved the mouse, the arrow followed. Once the participant felt the arrow's direction was correct, he nailed down the arrow by a left click. Similarly, the arrow would turn green if the indicated direction was precise enough. Otherwise, the participant had to release the arrow with a right click. When the users answer was accepted, the current trial would end and a new trial would begin.

The difficulty of this task mainly depends on five factors:

1. Landmark configuration

The first factor is the number and configuration of recognizable landmarks that the user can mentally map from the video to the model. People sometimes use the landmarks as references to help determine the target person's location and orientation. More references make the mapping easier.

2. Camera distance

The second factor is the distance between the target person and the camera. When the target is farther away from the camera, it maps to a smaller number of pixels on the video. In other words, a certain number of pixels on the video will map to a larger area on the model at a greater distance, and consequently makes it harder to do precise mental mapping.

3. Moving direction

The third factor is the walking direction of the target person. Walking from left to right (or from right to left) of the camera is often easier to understand than walking toward or away from the camera, because the depth cue provided by the latter situation is harder to interpret than the direction cue provided by the former situation.

4. Moving speed

Higher walking speed makes it harder to perceive the orientation and location of the target, but very low speed is also problematic. In an extreme case, where the video becomes a static picture and the target is not moving at all, the user can judge the orientation and location of the target person from a single frame without understanding the target's previous path. Since my research object is dynamic video instead of static picture, I wanted to avoid such an extreme case in the experiment.

5. Target position when video stops

The task can be of different difficulty depending on where the target person is when the video stops. If the target person's foot is still in the video, mapping is relatively easy because the user can utilize information shown in the static picture. Mapping is harder if the target person is out of the video, because the user has to totally depend on their memory of target's path in the video.

I decided to hold most variables in order to reduce possible confounding effects. However, I varied the first factor between trials to cover three difficulty levels, so that the

experiment designs would cover more abundant real world situations. Table 4-1 summarizes the details of the videos.

Number of Trials	Configuration of cameras and feature of videos	Difficulty Level
3	Camera shoots open area, with few references and landmarks	hard
3	Camera shoots open area, with references, e.g., chair, table	medium
3	Camera shoots a hallway, with references, e.g., doors and posters	easy

**Table 6-1. The three difficulty levels of a Type 1 integration task**

**Metrics:**

1. Total task time of the 9 trials. The time to indicate the location and the time to indicate the orientation were measured separately. Because the task time for each single trial was very short, using a relatively large number of trials can reduced the task time variation. Since we could not measure cognition time directly, I tried to minimize the time of interaction by simplifying the process of indicating the answer. Because the action involved in this task was as simple as clicking with the mouse and people’s performance of this action was normally consistent across designs, I did not expect that to be a major variation factor.
2. Total number of errors of the 9 trials. A marker that was laid outside of the error tolerance range was considered an error. A user could make multiple errors in one trail.
3. Reported mental workload. Since the action was very simple, the main part of the mental workload went to the mental mapping between the video and the model. For within-subject test, the reported mental workload was expected to be more reliable than the between-subject test in the previous user study.

**Hypotheses:**

1. Embedded designs would lead to shorter task time. The reason is that the display proximity of embedded designs allows faster linking between the video and the model.
2. The reported mental workload would be lower for the embedded design, which has shorter display distance between the video and its context. The shorter distance may lead to less eye movement and shorter cognition time.

**6.1.2.2 Virtual World Path Reconstruction**

**Rationale:**

The virtual world path reconstruction activity (Task 2 for convenience) was used to simulate a real world situation, where the Type 1 integration task is only one of the multiple steps to finish the activity. Task 2 contains the following tasks:

- Scan multiple videos to detect the target person. This task is video intensive.
- Mentally map the target’s location and orientation in the video to his/her location and orientation in the spatial context. This is a Type 1 video-model integration task.

- Indicate the target's path by laying markers on the model using the mouse. This task is model intensive.

Because of the complexity of the activity, the users can employ very flexible strategies. They dynamically decide when and how well the video-model mapping task will be done. They also evaluate the reliability of the mapping result using other information. For example, if the mapped orientation of the target is inconsistent with the natural extension of the current path, the user has to judge whether the target made a sharp turn or the mapping result was wrong?

### **Task Description:**

This activity simulates a suspicious person tracking situation, where the user not only needs to follow the suspicious person through multiple videos, but also needs to mark the suspicious person's path on the model by clicking a series of blue dots with the mouse (Figure 6-3).

There was a maximum distance between the two nearby markers so that the precision of the path can be easily judged. The user was trained on how to lay the markers. In the experiment, a circle showed up around the newly placed marker to indicate where the next marker could be placed. The participant could not put a marker outside this circle.

The subjects were not allowed to do the task during the first play of the videos. In this way, they were forced to focus their attention on observing the videos. Once the videos ended and a start message appeared, the participants could start to indicate the path. The videos replay automatically once they are end. Since all the videos were of the same length, they were always synchronized.

If a marker's location was not precise enough, it turned red and the user had to replace this marker with a new one before indicating the next marker. The time to correct the error is the time penalty. Once the user finished the whole path, a success message showed up. The final path, the total time to specify the whole path, as well as the time to specify each marker was saved.

I decided to control the task precision and measure the task time. In other words, I measured how much time it took for a participant to form a correct understanding of the path. For this activity, task time is easy to compare, but it is hard to find a reasonable way to quantitatively compare the precision of two reconstructed paths. Counting the number of errors is problematic, because an earlier error on the understanding of the path tends to trigger more errors in the later phase of the task. Another idea was to sample multiple points on the path and measure the average distance from the indicated location to the actual path, but a single error about the orientation of the target could lead to a totally wrong path and consequently a very large distance error. Either way, the measurement is dominated by a single error instead of the whole path. To overcome this problem, we needed to inform the participants of their error during the task. Therefore, I chose to control the task precision and measure the task time.

This study contained three trials, each of which used a different route. The size of the modeled environment and the length of the experiment limited the number of trials to a relatively small number. Ideally, the scene used in each trial should be totally different, to avoid learning effect.

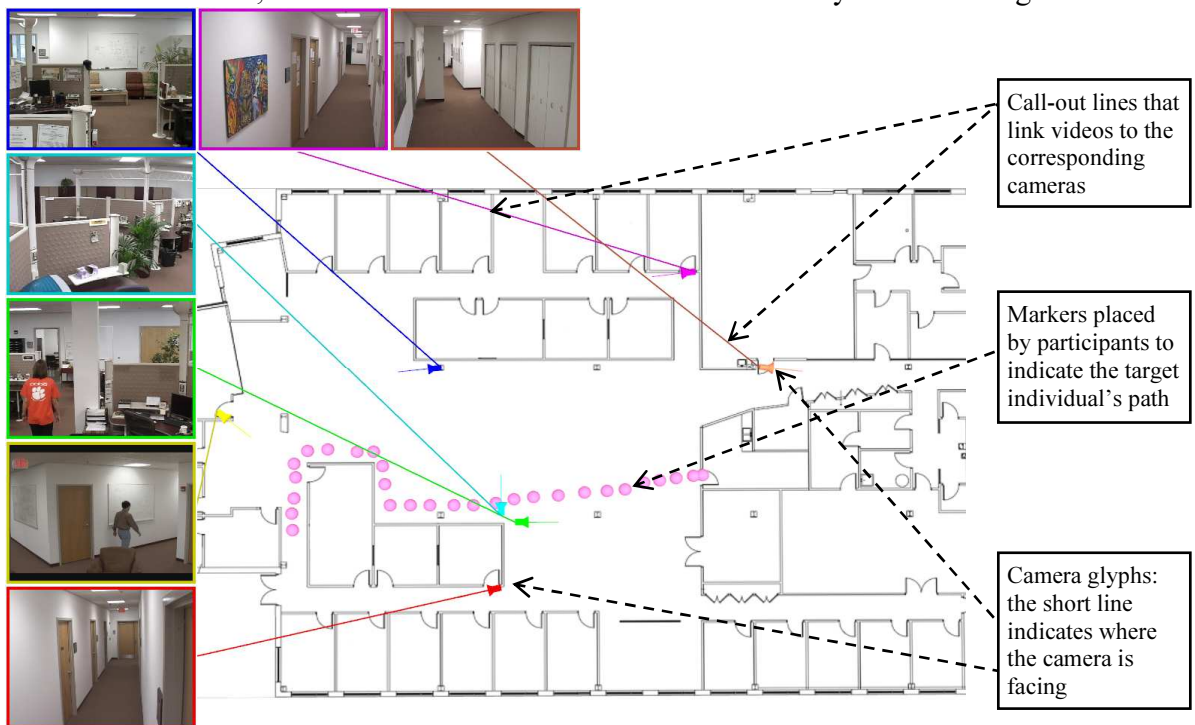
### **Metrics:**

1. The total time to specify the whole path.

2. The number of errors when performing the task.
3. Reported mental workload, which measures the workload of the whole activity.

**Hypotheses:**

1. The embedded design would lead to shorter task time; because the shorter distance between the video and its context makes it easier to match the landmarks in the video to those in the model.
2. The reported mental workload would be lower for embedded designs. In the associated video designs, the participants have to link the videos to their context by following the call-out lines. This seems to be a time consuming operation, which may increase the participants' stress in such a real time situation.
3. Task 2 contains multiple instances of subtask 1, and should be consistent with the task 1 result, so I used task 2 to check the external validity of the findings.



**Figure 6-3: The virtual world path reconstruction activity.**

**6.1.3 Procedure**

We used a within-subject test for the Type 1 integration task, and a between-subject test for the two path reconstruction activities. The two activities were evaluated using a between-subject test for two reasons. First, the path reconstruction activities covered a large portion of the real site and may have a pronounced learning effect between trials. Second, it is difficult to guarantee an equivalent difficulty level between two different trials using totally different paths.

A total of 24 participants between the ages of 22 and 35 finished the experiment. They were all non-residents of the site.

After a pre-questionnaire and spatial ability test, the subject was first given a guided tour around the building. The goal was to give the user a sense that she was monitoring the real site during the following tasks.

Session 1 of the experiment took about 30 minutes. The participants first took the training and practice for Type 1 integration task using one of the four designs. Then they finished 9 experiment trials of Type 1 task. They did the same for all the other three conditions. The sequences of designs were counterbalanced using Latin square design.

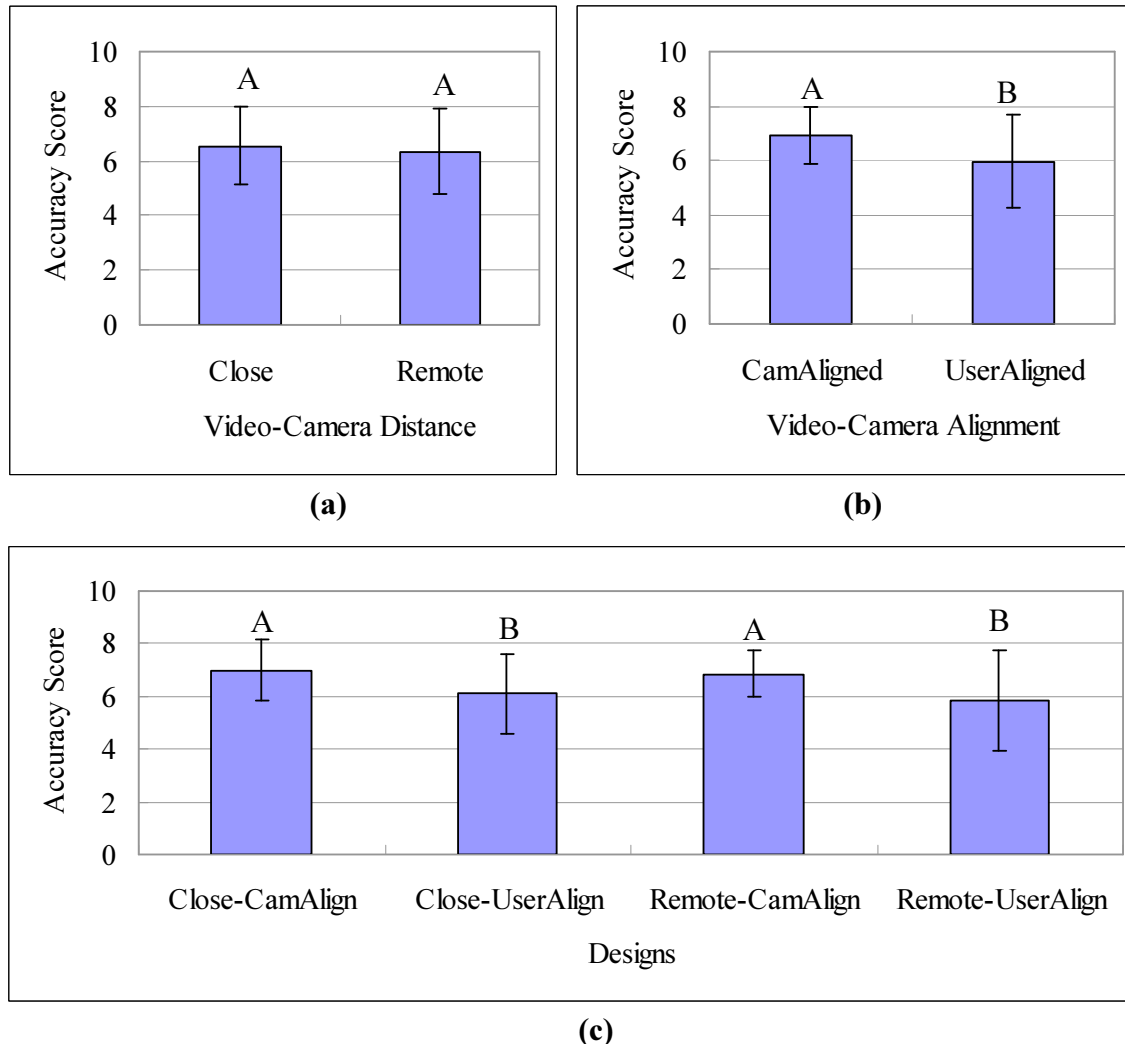
Session 2 took about 15 minutes. They were first trained to perform the virtual path reconstruction activity using one of the four designs. Then they did one practice trial. Following that, they finished 3 experiment trials of the path reconstruction activity.

Users were seated so that their heads are 20 inches from display. To prevent fatigue, we did not keep users from moving their head or body. In far distance conditions, the video and the camera's coverage area were at least 5-7 inches away. The video's size was 2.25 by 1.5 inches. The minimum distance between the video's border and the coverage area was 6 inches, or  $16^\circ$  view angle, in far distance conditions. In close distance condition, the distance between the center of the video to the center of the camera coverage area was at least 1.5 inches, or about  $4.7^\circ$  view angle. Therefore, the video and the coverage area could not be on the fovea at the same time in either condition. There was eye movement even when the video was next to the coverage area.



## 6.1.4 Results and Discussion

### Task 1 with controlled time:



**Figure 6-4: Task 1 accuracy with controlled task time. (a) Main effect of video-camera distance; (b) Main effect of video-camera alignment; (c) Interaction effect of video-camera distance and video-camera alignment.**

Repeated factorial ANOVA showed that video orientation had a significant effect on task accuracy ( $F(1, 81)=22.43, p<0.0001$ ). However, video-camera distance did not show significant effect. Also no significant interaction was found between the two factors. A *post hoc* analysis using Tukey's HSD test showed that cam-aligned conditions lead to significantly higher accuracy ( $p<0.0001$ ).

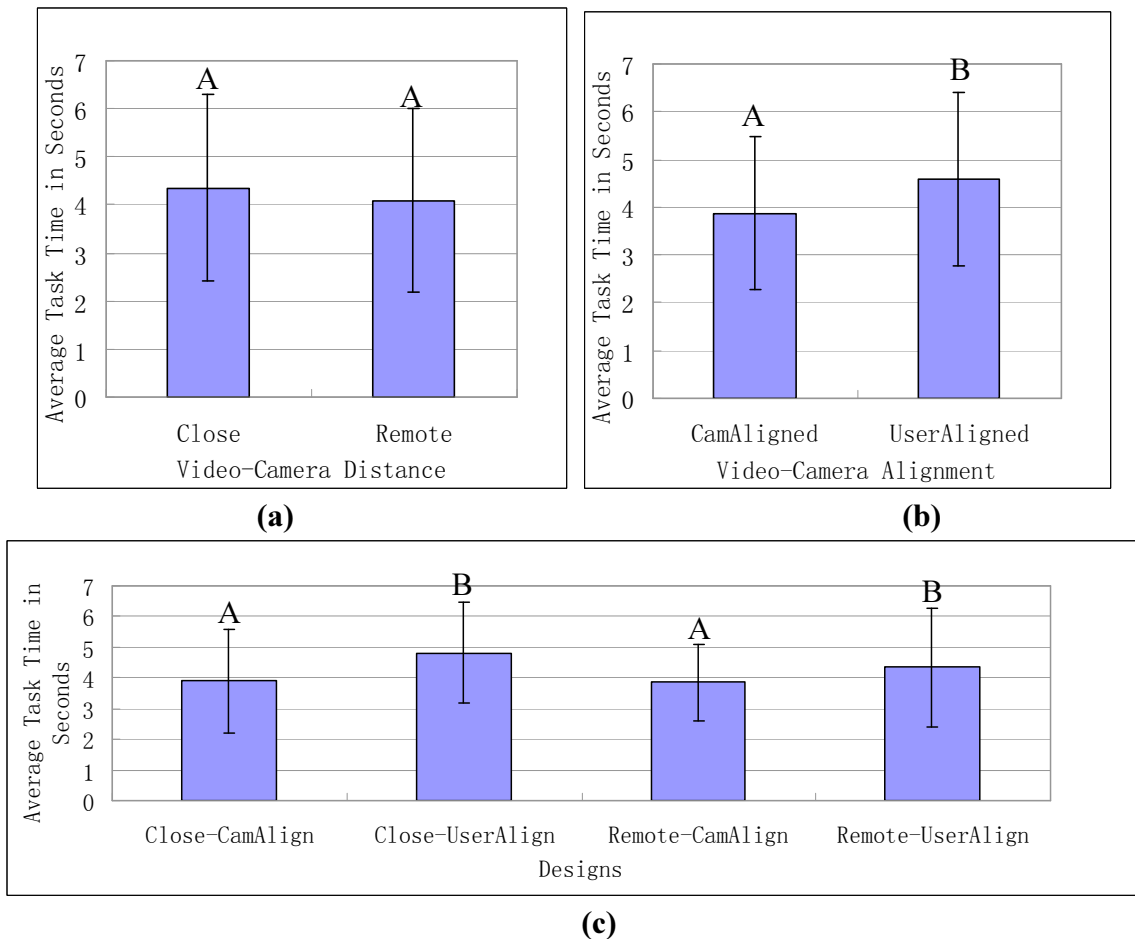
The participants' task accuracy had larger variation under user-aligned conditions. Considering the participants' spatial ability, we found that low spatial ability subjects often received the lowest score in RU condition. It seemed they did not have enough time to finish the mapping task. High spatial ability subjects, on the contrary, did as well in RU condition, as other conditions.

We found that high spatial ability participants did approximately equally well between conditions. It's the low-spatial ability people that made the difference. They did significantly better in camera-aligned conditions.

We also asked the participants to give a 1-4 rating (1 is the best) to the four designs. With regard to “easiness to use”, CC received the highest rating, followed by RC, CU and RU designs. For “easiness to learn” and “usefulness”, the sequence is the same.

**Task 1 with controlled accuracy:**

We ran repeated measures ANOVAs on the data collected from the remaining 25 subjects. Task time is strongly correlated with the video’s orientation ( $F(1,72)=9.46, p<0.005$ ). *Post hoc* analysis using Tukey’s HSD test shows that camera-aligned video leads to shorter task time ( $p<0.005$ ). Similar to Task 1 fixed time, video-camera distance did not show significant effect and the interaction between the two factors was not found to be significant.



**Figure 6-5: Task 1 time with controlled task accuracy. (a) Main effect of video-camera distance; (b) Main effect of video-camera alignment; (c) Interaction effect of video-camera distance and video-camera alignment.**

The fact that video-camera distance did not affect task performance as much as video-camera alignment has a very interesting implication. It implies that, compared with mental rotation, eye movement and mental translation of a visual element can be

performed better in parallel with other subtasks, which include watching the video and laying markers on the map. This implication needs to be confirmed with future experiments in the field of cognitive psychology.

We found that three subjects took too long (over 10 seconds per task on average) to finish the tasks and led to a very large variation in the data. They might have been too cautious and heavily sacrificed task time to precision. We think this is a phenomenon caused by experiment settings and can be corrected by excluding them during the data analysis process.

### **Task 1 Discussions:**

The video-model mapping task can be modeled into two mental steps:

*Coarse mapping*: mentally rotate the video or model to determine general orientation. Or use feature matching to identify landmarks in the videos.

*Fine-level mapping*: use landmarks and occlusion cues, to determine more accurate position. Landmarks can be used as a reference to judge the location of the target person, who was not shown on the model.

Not all the experiment trials needed fine-level mapping. In fact, for more than half of the trials, the participants did not need to recognize any landmark in the video. Of the 28 experiment scenes, I picked out 8 scenes that need a careful fine-level mapping to find the correct answer. After comparing task accuracy and task time, I found that camera-aligned condition even led to slightly longer task time. Inaccurate answers and video replay might be the reason. I observed many participants physically rotate their head and body to observe the videos. In these situations, the disadvantage of camera-aligned video dominates. Since user-aligned videos allow easier observation of the landmarks, some participants might do precise spatial judgment tasks better in user-aligned videos.

Mental rotation larger than  $90^\circ$  is hard and causes 87.5% of the total errors. For videos with less than  $90^\circ$  rotation, the performance difference between video-aligned and camera-aligned designs is not significant. Therefore, if the angle between the video and the camera's orientation is larger than  $90^\circ$ , the videos should be rotated to align with the camera to help mental mapping. An alternative solution is to flip the video horizontally if the rotation angle is larger than  $90^\circ$ . In this case, the interface should provide sufficient cues to inform the user whether a video is flipped or not.

### **Task 2 Discussions:**

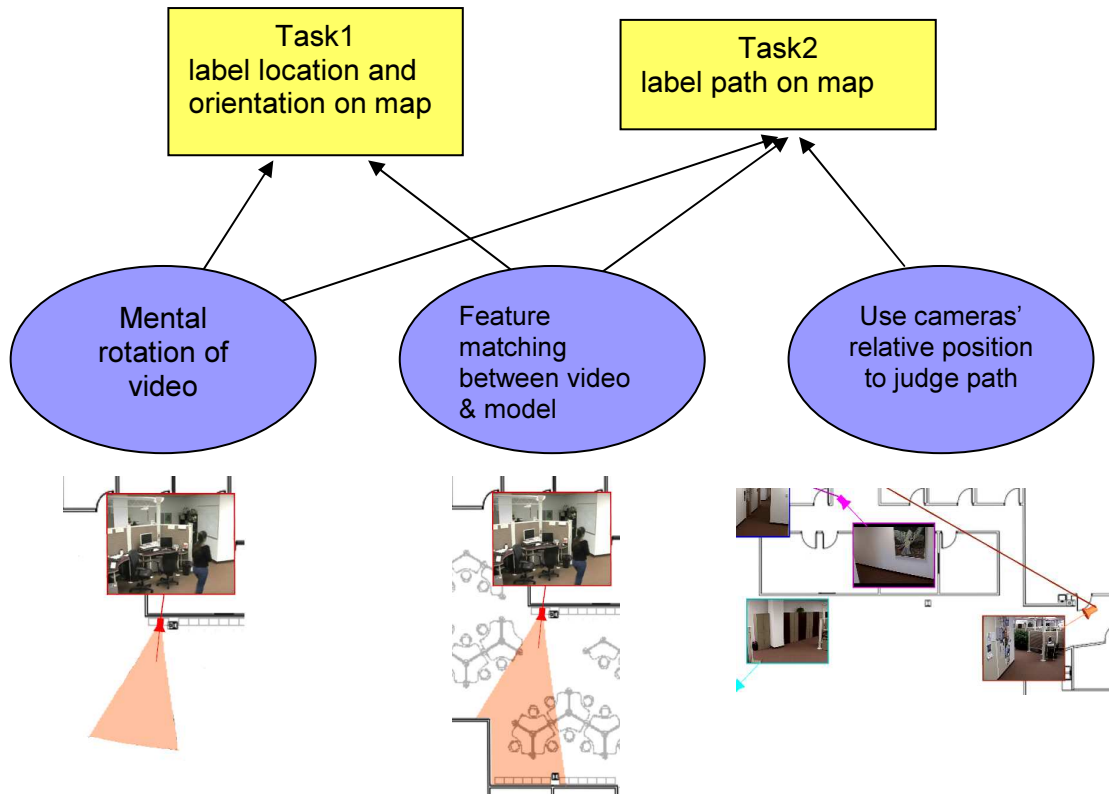
Task 2 was designed to be a between-subject test. Both camera-aligned video and closely-associated video led to slightly shorter task time, but the difference was not significant. Linking to the significant findings of the first experiment, video-camera distance might be a performance factor in realtime strategy, but not significant for non-realtime strategy.

The participants' task 2 performances did not significantly correlate with their task 1 performance. I believe user's flexible cognition in performing the two tasks explains the difference. After looking at the observation notes I took during the experiment and the interview record after the experiment, I found that the users used at least two different strategies in task 1 and at least 3 strategies in task 2.

For task 1, normally the user should mentally rotate the videos' reference frame to map the target's location and orientation onto the map. However, if the target is close to a landmark that can be easily mapped from the video to the map, the participants no longer

need to do mental rotation. Instead, they can use feature mapping directly. As shown in the middle image of Figure 6-6, when the target is close to the corner of a wall, the participant can easily find the corner of the wall on the map and indicate the location of the target.

For task 2, even more flexible strategy can be used. For example, even if the participant was not able to map the target's location from one video onto the map, he could still make a good guess on the overall path according to the appearance sequence of the target in the videos.



**Figure 6-6: The possible cognition processes of users when performing Task 1 and Task 2.**

Although the participants' cognition strategy might cause performance differences, the strategy could not be directly measured or recorded. To further investigate the effect of users' cognition strategy on their task performance requires atomic tasks under more strictly constrained conditions and is outside of this dissertation's scope. Even if we know this relationship, the strategy would be hard to fully control in a real usage case. Therefore, this dissertation focused on the effect of higher level designs on users' task performance when performing the mid-level tasks summarized in the task taxonomy.

## 6.2 Guidelines

The following design guidelines are summarized from the evaluation results:

## 6.2.1 Design According to Task Characteristics

**Guideline 2.1: If the task requires observation of the model without occlusion, then the videos can be placed farther away from the camera while keeping the orientation of the video aligned with camera.**

In this way, we can reduce occlusion of the video to the model without affecting performance too much; because this experiment clearly showed that spatial distance between the camera and the video does not have a significant effect on task performance. Also, the orientation and distance factor of embedded videos are not strongly correlated.

A special case of such a design is associated videos (videos are put outside the model) tilted in 2D or 3D to align with the camera's orientation. This is a counterintuitive design, but might be effective for tasks that require fast mapping of content in the video to the map. Interestingly, the ExoVis visualization [Tory 2003] proposed by Tory et al. is very similar to the tilted associated video design. Their comparative evaluation found that ExoVis design led to significantly better task performance than orientation icon, which is similar to associated videos.

**Guideline 2.2: If the task requires precise spatial judgment and the model shows plenty of landmarks, try both user-aligned and camera-aligned video layout.**

Some users might use feature matching instead of a mental rotation strategy when the landmarks are easy to observe. Since user-aligned videos allow easier observation of the landmarks, some participants might do spatial judgment tasks better in user-aligned videos. To support feature mapping strategy, the four sides of the video and the corresponding borders of the coverage area on the map can be color coded. In this way, the user can use the border as a reference to judge the subject's location on the map according to his location in the video.

**Guideline 2.3: In a complex activity that contains tasks other than spatial mapping, consider other cues that can bypass spatial mapping between video and model.**

For example, in the path reconstruction task, users can use the relative position between two videos to help judge the path even if they fail to map the target person's orientation from the video to the model.

## 6.2.2 Design According to User Characteristics

**Guideline 2.4: For low spatial ability users, use camera aligned design to make spatial mapping more intuitive.**

This can lead to faster mapping and fewer errors (supported by significant findings). An explanation is that a direct mapping operation replaced the costly mental rotation. High spatial ability users can do almost equally well under both conditions.

**Guideline 2.5: To support users of different preferences and working habits, both user-aligned video and camera-aligned video can coexist in a single user interface to allow flexible user strategy.**

As pointed out by Guideline 2.2, although camera-aligned video generally led to faster spatial judgment, user-aligned video can sometimes work better. The choice between the two depends heavily on the users' preference and the spatial cues provided by both the video and the model. Both designs can be supported on a single interface, as

long as the users clearly know which design they are using. The rotation between the two views can be animated to help users understand how the view is rotated to change between user-aligned video and camera-aligned video.

### **6.3 Summary**

Focusing on the video-model layout dimension of the design space described in Chapter 4, this chapter designed and evaluated Contextualized Video designs using the tasks from our Contextualized Video task taxonomy. Up to this point, we have finished the second of the four research cycles planned in our research approach (Section 1.7). The formal experiment findings addressed research question Q4 (What are the effects of various Contextualized Video designs on the performance of key Contextualized video tasks?).

During the formal experiments, we also used complex activities to check the external validity of the findings using the basic tasks selected from the taxonomy. This part of the experiment addressed research Q5 (How beneficial are Contextualized Videos in complex activities?).

## 7 Effect of Video Processing

Using automatic video analysis technologies, the target person's position and orientation can be extracted from the video and directly visualized on the model. Even though automatic video analysis algorithms are not yet fully reliable in general cases, we can still simulate the result of a perfect video processing algorithm and investigate how to combine the result with other Contextualized Video dimensions to better support surveillance tasks. This motivated me to investigate the effect of video processing on the design and usage of Contextualized Videos.

### 7.1 Experiment

#### 7.1.1 Designs



**Figure 7-1: Dynamic Path Visualization Prototypes used in the formal experiment. The path dynamically shows the target person's trajectory from his first appearance in a video to the current moment. (a) Dynamic Path Visualization without videos. (b) Dynamic Path Visualization with videos.**

This dimension contains a continuum of techniques that range between no processing to high abstraction. For example, if the camera's parameters are known, the path of the moving object in the video can be extracted and visualized on the model. In this way, the spatial mapping from the dynamic part of the video to the model becomes trivial. This design can be called *dynamic path visualization*.

I first implemented this design on the testbed using the background subtraction algorithm in OpenCV library. However, I decided to prototype this design in another way for two reasons: (1) the 3D position reconstructed from the video is not reliable due to occlusion, shadow, unknown target size, unexpected actions, etc; (2) the performance was not good enough for realtime use.

The second dynamic path visualization prototype was created by pre-computing the target's position at multiple critical points along the path (Figure 7-1). The pre-computed positions, along with their time stamp, are stored in a data file. In this way, the visualizations only need to interpolate between the stored positions at runtime. The dynamic path visualization design was combined with 2D embedded video design,

because the path can be seen clearly in a 2D design, and relating between the path and the video is likely to be easier than with associated design.

In dynamic path visualization design, the target person's current location and previous path can be seen clearly in a 2D map. Therefore, spatial mapping is no longer a problem. However, the detailed information of the target person and his surrounding environment has to be observed from the video. Tasks that require the user to relate the details observed from the video with the target's location on the model require some mental effort. Therefore, this experiment aimed to investigate whether the visual overload introduced by adding videos to the visualized path would outweigh its benefit to understanding the path. Two conditions were compared in this experiment: one with associated videos, the other without associated video (Figure 7-1).

### **7.1.2 Tasks and Hypotheses**

The task selected for this experiment was the real world path reconstruction task, where the participants first learned the path using Contextualized Video designs and then travel along the path at the real site.

This task can be connected with the virtual world path reconstruction task into a single activity. In the actual video surveillance application, the security guard often first tried to understand the path from the videos, and then went to the real site to catch the target, according to the learned path. I break this complex activity into two, and evaluate them separately, so that the cause of the performance difference can be easily attributed to a smaller portion. For example, if a participant fails to travel along the correct path in the real environment, I can easily tell whether it is a video-to-model mapping problem or a model-to-realworld mapping problem by using two separate activities. But I can never tell where the problem lies when using a combined activity.

The participant first observed the videos as well as the whole path of the target person in the Contextualized Video interface for a limited time. The path was shown on the model as multiple connected line segments. The videos showed 3-4 turns at easily identifiable landmarks.

After the participant learned the path, I measured the participant's walking speed. The participant was asked to walk a standard distance at a normal speed.

Then the participants were blind-folded and brought to the start of the path, facing the target person's orientation. They were asked to travel along the path of the target person and reproduce his/her actions at the proper location. Because of the limited size of the experiment environment, each participant performed three trials to avoid learning effect.

The real world path reconstruction task contains multiple tasks:

- Memorization of the route. This is a model intensive task.
- Mentally map the target's location and orientation in the video to a point along the route. This is a Type 2 video-model integration task.
- Mapping the landmarks on the model to their detailed appearance in the video. This is an integration task.
- Memorization of the detailed appearance of landmarks. This is a video intensive task.
- Travel in the real world according to the learned route. Since this task is not performed on the visualization, it does not appear in the task taxonomy. However, the visualization may affect the performance of this task by affecting the user's



encoding of the navigation process and the path in long-term memory.

We are interested in how much the designs would affect the overall performance of the activity. I controlled the task precision of recall and measure the time. The completion time, the location and the orientation of the action were measured. The participants were penalized by 1 minute for failing to remember a turn or action. The process was video captured and analyzed later to extract the time of decision making, and the time of deviation from the correct route.

### **Metrics:**

1. Time of completion of the path in the real world, time of decision making. The task time was normalized by the walking speed of the participant.
2. Precision of the path indicated in the real world. The precision was measured by the number of errors and the time of deviation from the correct route.
3. Reported mental workload.

### **Hypotheses:**

I hypothesized that dynamic path visualization with videos would lead to shorter decision making time during the realworld travelling process, because the landmarks are easier to identify when videos are present during the route learning process. Therefore, dynamic path visualization with videos would lead to shorter path reconstruction time in the real world.

## **7.1.3 Procedure**

This experiment was designed to be a within-subject test. It was carried out together with the second formal study described in Chapter 6.1. The task was performed right after the Task 1 and Task 2 used in the second experiment. The following steps were performed for this experiment:

1. Training and practice for the real-site path reconstruction activity for one design.
2. Finish 3 trials of the real-site path reconstruction activity. The users were blind-folded when brought to the real site.
3. Mental workload test and subjective rating of the designs.

## **7.1.4 Results and Discussion**

Between the two conditions, no significant performance difference was found on the time and accuracy of the realworld path reconstruction task. This result indicated that the participants performed approximately equally well even without seeing the videos. Therefore, for a relatively simple environment, the path and its context shown on the model can possibly provide enough cues to recall the route during the actual travel process.

On the other hand, interviewing the participants revealed the potential usefulness of videos. Most participants felt that the route shown on the map was the most helpful information. Still, about 2/3 of the participants reported using landmarks in the video to recall the path. More than half of the participants felt they can do equally well without looking at the video. They felt that looking at both the video and the map at real time is mentally overwhelming.

Although most participants felt that adding videos to the path did not cause significant visual overload, a small number of participants reported that always trying to integrate the route and video was mentally overwhelming.

Please keep in mind that we are assuming perfect automatic tracking algorithms that can precisely, reliably and quickly calculate the targets' location on the model. Many interesting questions need to be investigated if we integrate a real video processing algorithm with Contextualized Video interfaces. For example, it is unclear how the visualization will be used if the dynamically visualized information contains a lot of noises because of unreliable tracking or because of imprecise camera parameters, etc. Also, any video processing algorithm has a processing delay between the input and the output. How much lag between the visualization and the video output will change the users' usage pattern? All these are interesting questions to investigate in the future.

## **7.2 Guidelines**

The evaluation results indicate the following design guidelines:

**Guideline 3.1: If the information required by the task can be reliably extracted from the video, it should be visualized on the model.**

This experiment found that the extracted information visualized on the model attracted more attention than the raw videos. The dynamic path visualization was reported as the main source for memorizing and reconstructing the path. Therefore, we should try to visualize the path information on the model if the target's position can precisely be extracted through reliable video processing and can be reverse projected onto the model given the known camera parameters.

**Guideline 3.2: If the task requires detailed video information that cannot be fully extracted, raw videos are still needed.**

“Sight is the sense of choice” [Davies 2005]. Video processing made the choice for users by predicting users' information requirement and extracting the information from videos. However, users' information requirement might be unclear until he sees the video. Segmenting the target in the video from the background may lose important cues, especially when the target interacts with the environment. For example, if the person passes an office plant, the motion of the plant gives the observer an important cue about the motion of the person, but that cue is hard to reconstruct in the 3D scene.

**Guideline 3.3: If the task is to learn how to carefully maneuver through the environment, show both the path and video on the interface.**

Our experiment showed that although users tend to focus more on the visualized path, the videos provide richer cues for route and action recognition, e.g., actual appearance of landmarks, or when and how to perform a particular action. Presenting both the video and the extracted information allows flexible user strategy. Users might be able to gradually form their strategy to make use of both during long-term use of the interface.

**Guideline 3.4: Provide cues to help users manage their attention.**

When both the video and the dynamic path change over time, users may find it hard to manage their attention. Reducing dynamic information may help reduce mental workload. For example, the videos can be hidden most of the time. Only show videos at

important places, e.g., when making a turn. Chen et al. used such an adaptive method in their interface for driving direction learning [Chen, Neubert et al. 2009].

### **7.3 Summary**

Focusing on the video processing dimension of the design space described in Chapter 4, this chapter designed and evaluated Contextualized Video designs using the tasks from our Contextualized Video task taxonomy. Up to now, we finished the third of the four research cycles planned in our research approach (Section 1.7). The formal experiment findings addressed research question Q4 (What are the effects of various Contextualized Video designs on the performance of key Contextualized video tasks?).

## 8 Navigation

Navigation technique is the last dimension being explored in the design space (Chapter 4). Navigation is the act of “moving”, which is defined as “to change position or posture”[Merriam-Webster's 2000]. In Contextualized Video interfaces, navigation happens when the visual display changes in response to users' indication of movement. When indicating a movement, the user may have different navigation goals: (1) to find a better viewpoint (e.g. uncluttered, unobstructed, and undistorted) to observe a video or multiple videos, or (2) to find a better view to understand the spatial relationship between videos, or (3) to find a view for easier mental mapping between a video and the spatial context, or (4) to find a better view of the spatial context.

Initially, we planned to evaluate navigation techniques in terms of their support of the above task goals. However, as learned from prior experiments (Chapter 6.1.4), users' realtime strategy when performing a task can be very flexible. They may have multiple candidate navigation goals in mind. If one navigation goal fails, they may change to another navigation goal rapidly. For example, when following a person through videos, if the user can easily find the next video to observe in current view, then his navigation goal is to find a better viewpoint to observe the next video. Otherwise, his navigation goal may change to finding a better view of the spatial context, e.g., zooming out to overview in order to find the next video. If we overly constrain user strategy, e.g., not allowing zoom, the external validity of our finding is doubtful, as a well-designed interface normally supports a zoom function. Moreover, as the interface becomes more complex, the possible combination of navigation techniques increases exponentially. Summarizing the usage pattern of these techniques and explaining the patterns may be instructive to future designers.

Therefore, we decided to design and evaluate navigation techniques within a relatively complex interface.

### 8.1 Multi-View Interface

The designing experience and the guidelines summarized through controlled study allow us to create relatively complex designs that integrate more design features and provide more flexible usage. Designing complex interfaces is also an opportunity to examine the external validity of previous design guidelines.

The exploratory study described in Chapter 5.2 suggests multi-scale visualizations to support different tasks. Guideline 1.5, 2.5 and 3.3 suggest multi-view interfaces to support more complex realworld tasks and more flexible user strategies based on users' working habits. Combining the guidelines with creativity, I created a new interface containing four views:

(1) A video bank view shows an overview of all the videos. This view is similar to the traditional video surveillance interface. Selecting a video in this view will cause other views to update accordingly.

(2) A 3D context view shows the video in focus together with its nearby spatial context. The user can navigate using different techniques, zoom in and out, or click-and-link to other views.

(3) A peripheral view shows the videos that are neighbors to the focal video. A neighboring video (or neighboring camera) can be explained as follows: when a person walks out of the field of view of the camera in focus, he will appear in one of the neighboring cameras before appearing in any other cameras. The neighboring videos are determined according to their geographic configurations and the camera setup. The position of each neighboring video is calculated according to its camera's relative position with regard to the video in focus. In this way, the position of the peripheral videos gives the user a cue as to which video the target might appear next, once he exits the central video in focus.

(4) A focused video view shows the selected video enlarged at the bottom of the interface. The enlarged video is put below the context view in order to ease mental mapping between the video and the corresponding camera. When the user navigates, either manually or automatically, to look behind the camera in the context view, the enlarged video is visually aligned with the spatial context.

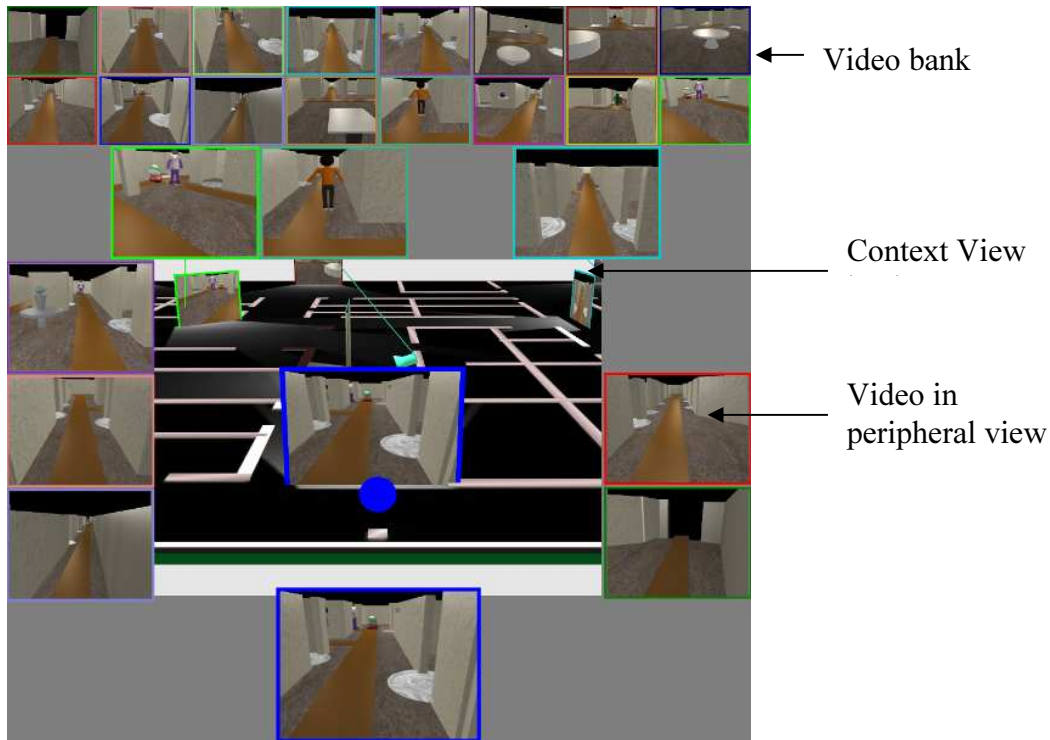
Click-and-linking is supported between the four views. When a user clicks a video in the video bank, in the 3D context, or in the peripheral view, the selected video will be enlarged in the focused video view. At the same time, all its neighboring videos are shown in the peripheral view.

When the context view automatically navigates to look behind the selected camera, the old central video and the newly selected video also smoothly animate in the peripheral view to accommodate the auto-navigation in context view. In this way, the user will not lose tracking of the target during the animation process.

These combined designs are called *multi-view visualization*, which is shown in Figure 8-1. Multiple navigation techniques were prototyped on this interface to support the navigation goals summarized at the beginning of this chapter.

As the interface becomes more complex, the cost of interaction grows [Lam 2008]. Particularly, navigation technique, as an important form of interaction to change view point and reduce occlusion, can have a big impact on the usability of these interfaces. For example, 3D embedded video allows more intuitive video-model mapping than 2D embedded video. However, it requires navigation to choose the proper view and to reduce occlusion. Since different tasks may require different sequences of views, navigation techniques need to be evaluated with regard to different tasks.

The videos in periphery and context view animate according to the user's position and orientation.



**Figure 8-1: Prototype for multiple view interface.**

## **8.2 Navigation Designs**

When discussing the navigation design dimension in Chapter 4.4, we highlighted two design factors that are especially important for Contextualized Video interfaces. One factor is the context of navigation, and the other is the mode of navigation. Navigation designs that preserve the working context and cause low time and cognition cost are likely to better support realtime tasks.

There are numerous design choices along each factor. In fact, they form a continuum. However, for evaluation purposes, we want to choose the most representative design choices for each factor. For navigation context, we chose two levels of context: overview, which shows a whole floor plan, and detailed view, which only shows nearby regions of the video in examination. For navigation mode, we also chose two levels of user control: a completely manual technique, or a semi-automatic technique, in which the user only specifies the next video to look at without control over the transition process. Combining these two design factors leads to four categories of designs, as listed in Table 8-1 and discussed in the rest of this chapter. All the navigation designs were prototyped in the multi-view interface.

**Semi-automatic Navigation in Detailed View (AD):** Shown in Figure 8-1. Navigation happens in detailed view, which contains one to several cameras and their nearby context.

The videos in context view were embedded as a fixed canvas in the model. The canvas was put on the near plane of the camera's view frustum. For now, we choose not to use dynamic embedding; as we are not sure how well dynamic embedding can be done without 3D reconstruction from the video, as the depth of the object in the video is

unknown. User's view point is put a little above and behind the camera's position, so that more context can be seen.

The videos that are close to the center video are displayed on peripheral view as in Girgensohn's spatial multi-video player. The positions of the peripheral videos approximate their correct spatial relationship with the video in center view. The videos move accordingly when the user changes view position.

In the detailed view, users can select a camera or video in peripheral view to navigate to look behind the camera along a pre-defined route. Automatic navigation time was controlled at 1.5 seconds.

Users could click a video in the video bank to navigate to look behind the corresponding camera. The navigation might cover a longer distance and may cause disorientation. A "zoom out – navigate – zoom in" technique can give a better sense orientation, but this technique could not be used here because it involves navigation in overview. This problem is an inherent drawback of detailed view navigation.

Users can zoom out to observe the whole environment from a higher observation point, which is farther away from the camera's location along the view axis. If the view is mainly a side view, the observation point will be put higher in order to reduce occlusion. However, the users are not allowed to navigate in the overview.

	<b>Navigation in Detailed view</b> + large video - small context	<b>Navigation in Overview</b> - small video + large context
<b>Manual navigation</b> + control on intermediate view - complex control	<b>Manual Navigation in Detailed View (MD)</b>	<b>Manual Navigation in Overview (MO)</b>
<b>Semi-automatic navigation</b> - no control on intermediate view + simple control	<b>Semi-automatic Navigation in Detailed View (AD)</b>	<b>Semi-automatic Navigation in Overview (AO)</b>

**Table 8-1: Navigation technique categories with their pros and cons.**

**Manual Navigation in Detailed View (MD):** The visual interface is the same as AD, but with different navigation controls. Navigation also happens in detailed view. However, the user uses the keyboard and mouse to travel around the detailed view.

This navigation control is similar to many 3D first-person shooter games, in which users hold the "w" key to go forward, the "s" key to go backward, the "a" key to shift left, and the "d" key to shift right. They drag with the left mouse button to rotate.

He can still click a video thumbnail in the video bank to automatically navigate to look behind that camera, but he cannot do this with a camera in the detailed view or a video in the peripheral view. This is the difference between MD and AD.

The user can move away from behind the camera and observe the model from behind the camera at a different angle. If the user moves forward and passes the video canvas, that video will be shown in peripheral view. Users can zoom out to observe the whole environment from a higher observation point. However, they are not allowed to navigate in the overview.

**Semi-automatic Navigation in Overview (AO):** The interface is shown in Figure 8-2. Navigation happens in overview, which contains all the cameras and the whole scene. They can zoom in to see the local scene from the same view angle, but cannot navigate in the local scene. Selecting a camera in the model or a video thumbnail on the video bank triggers automatic navigation to look behind the camera. The navigation path is along a bounding sphere's surface. The selected video will be highlighted and will be put in the peripheral area nearest to the focal point of the camera. The nearby videos of the selected video will be shown in the peripheral area as well.

**Manual Navigation in Overview (MO):** In this design, the user controls the observation point on the surface of the model's bounding sphere (without cap to avoid polar point problem) by dragging with the left mouse button. The user can click a video on the model or on the video bank to enlarge it in the peripheral area. The highlighted video will be put nearest to the focal point of the camera. The nearby videos of the selected video will be shown in the peripheral area as well. In this condition, the overview of the model will not change as the user selects a video. Users can zoom in to see the local scene from the same view angle, but cannot navigate in the local scene.

The following features are common among all the navigation designs:

Use the SPACE key to zoom in or out

Use the SHIFT key to hide the videos in the context view

Use the left mouse button to select a video to enlarge at the bottom; may also automatically navigate in the context view to look behind the camera when using semi-automatic navigation designs.





**Figure 8-2: Prototype for overview navigation interface. All the cameras and their coverage area should be shown in the 3D model. The red camera is selected and highlighted.**

### **8.3 Experiment**

The goal of this experiment was to understand the effect of the two identified design factors of navigation design on various realtime tasks selected from the task taxonomy. The specific research questions were:

- (1) In which view should navigation be performed, in detailed view or in overview?
- (2) How much control should we give to the user, full manual control or semi-automatic control?

The navigation techniques were designed to be representative and unbiased. We focused on realtime tasks because of the special challenges they pose and the importance of these tasks in multiple domains including video surveillance and tele-collaboration.

#### **8.3.1 Experiment Design**

We selected one representative design from each category. These four designs differed only on the two design factor variables; the other factors were kept equal and unbiased.

For both detailed view navigation techniques, we allow zooming out to overview and observe, but do not allow navigation in the overview. Similarly, for both overview navigation techniques, we allow zooming in to detailed view around the selected camera to observe, but do not allow navigation in the detailed view.

A real application is likely to allow users to navigate in both overview and detailed views. But for a particular task, given the realtime and multi-tasking feature, navigation

in the current view might be better, as there is time and attention cost to switch views. Also, in a real usage case, users are often trained, or set out by default, to navigate in one particular view. Assuming users have already made the choice, we investigate how the choice affects their performance and cognitive load. It would be interesting to allow navigation in both views and observe how users make the choices. However, it involves too much user strategy, which may lead to large variance in the performance data. We left it as future work.

A real application can also allow both manual and semi-automatic navigation. As discussed in the term explanation section, the tradeoff between the two methods has not been investigated for video surveillance tasks. Most existing techniques use semi-automatic navigation, but we suspect the ability to freely choose a view may give manual navigation an edge for some tasks. For example, an experienced user can manually follow a target through multiple videos and form a first-person understanding of the target's travelling process.

Depending on the visualization design, navigation can be performed in 2D display space, as in Girgensohn et al.'s Spatial Multi-video Player [Girgensohn, Shipman et al. 2007]. It can also happen in 3D model space, as in the auto-rotation technique we created before. In this experiment, I investigate navigation within the 3D model space that contains embedded videos. 3D model space navigation has multiple potentials: (1) 3D models allow more flexible viewpoints than 2D map; (2) 3D models allow better alignment between the video and the virtual environment; (3) 2D maps can be thought of as a special case of a 3D model (top-down view).

### **8.3.2 Scene Generation**

After considering the tradeoffs carefully, we decided to use a simulated environment (a virtual model not corresponding to a real world location, and simulated videos using animations of virtual characters rendered from the camera locations in the model) to compare the four designs.

Using a real site and captured videos for the experiment better resembles a real world situation. However, through previous experiment preparation experience, we found two major limitations to using a real site for a comparative experiment:

(1) It is hard to set up a site that allows us to create many unrepeatable trials of similar difficulty, particularly if each trial involves tens of videos. The size and layout of the site is often limited. Many objects on the real site cannot be easily moved and put back. Some unique details of the environment, e.g., a picture on the wall, can be easily remembered by the user, causing severe learning effect.

(2) When the scene is complex, scene and camera setup are very time consuming and error-prone. For example, if a scene requires eight actors to walk along the predefined path, and one actor makes a mistake, the whole scene has to be recaptured.

Using a virtual environment can overcome the above limitations. The site can be easily created and freely modified according to task requirement. Therefore, variance between trials can be better controlled. Simulated videos do not contain as many details and variations as real videos. They are harder to remember. This is desirable for reducing learning effect. Camera parameters are known precisely and the scene can be easily animated. Correcting an error also takes a much shorter time. In summary, given the limited resources, a virtual scene is more desirable for a controlled experiment.

Although the videos are simulated, we believe the result of this experiment should carry well to a real environment because of the following reasons:

(1) The motion and the details shown by the simulated videos are adequate for performing all the three tasks used in this experiment. We carefully set up the virtual environment so that all the generated videos provide enough visual cues for the user to perform the task in all the four conditions. All the trials were also tested in the pilot study.

(2) The tasks do not need absolute depth judgment. The simulated videos are designed to provide the major depth cues that exist in real videos: occlusion cues, motion parallax, depth from motion, perspective, relative size, and texture gradient. Since humans judge depth according to the weighted combination of cues [Landy, Maloney et al. 1995], the lacking of a few visual cues, e.g. aerial perspective, should not cause them to misjudge relative depth. Lacking certain cues may make it harder to judge absolute depth. Therefore, we avoided such situations when designing the tasks.

(3) We did not find any case in which using virtual videos instead of real videos caused any bias towards any particular navigation technique compared in this experiment.

The simulated environment was created in Blender, a 3D modeling tool. It has proper textures and lightings so that the video are not overly simplified and looks similar to a real video. 16 videos were used to cover a single-floor office environment, containing hall ways and open areas. The cameras were configured to have good coverage (over 95% of the open area and hall ways) of the monitored environment. A small amount of blind area (gaps between cameras) was allowed. 8 animated 3D characters with distinguishable visual features were used in the scene.

### **8.3.3 Tasks and Hypotheses**

This experiment contained three realtime tasks selected from the task taxonomy (Chapter 4.5).

#### **Task 1: Video level tracking**

The participants followed a target person through a dense and dynamic crowd of 8 people across 8-10 cameras. To measure how well the participants could keep track of the target person, they were asked to click the video to specify the target person's location on the video. As the participants saw the target appear in a video, they used the right mouse button to lay a marker on top of the videos to indicate that they saw the target in the video. User performance was measured by the number of errors they made. Right-clicking a wrong video or missing a video were both counted as one error. The target might be captured by more than one camera. The participants were instructed to "at least click the video showing a good view of the target, and try to click all the videos." A video shows a good view if it can show the target clearly, sufficiently large and for a long enough time for observation.

In this task, the participants did not need to form a comprehensive understanding of the path traveled by the target. Instead, their focus of attention was on monitoring the videos in focused video view, context view and peripheral view. When the target walks out of the field of view of the current video, the participants must scan the neighboring videos shown in peripheral view and context view. If they lost track of the target in these

views, they went back to the video bank, which showed an overview of all the videos in miniature.

We hypothesized that detailed view navigation designs would out-perform overview navigation conditions, as the users may not even need to look at the large context for this task. Between the two detailed view navigation designs, AD was likely to lead to better performance, as this task requires the users to constantly focus on the video content instead of the navigation process. To navigate to the next video, MD needs more cognitive effort than AD.

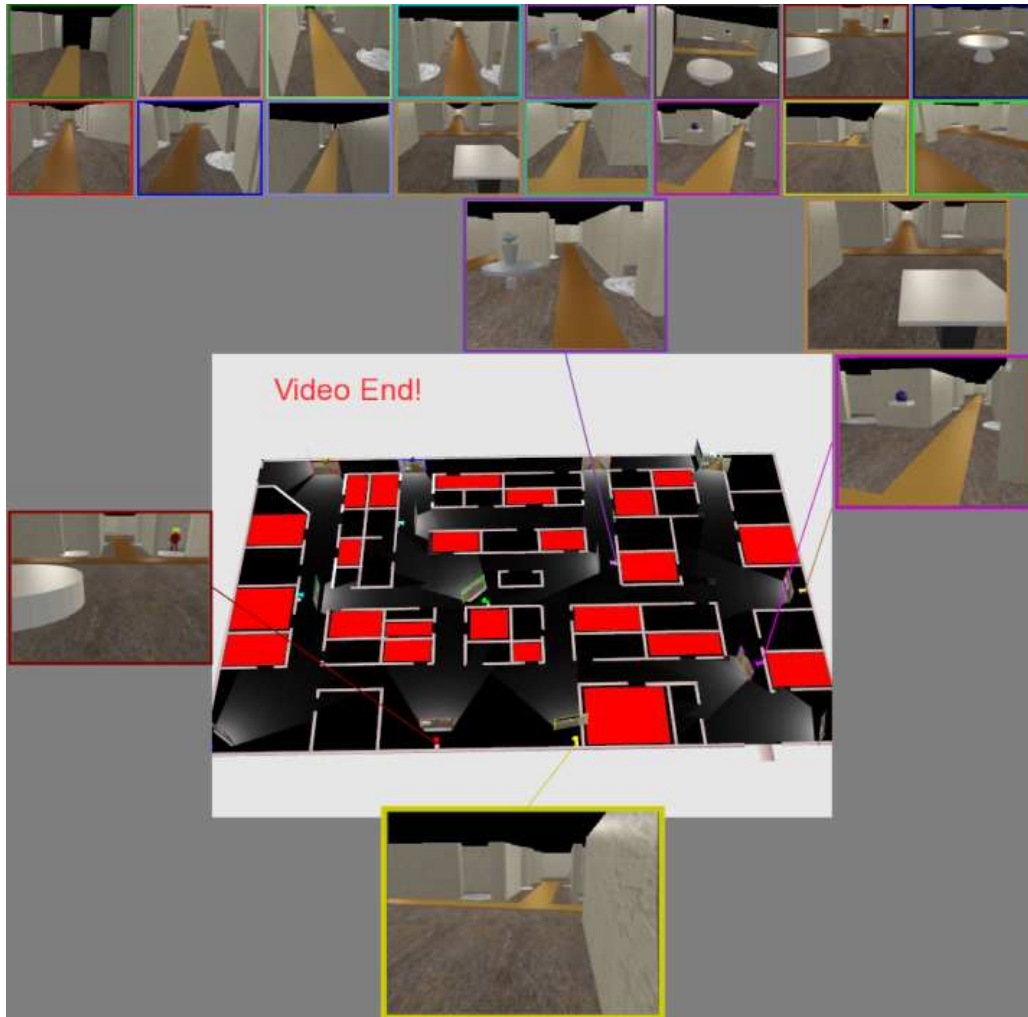
### **Task 2: Model level tracking: Procedural + Structural/abstract information task**

The target person walked around within the model. He visited 7-9 rooms on the path. All the rooms had only one door. All the processes of entering and exiting a room were captured by 6-8 videos. Some of the rooms were secured rooms, which were painted red on the model, but could not be distinguished from regular rooms on the video. There are similar real world situations, where the rooms with high security level may not be easily identified from the video. Often these rooms may just have a special tag on the door. But they can be visualized more clearly on the model.

Whenever the target entered a room, the participants had to say the color of the room. Only one character walked around in the scene, so the participants did not need to divide their attention on discrimination of the target. The number of correctly recognized rooms were counted and compared. Figure 8-3 shows the interface for Task 2.

This task is an example of a Type 1 integration task described in our task taxonomy (see Chapter 4.5.2 and Figure 4-12). It required the users to correctly observe which doors the target entered and to mentally map the doors shown in the video to the doors shown on the map. Therefore, the user needed to integrate the dynamic and procedure information learned from the video with the structural information from the model.

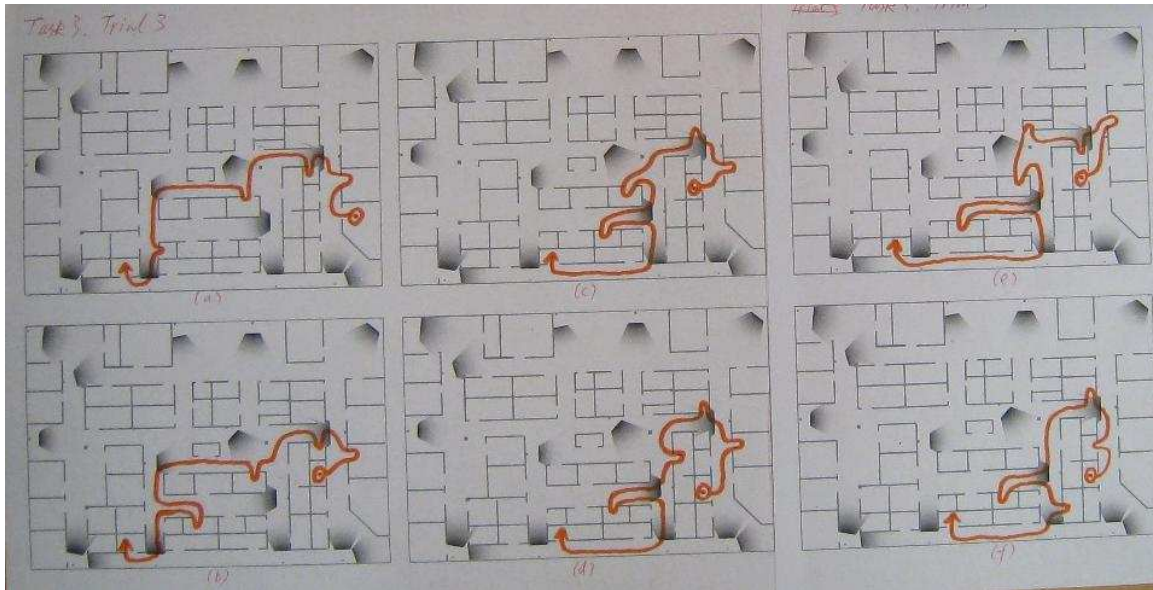
We hypothesized that detailed view navigation would out-perform overview navigation. Even though the users had to see the model, only the nearby context was needed. We thought MD might do better than AD in this task. In AD, when the videos occlude the model, the participants had to see the information in overview. In contrast, the user might manually shift to the left or right to look behind the video in MD. This within detailed view navigation may need less effort than zooming out to overview.



**Figure 8-3: A screenshot of Task 2 with AO technique.**

**Task 3: Model level tracking: Procedural + route information task**

Similar to Task 2, the users saw the target person visit 4-5 rooms and make 4-5 turns along a path. After the video finished, we showed the participants 6 maps with the path and the location rooms visited by the target labeled on the map. Only one of the 6 choices was fully consistent with the actual path. The participant had to select the correct choice. If they could not decide which was correct, they were asked to assign a possibility value to each choice. Those choices excluded because of the features remembered by the participants were given a possibility value of 0. An example map is shown in Figure 8-4. This task required the participants to follow the target through the whole path while observing and remembering the turns and rooms along the path. It was more challenging than the first two tasks.



**Figure 8-4: A sample map containing 6 choices. The path is drawn in color. Some choices differ in their overall shape and other choices differ in details.**

This task is an example of Type 2 integration task described in our task taxonomy (see Chapter 4.5.2 and Figure 4-12). The participant had to judge the route of the target according to the camera that the target entered and the target's position within that camera's view. He also needed to register the door that the target visited to a particular point along the route. Therefore, the user needed to register the dynamic and procedure information learned from the video with the route information learned from the model and the video.

We hypothesized that overview navigation is better for this task. Between MO and AO, I think MO might be better; because semi-automatic navigation in the model can cause disorientation and prohibit understanding of the path. If users are well-trained, manual navigation might exceed semi-automatic in both overview and detail view navigation, because they have less chance of becoming disoriented.

The difficulty of the three tasks was designed to increase from Task 1 to Task 3, which is consistent with the learning curve. By the time they reached Task 3, their skill in using the techniques also became better, allowing them to do Task 3 comfortably.

### 8.3.4 Procedure and Settings

16 participants (9 females) between the ages of 22 and 46 (mean 28.0) finished the experiment. All participants were Asian. All participants used computers regularly and none had worked in a video surveillance job. Two participants had video processing experience.

The interface was displayed in a 1024×1024 window on a 19-inch LCD display of resolution 1280×1024. The video thumbnails in the video bank were controlled at a proper size (120×90 pixels) so that the users could detect motions in the video, but could not clearly discriminate the target from distractions when, the target was far away from the camera. The size of context view is 586×458, focused video view 266×200, and peripheral video 200×150.

## **Spatial Ability Test**

People vary in their way of processing spatial information. This difference is often called spatial ability difference, which may affect their usage preference, interaction strategy and task performance for Contextualized Video applications. In order to understand the relationship between people's spatial ability and the above aspects of the experiment, all participants were asked to perform a 3-minute long Cube Comparisons Test from ETS [Ekstrom, French et al. 1976].

## **Questionnaires**

The participants filled out demographics questionnaires after the spatial ability test. This step was purposely planned after the spatial ability test in order to reduce the latter's priming effect on the experiment trials.

After each task, the participants also filled out post-task questionnaires to rate the four navigation designs with regard to "easiness to learn", "easiness to use" and "usefulness" at a scale of 1-4, with 1 being the best.

## **Experiment Trials**

At the beginning of each task, we first trained the participants to perform the task. Then they did 1 or 2 practice trials for each navigation technique, followed by 2 experimental trials. After they had used all the 4 navigation techniques, they filled out the post-task questionnaires. All the participants finished all the three tasks in sequence.

## **8.3.5 Results and Discussion**

The experiment used a randomized complete block design. Users are the blocks. Their sequence of navigation technique usage was counter-balanced using Latin square.

### **Task 1**

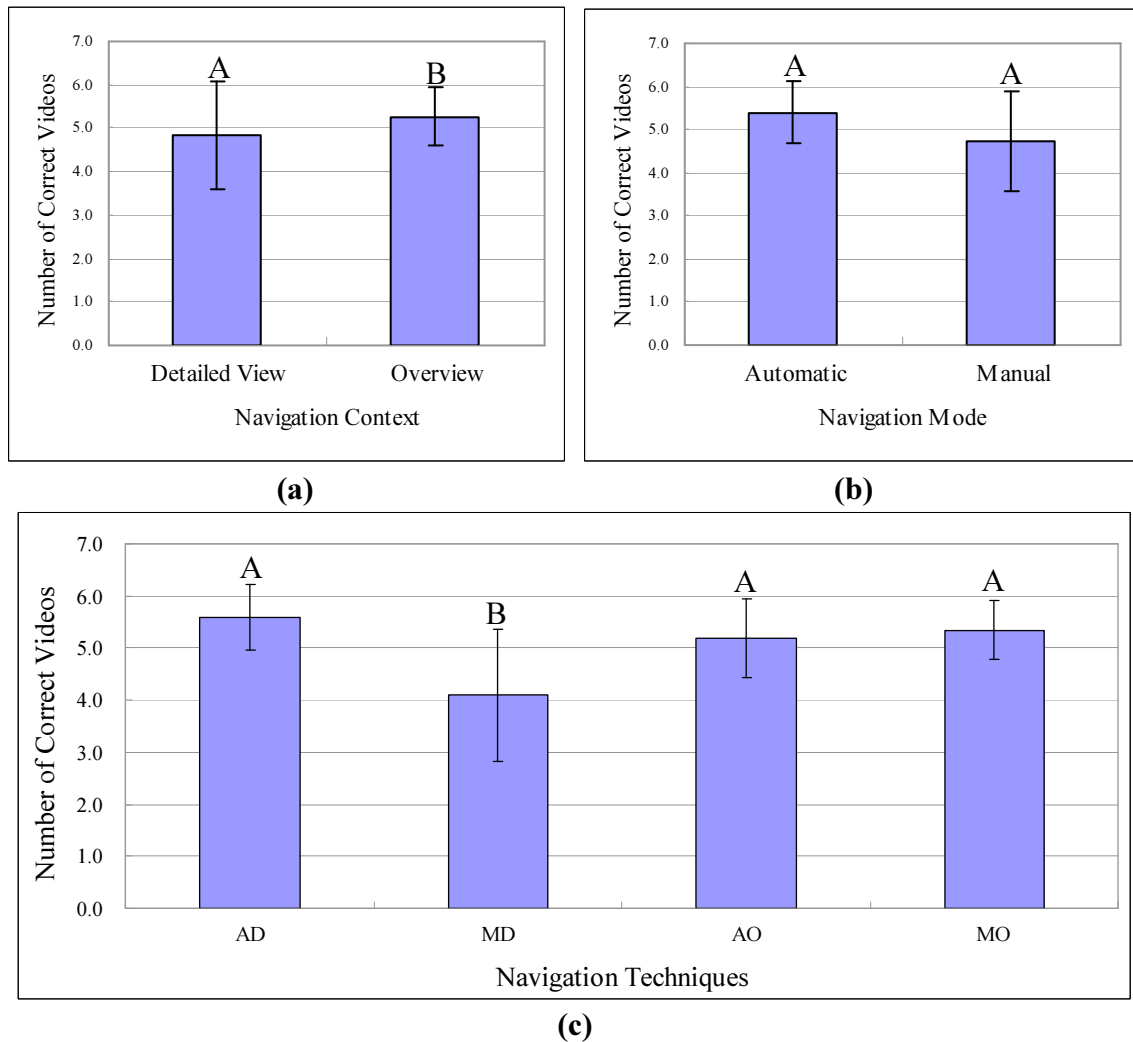
Since the task time was fixed (the task ended once the videos stopped playing), we measured the participants' task accuracy. The videos that were marked by the participants were recorded. By comparison with the correct answer, the number of missing or incorrectly marked videos was calculated for each navigation technique. Then the numbers are normalized according to the number of correct videos for each trial. As shown in Figure 8-5, two-way factorial ANOVA showed significant interaction between the two factors ( $F(1,60)=15.13, p<0.001$ ). Navigation mode (manual or semi-automatic) had a significant effect on the correctness of the marked videos ( $F(1,60)=9.96, p<0.01$ ). Semi-automatic navigation led to better video level tracking performance. Navigation context had an almost significant effect on task accuracy ( $F(1,60)=3.93, p=0.0521$ ).

We hypothesized that detailed view navigation designs would out-perform overview navigation conditions. Between the two detailed view navigation designs, we hypothesized AD would lead to better performance, as MD needs more user control than AD.

*Post hoc* analysis showed a somewhat different result: MD led to significantly more errors than others ( $p<0.001$ ). It is the poor performance of MD that caused the main effects of the two factors.



According to the participants' ratings and onsite observation, MD is the hardest to learn and hardest to use for most users. They often lost the target during the tracking process, due to losing control of the navigation. Therefore, this finding is not surprising.



**Figure 8-5: Task 1 result: the mean number of correctly selected videos in Task 1 (normalized among trials). (a) Main effect of navigation context; (b) Main effect of navigation mode; (c) Interaction effect of navigation context and navigation mode.**

Task performance in the MD condition also had the largest variation. This can be explained by the user strategy in this condition. Participants who tried to manually navigate in the context view often got lost in the middle and missed some videos. Since this task did not require the participants to understand the target's location in the spatial context, more than half of the users did not navigate in the context view at all. They depended on the static context view, the peripheral view, as well as the video bank to find the target. Recorded clicking patterns (see Table 8-2) indicate that participants clicked the video bank more often in the MD condition than in any other conditions (twice as often as in the AD condition). Since videos in the video bank are small and hard to perceive, users preferred to observe video in the other three views. Users in the MD condition got



lost more often than other conditions, so they had to look at the video bank to reacquire the target back.

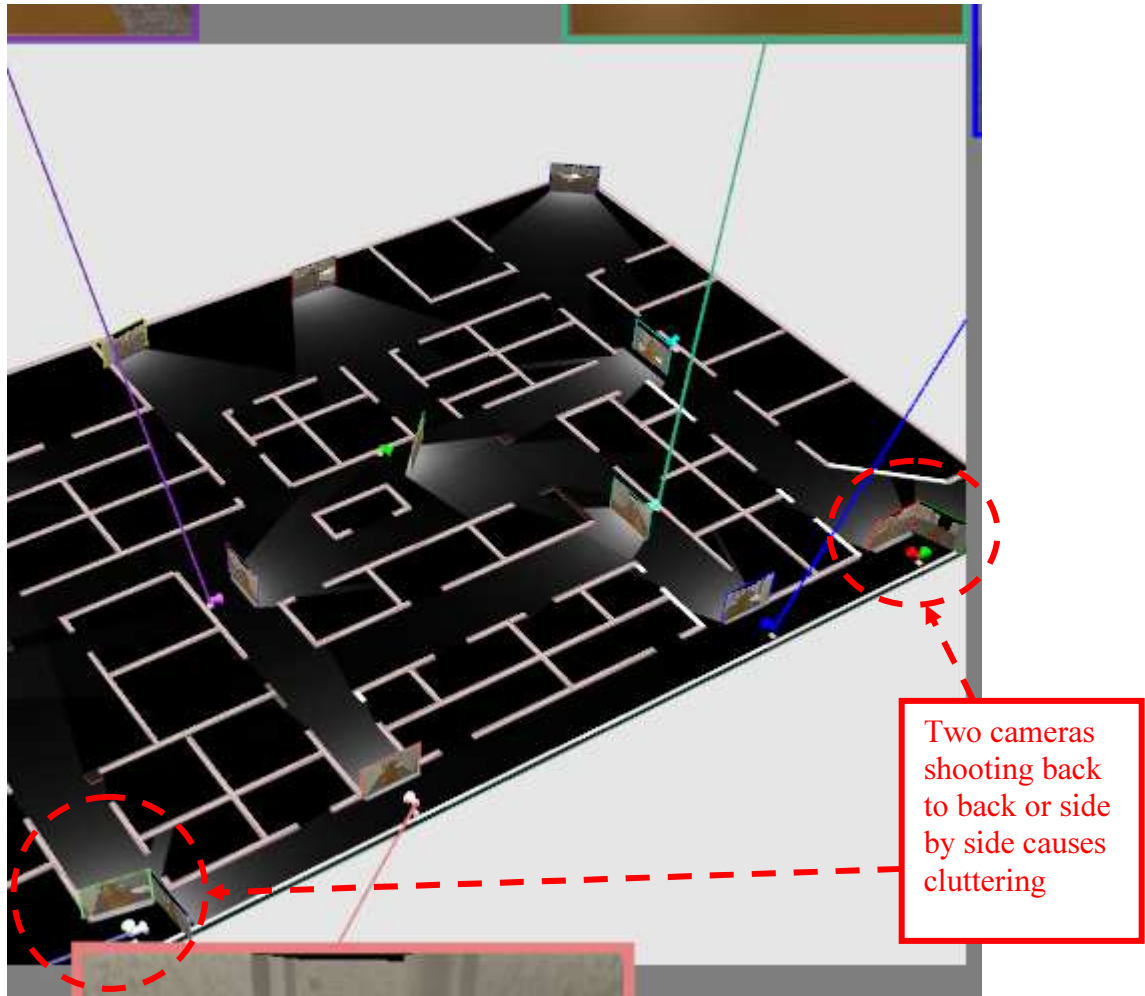
While *post hoc* analysis between AD, AO and MO did not show significant differences, AD led to relatively fewer mean errors than AO. Although the difference is not significant ( $p=0.1532$  for Tukey-Kramer test between AD and AO), we can see a trend. This is consistent with our initial hypothesis.

Recorded clicking pattern (see Table 8-2) indicates that participants clicked the video bank 42% more often in the AO condition than in the AD condition, and clicked the peripheral videos 44% more often in the AO condition than in AD condition. These facts indicate that the participants lost the target more often in the AO condition. When they lost the target, they had to either find it in the video bank or click one of the peripheral videos to bring up its neighboring videos.

One explanation for the performance difference is that the context overview of AO cues the participants to use a strategy that led to more error. The overview shown in the context view clearly shows the videos' spatial relationship, which gives the user a strong cue to try to find the next video from this overview then follow the callout lines to find the corresponding video in peripheral view. Unfortunately, the videos shown in the context view in the AO and the MO condition are small, consequently hard to find and click. I observed that some participants selected the wrong video where two cameras were shooting back to back. Unlike AO, the detailed view in AD does not show all the neighboring videos in the context view. Therefore, the user directly relied on the peripheral view to find videos.

Task	Tech	Video Bank Select	Context View Select	Peripheral View Select	Total
1	AD	35	27	113	175
1	MD	69	67	64	200
1	AO	50	34	163	247
1	MO	55	42	130	227

**Table 8-2: The total number of left mouse clicks on three views of Contextualized Video interfaces recorded in Task 1.**

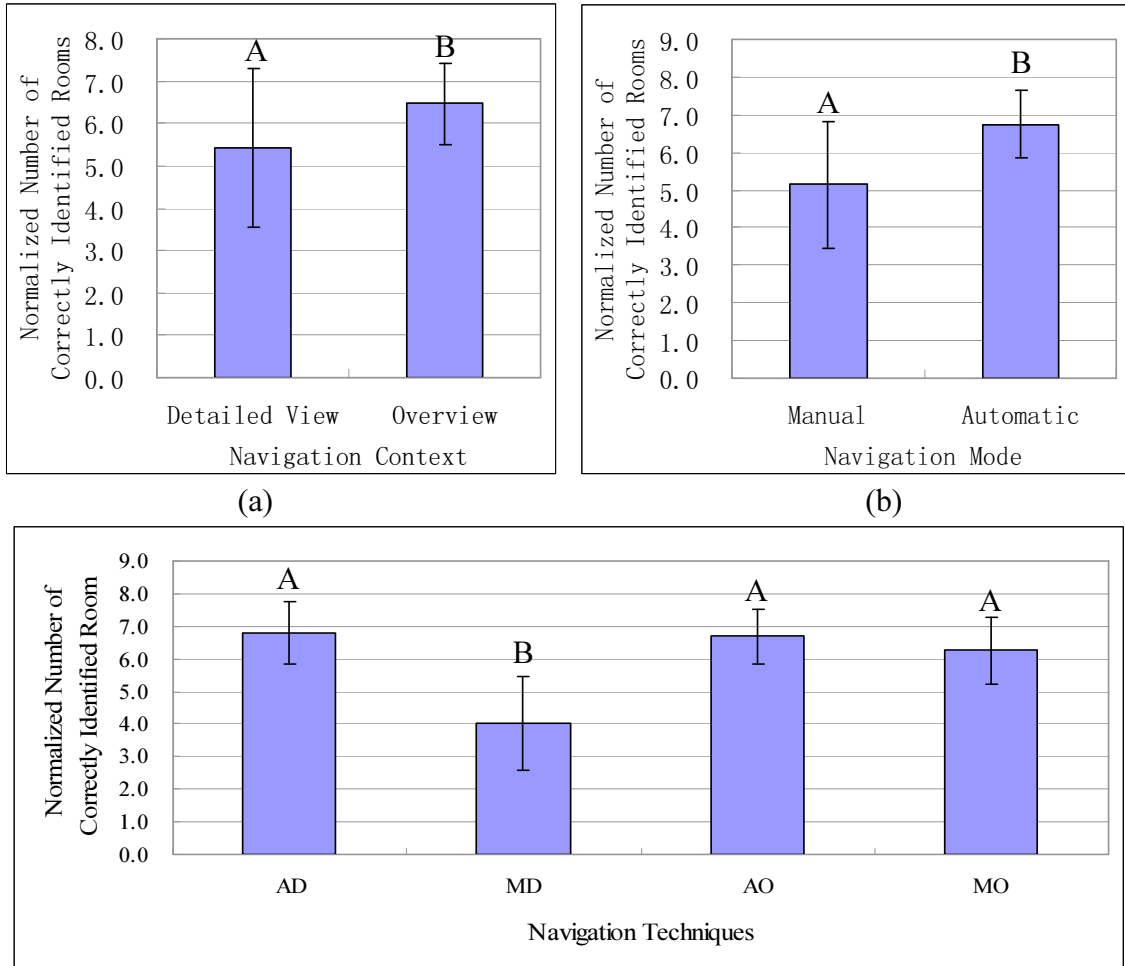


**Figure 8-6: Video and camera cluttering problem in AO condition.**

### Task 2

Task 2 measured the number of correctly identified rooms. The numbers are normalized according to the total number of rooms visited in each trial. If a room was not identified or incorrectly identified, it was treated as incorrect answer. As shown in Figure 8-7, two-way factorial ANOVA shows significant interaction between the two conditions ( $F(1,60)=18.32, p<0.0001$ ). Both navigation mode and navigation context had a significant effect on the number of identified rooms (navigation mode:  $F(1,60)=34.87, p<0.0001$ , navigation context:  $F(1,60)=14.79, p<0.001$ ).

When participants were using the MD condition, they correctly identified significantly fewer rooms than while using the other three conditions ( $p<0.0001$ ). The differences between other pairs of conditions were not significant. Therefore, the main effects of the two factors are caused by the poor performance of the MD condition. We hypothesized that the user might manually shift to the left or right to look behind the video in MD, this did not happen very often because the participants did not have enough time to navigate to look behind the camera. Navigation control was still the bottleneck for participants without extensive first-person shooter game experience.



**Figure 8-7: Task 2 result: the mean number of correctly identified rooms (normalized among trials). (a) Main effect of navigation context; (b) Main effect of navigation mode; (c) Interaction effect of navigation context and navigation model.**

We hypothesized that detailed view navigation would out-perform overview navigation, as detailed view showed the nearby context more clearly. The result did not show such trend. Although the participants used different strategies under the other three conditions, their performance did not differ much. Being able to align the video and the model, and to see the model without occlusion are both important factors for the success of this task. AD might have allowed easier alignment, but the occlusion between the video and the model was more severe than the AO and MO condition. Users who hid the videos of context view achieved good performance in this task.

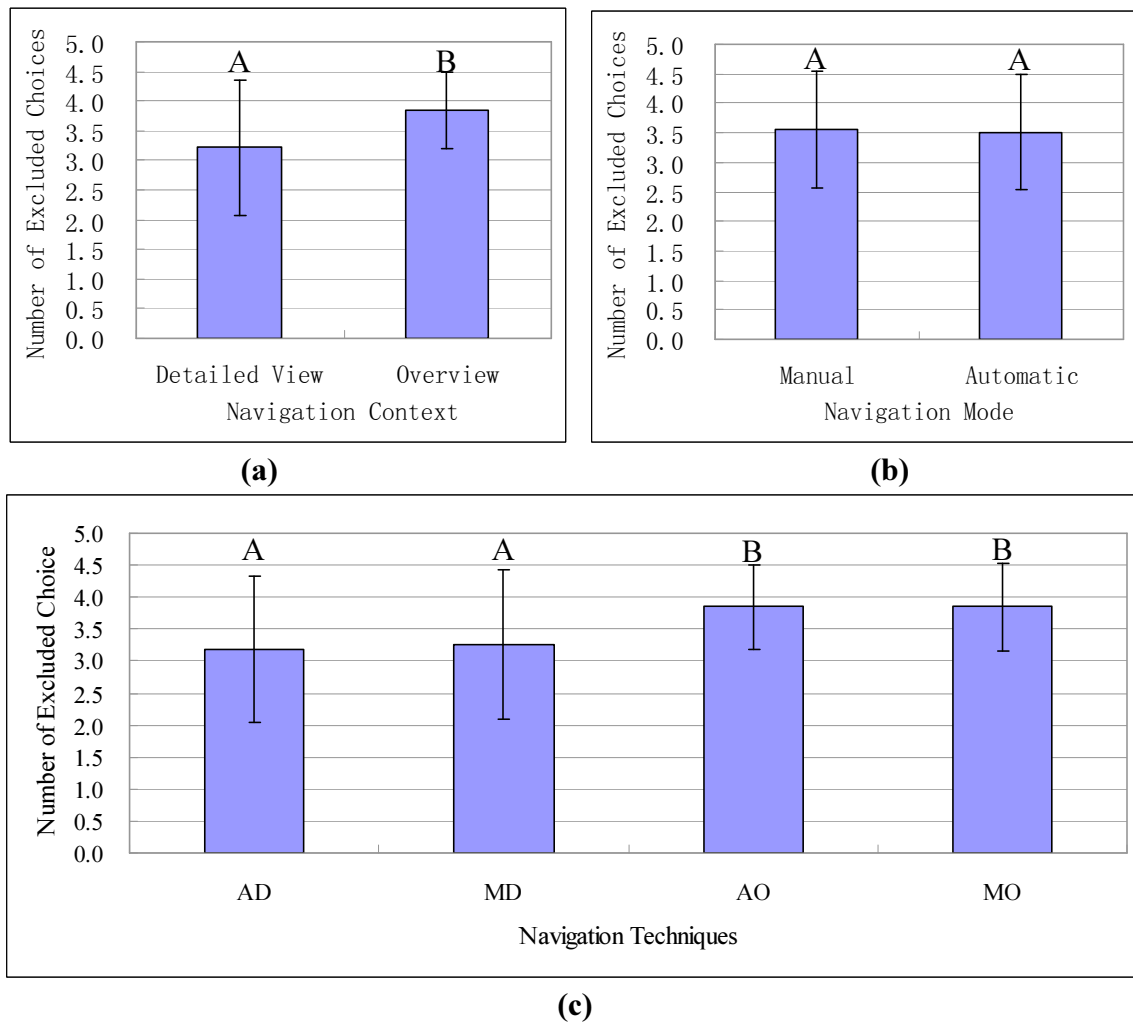
AO gets absolutely the highest preference rating, followed by AD and MO. MD is the least preferred condition. This rating is reasonable for Task 2, as AO has less occlusion than AD and easier control than MO and MD.

### Task 3

Taking the mean number of excluded choices as the criterion, linear model ANOVA showed significant difference between navigation in detailed view and navigation in overview ( $F(1,60)=6.99, p<0.05$ ). Tukey's HSD test showed that AO/MO led to higher

task accuracy than AD/MD (Figure 8-8). Consistent with our hypothesis, overview conditions had an obvious advantage for this task. Even under AD and MD conditions, the participants often zoomed out to see an overview while the video played. Overview not only shows a larger context, but also reduces occlusion between the video and the model. The participant's subjective rating is also consistent with the finding: they preferred the two overview designs over the two detailed view designs.

We initially used the number of correct answers (amortized by the reported confidence levels) as the performance criterion. But because the number of trials is too small, the variation turned out to be quite large. Since we asked the user to exclude as many choices as possible using the features they remembered along the path, the number of choices they excluded is a quantitative indication of how well they learned the route.



**Figure 8-8: Task 3 result: the mean number of excluded choices (6 choices in total). (a) Main effect of navigation context; (b) Main effect of navigation mode; (c) Interaction effect of navigation context and navigation mode.**

Overall, neither user performance nor user preference showed significant difference between automatic and manual navigation. We hypothesized that semi-automatic navigation in the model can cause disorientation and prohibit understanding of the path.

In reality, about half of the participants preferred AO over MO. Because the automatic rotation reduces interaction overhead, they felt that the autorotation without translation in AO does not cause as much disorientation as in the AD condition, which translates and rotates at the same time. However, some other users felt differently. They preferred to manually rotate the model in order to align it with the video. User preference is not correlated with their spatial ability or gender. Therefore, we believe that allowing low degree of freedom (DOF) control, e.g., 1 or 2 DOF, may be preferred by some users.

Unlike Task 1 and Task 2, the performance under the MD condition was not worse than under the AD condition in Task 3. Some participants even reported MD as more useful than other techniques for this task. They felt that having full control over the navigation process helped reduce disorientation. The relative performance difference of MD between tasks might also be due to the longer learning time of MD. In general, manual navigation took longer to learn. The more degrees of freedom in control, the harder it is for the user to form an effective working strategy. Many participants still felt frustrated when trying to navigate to look behind the camera, even after three practice trials. For the other three techniques, they normally needed less than two practice trials. Since Task 3 was always the last task to perform, the participants learned to use this technique better. Many participants reported this feeling when performing the practice trials.

Overall, MD is still the least preferred technique, apparently because of the control overhead. Users' ratings with MD are correlated with their spatial abilities and 3D game experience. Two users with extensive 3D shooting game experience did especially well with the MD condition. Some users only pan and zoom when using the MD condition. They felt it to be too difficult to look behind the target camera while, at the same time, translating and rotating. So they performed mental rotation instead of physical rotation in the context view.

## **Correlations**

Spearman correlation analysis showed that Task 2 performance is strongly correlated with user's spatial ability test score ( $p < 0.05$ ). However, the correlation is less strong for Task 1 ( $p = 0.38$ ) and Task 3 ( $p = 0.18$ ). This difference can be explained by the amount of spatial mapping needed for each task. Task 2 requires precise spatial mapping between the video and the model. Task 1 does not require precise video-model mapping. Task 3 allows other strategies to bypass video-model mapping.

Since a strong correlation was found between spatial ability score and Task 2 performance, we added spatial ability as a created variable with two levels (used a mean split on the spatial ability score, because the distribution was not normal) and performed general linear model ANOVA on the data again. We found that spatial ability has a significant effect on Task 2 performance ( $F(1, 14) = 5.15$ ,  $p < 0.05$ ). But the interaction between spatial ability and navigation mode ( $p = 0.23$ ) or navigation context ( $p = 0.96$ ) is not strong. No significant effect was found for Task 1 and Task 3.

Gender was found to be correlated with Task 3 performance ( $p = 0.062$ ) but not Task 1 ( $p = 0.91$ ) and Task 2 ( $p = 0.20$ ). In general, males did better than females. This cannot be explained by the correlation between gender and spatial ability test score.

Task 2 performance is strongly correlated with users' prior gaming experience ( $p = 0.02$ ). The other two tasks are not strongly correlated ( $p = 0.15$  for task 2 and  $p = 0.20$  for tasks 3).

## 8.4 Guidelines

The evaluation results indicate the following design guidelines:

### 8.4.1 Design According to Task Characteristics

**Guideline 4.1: If the task focuses on video level tracking, and no spatial understanding is needed, then choose semi-automatic navigation in detailed view.**

Showing an overview of the context may cue the user to judge the camera's relationship from the overview, which may pose a higher cognition workload. This is likely the reason that caused the performance difference between AD and AO on Task 1 in our experiment.

**Guideline 4.2: If the realtime task requires precise spatial mapping between video and model, especially from the camera's point of view, then semi-automatic navigation is preferred.**

The interaction cost of manual alignment might reduce task performance, particularly if the manual control involves both translation and rotation. Although disorientation was a reason for some users to give semi-automatic navigation a low preference, users' performance indicated that its low interaction cost outweighed the disorientation problem. Furthermore, several methods can be used to reduce the disorientation caused by semi-automatic navigation: (1) Avoid translation and rotation at the same time. (2) Always keep the starting point and destination in the view. (3) Visualize the navigation process in an overview.

**Guideline 4.3: If the task is realtime route or procedural learning, then showing a consistent overview that contains the whole route or procedure is preferred.**

Our experiment found two strategies to use an overview during this type of tasks:

(1) Mainly focus on detailed view, and zoom out on demand to mentally register the details to overview. The main overhead of this strategy is to hold an overall picture of the route or procedure in their mind. This strategy was used in AD and MD condition.

(2) Mainly focus on overview, and zoom in on demand to do some tasks requiring more details. In this case, the user needs to specify the spot to zoom in, and may also need to specify the view angle. This strategy was used in AO and MO condition.

Strategy (2) seems to be more effective according to Task 3 results in our experiment. Even in AD and MD conditions, multiple users mainly kept to zoomed out view, and only went back to detailed view for navigation.

### 8.4.2 Design According to User Characteristics

**Guideline 4.4: Between semi-automatic navigation and manual navigation, users may have quite different preferences. Support both techniques if possible.**

We found that the preference between semi-automatic and manual navigation is not correlated with the users' spatial rotation ability. Therefore, we believe this is a personal preference.

**Guideline 4.5: Pure manual navigation with more than two DOF is not recommended for non-expert users.**

Our experiment showed that manual navigation took a longer time to learn in general. The more degrees of freedom in control, the harder it is for the user to form an effective work strategy. According to our evaluation experience, for a technique with more than 3 DOF control, hours--even days--of practice should be expected before the user can form an effective work strategy. In this experiment, many participants still felt frustrated when trying to navigate to look behind the camera even after one hour of practice and usage.

**Guideline 4.6: If manual navigation is necessary and the user has no prior experience of the navigation technique, we recommend giving the user enough time to find a good usage strategy through both guided training and free practice before evaluation of the technique.**

Free practice gives the user the chance to discover a comfortable strategy for himself. The purpose of guided training is to inform the user about the strategies adapted by other users. Combining both approaches, the users have a better chance of finding a good usage strategy that can demonstrate effectiveness of the technique.

### 8.4.3 Other Guidelines

**Guideline 4.7: Peripheral videos are helpful for time-critical video-level tracking and video-model integration tasks.**

As shown by the clicking patterns, peripheral videos were used more frequently than any other views for selecting of focal videos. Girgensohn et al. showed that peripheral videos in a 2D interface support video-level tracking better than regular video banks [Girgensohn, Shipman et al. 2007].

**Guideline 4.8: Avoid complex navigation containing both translation and rotation; particularly avoid such navigation if the landmarks cannot be seen in the intermediate view. This is true for both manual and semi-automatic navigation.**

Even though translating and rotating at the same time can make the navigation faster, it may cause severe disorientation if the users cannot see the landmarks in the intermediate view. The landmarks include the starting point, the destination, and any other salient features that can be referenced by the user to judge spatial relationships.

**Guideline 4.9: Reduce the number of videos in the peripheral view when animating the videos. Users may lose track of the target shown in the peripheral view because of severe visual clutter caused by many videos moving around.**

In the designs used in our experiment, the videos in the peripheral view animate to accommodate the navigation in the context view. We found that it is much easier to track the target video if we hide the other videos and only animate the previous video in examination and the currently selected video. Since the users' attention was focused on only these two videos, this design can help the users to manage their attention by hiding unused information.

## 8.5 Conclusions

Summarizing the design practices and experiment results, we believe semi-automatic navigation is useful for most Contextualized Video tasks. There are multiple possible ways to reduce disorientation in semi-automatic navigation: (1) Avoid translation and

rotation at the same time. (2) Always keep the starting point and destination in the view. (3) Visualize the navigation process in an overview. Following methods (2) and (3), we prototyped an overview visualization in Chapter 9 to help users understand the navigation process.

High DOF (degree of freedom) manual navigation is hard to learn and also inefficient for global navigation, but it can possibly be combined with semi-automatic navigation to provide more flexibility for examining a local view. Investigating the combination and the mode change between semi-automatic and manual navigation is a promising research direction.

The choice between navigation in overview and detailed view is task dependent. As summarized in Guideline 4.3, if the task requires the user to understand the whole, then an overview of the whole scene is desirable and navigation in overview is preferred. This is supported by Task 3 results. If the task only requires the user to examine details, then detailed view is enough. This is supported by Task 1 results and summarized in Guideline 4.1. Overall, overview navigation seems to be able to support both tasks better than detailed view navigation. However, if the whole scene involved in the task covers a very big area or multiple floors of a building, showing the whole overview would make some elements on the overview too small to be identified or selected.

We did not investigate the effectiveness of supporting navigation at all level of detail in Contextualized Videos, because it involves too much user strategy, which may lead to large variance in the performance data. We leave it as future work.

Considering the interaction between the two factors, we think it is inherently easier to navigate in overview because route planning is easier in overview, which shows all the information required to reach a destination. The MO in our experiment also used fewer DOF (degrees of freedom) control than MD condition. Therefore MO is much easier to learn than MD in our experiment. However, MO does not necessarily have fewer DOF than MD. If both view angle and view point need to be specified in MO navigation, then the MO navigation will need four DOF control, which is the same as the MD navigation used in this experiment.

Focusing on the navigation dimension of the design space described in Chapter 4, this chapter designed and evaluated Contextualized Video designs using the tasks from our Contextualized Video task taxonomy. We have now finished the last of the four research cycles planned in our research approach (Section 1.7). The formal experiment findings addressed research question Q4 (What are the effects of various Contextualized Video designs on the performance of key Contextualized video tasks?).

These guidelines are summarized for desktop displays. If a designer is designing for a Contextualized Video interface for body-sized or wall-sized large displays, these guidelines may not be directly applicable. Multiple challenges need to be considered: (1) the peripheral videos may take significantly longer time to scan as their angular distance with regard to the user increases. (2) The user may not be able to see an overview of the model without body or head movement.



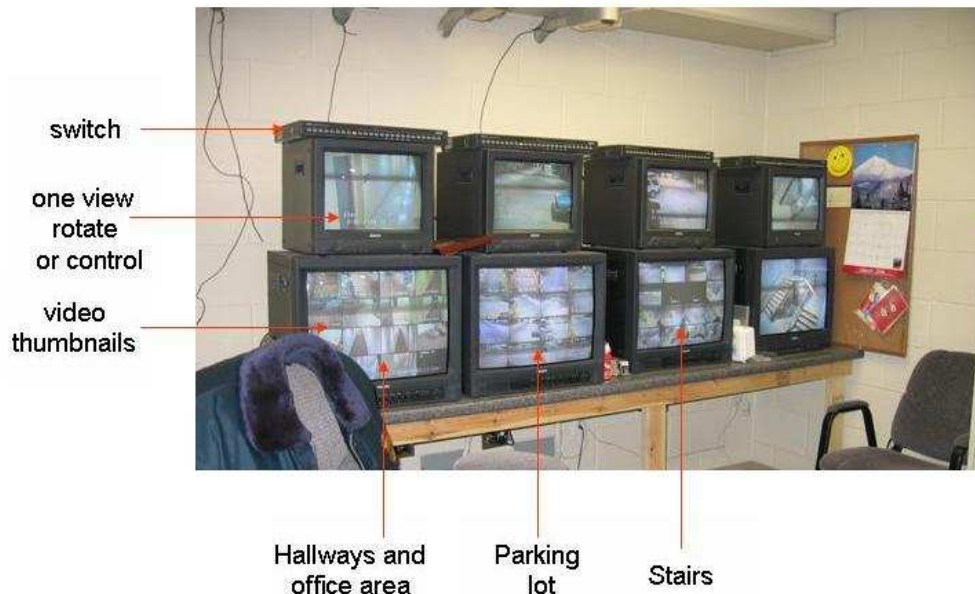
## 9 Application of Guidelines

The design guidelines summarized in this dissertation are intended to help visualization designers understand the tradeoffs between different design choices. The guidelines also provide a framework for understanding the tasks at hand and for creating proper interfaces based on the designs evaluated in this dissertation. However, visualization design is not a pure engineering problem [Shneidermann 2003]. Guidelines do not automatically lead to a good design. The designer also needs to successfully apply other design factors, including creativity, design aesthetics, knowledge of perceptual psychology, and a deep understanding of the users' work flow. This chapter demonstrates how to utilize the design guidelines in concert with other factors to create a new interface for a real world building security surveillance application described in the field study chapter (Chapter 3).

### 9.1 Problem Scenarios

Following the scenario-based development (SBD) framework introduced by Rosson and Carroll [Rosson and Carroll 2002], the design practice starts with the construction of problem scenarios that can help me discover the problematic aspects of the current system. These problem scenarios should resemble the typical building security tasks I summarized during the field study. Therefore, the following scenarios are constructed:

#### 9.1.1 Monitoring and event handing



**Figure 9-1: The current video surveillance system used in Building A.**

JB is an experienced security guard. During a usual day, JB monitors the whole 5-story mixed-use building using a surveillance system with 55 cameras, as shown in Figure 9-1. He had been monitoring these videos for half an hour. He glanced at the

videos every few seconds. To keep himself at a high vigilance level, he had been drinking coffee and listening to music.

He suddenly noticed one male with a red shirt on the top floor of the parking area smoking and talking over the cell phone. JB enlarged that video on the top monitor and took a close look at this target. JB did not recognize him. He might not be one of the residents of this building. So JB decided to keep an eye on him. After 5 minutes, a male with a white shirt appeared in another video. JB took a close look and did not recognize him either. JB decided to follow the male in the white shirt. He enlarged this video on the top monitor. He still has to keep an eye on the male in the red shirt in another video.

Since there are some blind areas between surveillance cameras, whenever the male in white exited from one video, JB had to imagine where he might be and in which video he might appear next. He was not sure which entrance the male in white had initially used to enter the premises. He might have missed a video, but he had no time to check now.

The two observed subjects finally met on the parking lot. They started to talk while walking around, and entered a corner that can not be observed by any camera. JB tried to determine which were the nearby videos, and decided to look at three videos on the monitor. Since these three videos are at different monitors, he had to look back and forth between them. After 5 minutes, JB had still not seen these two subjects again in any video. He decided to go to the site and see if they might be engaged in some mischief.

Now he needed to plan a route to go there. JB has to consider not only the time to arrive there, but also the chance of missing them if they were to leave the building from any exit. JB thought about the tradeoffs for a few seconds and picked a route that would allow him to observe the two subjects from faraway if they attempted to exit the building through a nearby emergency exit.

Upon arrival at the site, he found the two subjects drinking beer at the corner. Since JB did not find them stealing anything, he politely but firmly escorted them from the building.

**Claim 1:** Viewing only videos:

- + keeps the display easy to understand
- requires the user to have excellent mental models of the spatial context. A spatial context display is desirable when the activity to be analyzed is complex.

### 9.1.2 Creating a daily report

JB wrote a daily report to summarize tonight's suspicious events. He wrote:

“Two guys hang around on top floor of the parking lot between 11:30 pm and 12:00 am. I talked them away at 12:05am. Please take a look at the video. ”

“Office 409's door was open when I patrolled at 9:30 pm. No one was in it. I locked the door. I looked at the video and found that a tall male with glasses and black hair was the last to exit the office at 8:48 pm. He seems to be working in this office. Please check with them during the day.”

The property manger came in during the day, read these reports and went to watch the video. It took her 20 minutes to find the proper video segments and finish watching it.

**Claim 2:** Annotating and retrieving of videos by hand:

- + allows the user to freely pick any video segment to annotate and watch
- is time consuming, can be automated by dragging the video onto a timeline

- lacks the spatial and time context of the event, can be annotated within a spatial context and timeline give.

### 9.1.3 A security guard's first week in the building

Building A currently has only one building security guard, JB, who worked the night shift. The property manager for Building A decided to hire another building security guard, SH, to work on weekends.

JB walked SH around the building and told him the camera settings and the coverage areas. SH then came back to the office and looked at those videos, trying to register the videos to their actual location. It took about a week before SH was able to form a precise understanding of the placement of all the surveillance cameras and their spatial relationships.

JB also told SH what locations he should pay particular attention to, as well as the working habits of the residents of the building, etc. Since there is no record of such knowledge in the surveillance system, SH later forgot some of this.

**Claim 3:** Learning the camera configuration of a site by walking through:

- + Can get detailed knowledge of individual cameras and first-person view of the site
- Lacks an overview of the whole site, need to form the survey knowledge through long term walking through the real site. A spatial context display with the camera's configuration clearly labeled is desirable for learning.

**Claim 4:** Transferring informal and tacit knowledge person to person:

- + Is a fast and intuitive way to transfer knowledge.
- Hard to guarantee that all knowledge is successfully transferred. Need a way to easily record and transfer informal and tacit knowledge.

## 9.2 Interface Design

Four claims were made regarding the three problem scenarios. The downsides of the claims can be viewed as the requirements for the new system. We will propose a new Contextualized Video interface that fulfills these requirements. To design the new interface, we first analyze the tasks involved in the problem scenarios, then look for design guidelines according to the user and task features, and finally construct the interface according to those guidelines.

I suppose the building I visited in the field study has updated their CCTV system to an IP solution and the system can automatically detect motions in videos. As discussed in Chapter 3, IP based surveillance systems are the future trend of the market.

### 9.2.1 From Problem Scenarios to Tasks

Analyzing the monitoring and event handing scenario in Section 9.1.1 using our task taxonomy in Section 4.5.2, the following tasks can be extracted:

**Task 1:** Monitor all the videos, detect motions and suspicious behaviors. This is apparently a video intensive task in our task taxonomy.

**Task 2:** When the security guard finds a subject in the video, he wants to enlarge the video to observe closely to see if the subject is a resident of the building and whether the subject's behavior looks suspicious. This is also a video intensive task. More specifically, the user wants to acquire details and procedure information from the video.

Once the security guard noticed that the subject exhibited suspicious behavior, he started to follow the target from one video to another, in order to learn which places the subject visited, understand the subject's activities, and even to judge his intention from his behaviors. The tasks involved in this process are mainly video-model integration tasks, as the security guard needs to think of the subject's behavior in a larger spacial context.

**Task 3:** When the subject disappeared from video, the security guard had to judge where the subject is in the model (a mental model if the model is not shown on an external display), at the moment he walked out of the camera's coverage area. This task involves integrating dynamic information of the subject in the video with the structural information (the camera's location and coverage area) in model, hence is a Type 1 integration task in the taxonomy.

**Task 4:** If the security guard had noticed the subject holding a laptop, he would need to think which room the subject visited and what he did in those rooms. In this case, the security guard has to recall the previous path of the subject on the model (route information) as well as his behavior (procedural information) observed from the video. This is a Type 2 integration task in our taxonomy.

**Task 5:** If the security guard decides to expel the subject from the building, he needs to plan a route to reach the subject. Route planning is a model intensive task in the task taxonomy, as he already knows the subject's location in the model and no longer needs to look at the video.

There might be other tasks in this problem scenario. We listed the representative ones for demonstration purposes.

## 9.2.2 From Tasks to Guidelines

Having those tasks in hand, we go over the guidelines summarized in Section 5.4, 6.2, 7.2 and 8.4. We list the guidelines here if the task and user characteristics match the guidelines:

### **Guideline 1.1: Emphasize videos for tasks that mainly require video information**

Since Task 1 requires motion information from possibly any video, then an overview display of all the videos should be presented.

However, monitoring (vigilance task in psychology) is a very boring task. Humans' vigilance levels often decrement sharply after 30 minutes of watching [Mackworth 1948] [Harris and Chaney 1969]. Since automatic motion detection is technically feasible, automatic motion detection can be used when activities are infrequent. Therefore, we plan to have an overview of the videos from which motions are newly detected.

Task 2 requires detailed information from the video, hence an enlarged view should be provided on the interface. As stated in the problem scenario, the security guard wants to keep an eye on the overview videos while closely observing a particular subject. Therefore the enlarged view should not occlude the overview, and should be put close to the overview to support multi-tasking.

### **Guideline 1.4: Use embedded designs for time-critical Tasks**

### **Guideline 1.5: Use combined designs to allow flexible user strategy**

Since all the tasks listed above are time-critical, we choose to use embedded design. Also, as the user is an expert, we also use associated design to allow flexible user strategy.

**Guideline 1.9: Provide rich context cues for users who are familiar with the environment**

Since the security guard is very familiar with the building, we can visualize the cues listed under the guideline. Even the residents' photos can be linked to their offices on the model, so that when the security guard clicks a room, the photos the residents will pop-up, enabling the security guard to quickly judge whether the subject who just visited the room is a resident or not.

**Guideline 1.3: Maximize the videos and their nearby context for integrative tasks**

**Guideline 2.2: If the task requires precise spatial judgment, and the model shows plenty of landmarks, try both user-aligned and camera-aligned video layout**

**Guideline 4.2: If the realtime task requires precise spatial mapping between video and model, especially from the camera's point of view, then semi-automatic navigation is preferred**

**Guideline 4.7: Peripheral videos are helpful for time-critical video-level tracking and video-model integration tasks**

Task 3 and 4 are integrative tasks.

Task 3 requires precise spatial judgment. Also, we will provide many landmarks according to Guideline 1.9. Therefore, we should follow Guideline 2.2 to provide both user-aligned and camera-aligned video layout. Following Guideline 4.7, peripheral videos should also be provided. For navigation, Guideline 4.2 suggests semi-automatic navigation.

**Guideline 4.3: If the task is realtime route or procedural learning, then showing a consistent overview that contains the whole route or procedure is preferred**

Task 4 is a realtime route learning task. According to Guideline 4.3, we decide to provide an overview of each floor together with the videos that monitor this floor.

**Guideline 1.2: Emphasize model for tasks that mainly require contextual information**

Task 5 is the route planning task listed under this guide line. Therefore, when users perform this task, we want to show an overview of the model, assign more physical display space for the model, put the model at the center of the display, etc.

### **9.2.3 From Guidelines to New Interface**

Those guidelines define the important features of the interface. These guidelines suggest a multi-view, highly dynamic interface to support the workflow composed of multiple tasks. The workflow can be represented as:

- monitor all videos → closely observe a video**
- judge the subject's location on the model**
- understand the path and behavior**
- plan a route**

Now we can convert the workflow to visualization requirements, according to the guidelines:

- video overview → video detail**

- video with local context
- model overview with related videos
- model overview

Having this dynamic display requirement in mind, we start to design the layout of the display, which should directly support this workflow by providing the proper context, depending on the information requirement of each task and a smooth transition between tasks.

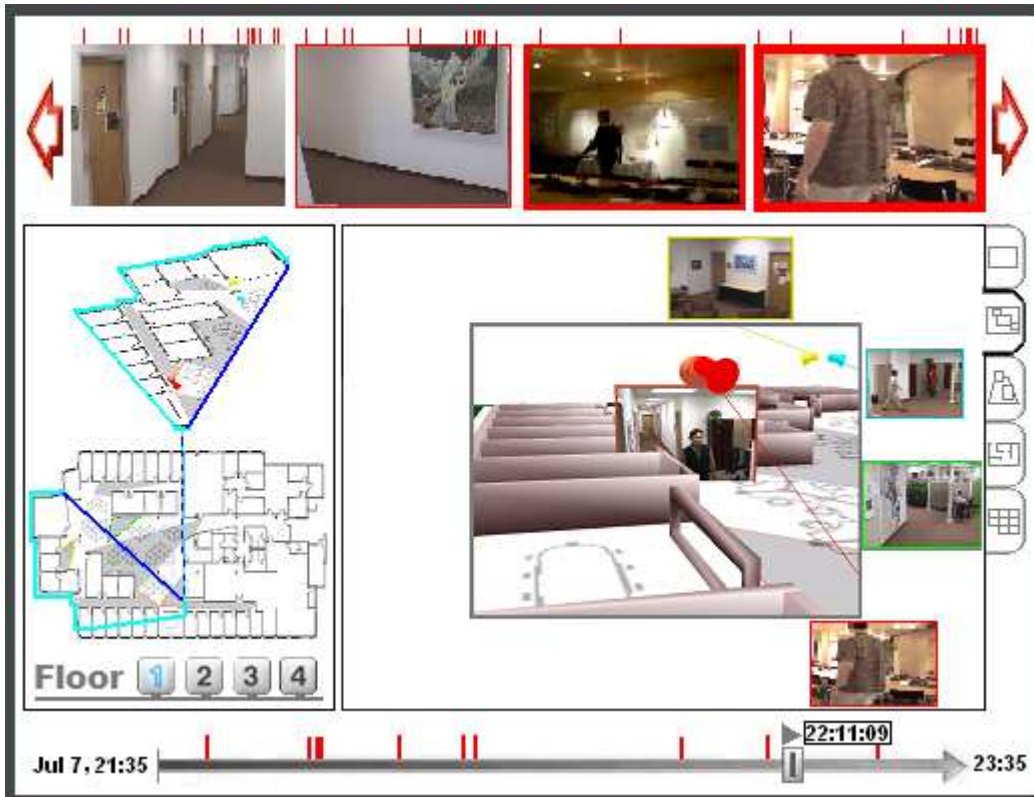
The design process is less organized than the previous steps. This step involves creativity, aesthetics, domain knowledge, experience and tacit knowledge that are not easily articulated. It contains many design-evaluation cycles. Often, we need to revisit the guidelines to remind us of the features we would like to have. The guidelines may lead to quite different requirements for different tasks. Designers need to understand the high level application requirements as well as the workflow in order to make tradeoffs between those requirements. We leave it as future work to explore effective design procedures and formalize the design process.

The rest of this section describes the interface we prototyped for this application:

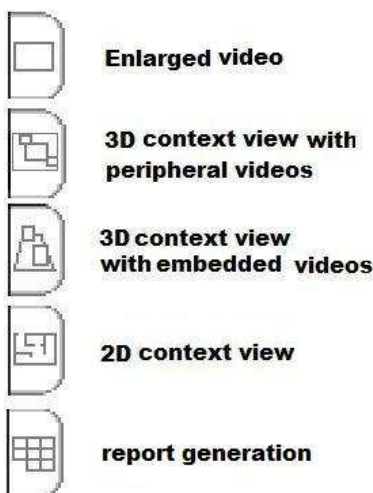
**Hardware configuration:** The hardware configuration of this new building security system includes a working display and multiple monitoring displays, each showing a 4x3 grid of videos. The interface for the new system is shown in Figure 9-2:



**Figure 9-2: Interface for the new system, which contains a monitoring display and a working display. The working display contains four areas: The area on top is the Popup Videos area. At bottom is the Timeline. Of the two areas in the middle, the right is the Detailed View, which contains five tabbed panels. The left area is called Context Overview.**

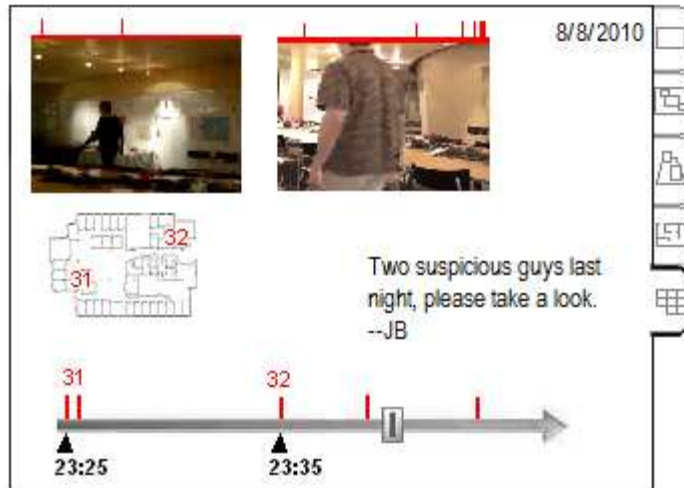


**Figure 9-3: An enlarged view of the working display. The working display contains four areas: The area on top is the Popup Videos area. The area at bottom is the Timeline. Of the two areas in the middle, the right one is the Detailed View, which contains five tabbed panels. The left area is called Context overview.**



**Figure 9-4: Description of the five tabs in the Detail View panel.**





**Figure 9-5: Report generation panel.**



**Figure 9-6: 2D Context View panel.**

**Monitoring Display:** The monitoring displays show an overview of all the videos captured in realtime. The displays are placed on one wall of the office, above and behind the working desk. The security guard can look at the monitoring displays from most places in this office. In this way, he need not be sitting in front of the working display to accomplish his monitoring task.

Security guards do not stare at the monitors all the time. They may not even sit in front of the monitors all the time. They need rest and something that can help them to keep awake during the boring job. Therefore, the interface should provide an overview of all surveillance videos, one that would allow the security guards to catch the information during a glance (Domain Knowledge 2 and 3 in Chapter 3.4). I choose to use the current monitoring system as an overview, because: 1) the current system meets the above requirement. 2) users are familiar with the current system.



**Working Display:** The new system features a working display, which supports far more functionality than simply showing a detailed view of the selected video (as in the old system). The security guard can perform smooth workflow, from computer-aided monitoring, to tracking, route planning, and report writing tasks, using the working display.

**Video Overview:** This view shows an overview of the videos. Guideline 1.1 in Section 5.4 also suggests an overview of all the videos for motion detection task. As mentioned in the previous section, this overview should be put close to detailed view to support multi-tasking. It can either show a shadow of all the videos on the monitoring display, or show a popup view of the videos where motion is detected. Popup view can be turned on during nights and weekends, when few people are in the building. During work hours, when many people walk around the building, the Popup view will be updating too rapidly to give the security guard a chance to observe the latest video.

**Popup Videos View:** The system automatically detects motions in a video and pops up those videos with motion. The latest video always pops up in the left side of the popup view, pushing the other videos in the view to the right. At the same time, the system will alert the users with a beep sound.

When a person walks through multiple videos, the user will see the corresponding sequence of videos in the popup view. This design can reduce the mental workload of searching for a target in a bank of videos.

Each popup video has a timeline on the top to indicate the detected events. Clicking an event causes the event to replay in the detailed view panel. A popup video can be locked so that it will not be pushed out of the view. When too many videos are locked, the videos will shrink to fit into the view.

**Detailed View Panel:** When the user detects suspicious activity, he will analyze the activity and plan a response in the detailed view panel. This panel comes from Guideline 1.3 Section 5.4. The user can freely switch between the five views based on the task requirement. Transition between views is smoothly animated both on the detailed view panel and on context overview.

**Enlarged videos:** The selected video will be enlarged in this view, so that more detail can be observed. This panel comes from Guideline 1.1 in Section 5.4. Popup videos can be selected by direct clicking. To select videos in the monitoring display, the user will move the mouse all the way to the top of the working display. A shadow of the monitoring display will slide down from the top. The videos in the shadow can be selected like a drop down menu. The user can also enlarge videos from other views in the detailed view panel.

**3D context view with peripheral videos:** This view is particularly useful for tracking a target person through multiple videos. The video being examined is shown in a larger spatial context, so that the target's relationship with the nearby environment can be judged intuitively, even if part of the environment is not captured in the video. The peripheral area contains videos from cameras adjacent to that of the central video. A

video's position in the peripheral area is based on its camera's relative direction from the central video's camera. In this way, when the target person walks out of the current video, the security guard can easily pick the proper adjacent video in which to observe the target. When the user clicks a peripheral video, the view will automatically navigate the user to look behind the camera of the selected video (Guideline 4.2, Chapter 8.4). During the navigation process, the context overview on the left will update simultaneously to minimize disorientation problem.

As Guideline 1.4 (Chapter 5.4) pointed out, embedding the video into the model helps spatial mapping, but may cause a severe occlusion problem. Navigation and filtering out of some visual elements are two ways to reduce such occlusion. This interface supports both approaches through low DOF manual manipulation (Guideline 4.5 suggests using low DOF manual control if necessary, but avoid high DOF manual control) using a standard mouse and keyboard: dragging the model to rotate and look behind the video; zooming using the mouse middle wheel; reducing the wall's height by dragging the top of the wall; highlighting all the possible paths between nearby cameras' coverage areas using a shortcut keys; switching between camera-aligned video and user-aligned video using another shortcut key.

**2D context view:** This view shows the 2D floor plan together with camera information in detail. This view mainly supports the route planning task. Since this view does not show any video, the currently selected video, as well as all the other videos within that floor are shown in the overview panel on the left. When users select an area on the map, the videos capturing that area will be highlighted in the overview, as shown in Figure 9-6.

The interface allows the user to add and modify information to be visualized on the model, e.g., social information, personal notes, etc., so that the visualization will hold whatever information the user decides to externalize. It provides an extra memory. Design Guideline 1.9 (Chapter 5.4) states that as users become more familiar with the environment, they develop spatial and social knowledge of that environment. They can utilize more information from the visualization in the tasks. Therefore, the design allows the user to add information to the visualizations. It is also a way to accumulate knowledge and to communicate information among multiple users.

**Report generation:** This view is designed to support fast creation of a work log and report. It can be used freely as a sketchpad. Users can write notes. They can drag a video into this view by putting the video onto the tag of this panel. The associated spatial context and time stamp of the video will be automatically added to this view. This function is designed to overcome the downside of Claim 2.

### **Context Overview:**

The context overview is designed to keep users well-oriented while transitioning between an egocentric view and an exocentric view. This interface provides an innovative 2-Stage "Detail + Context" view to help reduce disorientation. The visualization highlights the coverage area of the detailed view in the overview. It further decomposes the complex mapping process into two stages. It guides the user to first map the 3D context into a 2D view with the same view angle, then map the 2D view to the larger 2D map by providing salient visual cues. These cues are designed to help the user use a

feature matching strategy instead of the more costly mental rotation. This design is inspired by Design Guideline 2.2 (Chapter 6.2): users may use feature matching to bypass mental rotation if the matching features between the video and the model are both salient and close to the target object to be mentally mapped.

Context overview uses 2D maps, because the user is mainly interested in the topology of the space when looking at the building from an exocentric view.

Context overview supports realtime position labeling, so that the security guard can mark a path in realtime while he follows the person through multiple videos. The path will also appear in the detailed view to allow further analysis. This function is inspired by Design Guideline 1.4 (Chapter 5.4): embedded video allows real time path reconstruction strategy.

**Timeline:** The timeline allows the user to review the activities. The user can drag the time bar to watch the video from any selected moment. The red bars on top of the time line show time stamps of when activities were detected.

## **9.2.4 Other possible features**

A video surveillance system that needs to integrate with other systems in a real world application will normally contains more components. During the field study, I found that the security guards' office in that building has control over a CCTV (Closed Circuit Video Surveillance) system, a fire alarm system, and a parking lot gate control system. In this case, the benefit of Contextualized Videos is more obvious.

A real interface would likely contain additional components, e.g., a system control menu. Since these wider systems are not directly related to the guidelines, I do not present them in this prototype.

## **9.3 Usage Scenarios**

This section describes three usage scenarios with the new interface. These three scenarios correspond to the three described in Chapter 9.1.

### **9.3.1 Monitoring and event handling**

A building security guard JB was working on Thursday at 11:00pm. During this time of the night, there are normally very few activities going on in the building. JB set the system into automatic motion detection mode, so that whenever any motion is detected in a video, it will pop-up in the left side of the popup view and a beep will sound to alert him.

After a while, JB heard the motion alarm and came to look at the popup video. From the sequence of automatically popped-up videos, he saw a person exit an office, take the elevator, and then exit the building. JB tracked the target through multiple videos without any video selection.

An hour later, the motion alarm beeped again. JB saw one male in a red shirt on one video. From the context overview, he knew that the video captures the top floor of the parking area. JB enlarged that video in the detailed view and took a close look at this subject, who was smoking and talking over a cell phone. JB did not recognize him. He felt the subject might not be a resident of this building. So JB decided to keep an eye on

him. After 5 minutes, a male with white shirt appeared in another video. JB took a close look and did not recognize him either. JB decided to follow the male with white shirt. He clicked the 3D context tab and the view smoothly switched to the context view. JB locked the video containing the male with the red shirt, so he could still keep an eye on him, while focusing on the other subject.

Since there are some blind areas between surveillance cameras, JB used the context view to judge where a subject might be. JB can easily tell what the next video to observe should be from the nearby videos shown in the peripheral view.

When JB felt that he might have missed some activities of the male in white, he replayed the video by selecting the event bar (red bar) on the timeline.

The two subjects finally met on the parking lot. They talked as they walked around, eventually vanishing into a corner that can not be observed by any camera. JB looked at the context overview on the left to get an idea of where that location might be. The detailed context view also gave him a notion of all the videos in which the two subjects were likely to appear next.

After 5 minutes, JB still did not see these two persons. He decided to go to the top floor and take a look. He switched to the 2D map view. When planning the route, JB had to consider not only the time to arrive there, but also the chance of observing them from faraway if they attempted to leave the building through a nearby emergency exit. The map clearly shows where the blind area is, all the nearby exits, and the viewing distance from any location. JB picked a route that allowed him to see the blind area pretty early. He can go to check the nearby exists if the subjects are no longer in the blind area. The outcome was the same as in the earlier scenario described in Section 9.1.1.

### **9.3.2 Creating a daily report**

After coming back to the office, JB clicked the event bars on top of those recently popped up videos to review the events. He then dragged the events containing the two subjects into the report generation panel. The video segment containing the events, the time stamp and the spatial context containing these videos were automatically recorded in this panel. JB then typed a short report: “Two guys drank in the parking lot. I expelled them out.”

To report another event, JB wrote on the report generation panel, “Office 409’s door was open when I patrolled at 9:30 pm. No one was in it. I locked the door. This is the last person to exit the office.” He dragged that event to the above sentence on the report and a link was automatically added allow a subsequent reviewer to play the corresponding video segment.

The property manger came in during the day, read these reports and clicked the events bar on the timeline to watch the videos. It took her only 10 minutes to finish watching the videos.

### **9.3.3 A security guard’s first week in the building**

JB walked SH around the building and told him about the camera settings and the coverage areas. SH then came back to the office and looked at the visualization to register the videos to their actual location. He can clearly see the placement and coverage area of the cameras. With the help of the spatial context visualization, SH felt it easier to form a consistent mental model of the whole monitoring system.

JB also showed SH the events captured by each video, to help SH understand the working habits of the residents of the building, which areas he should pay particular attention to, etc. The visualization helped JB transfer his personal knowledge to SH.

## **9.4 Summary**

This chapter demonstrated how to use the design guidelines summarized in Chapters 5 to 8 to create a complex interfaces for a specific application. This chapter, together with the formal experiments that used complex activities to check the external validity of the findings using the basic tasks, addressed research Q5 (How beneficial are Contextualized Videos in complex activities?).

## 10 Summary and Future Work

Contextualized Videos integrate raw videos with a visualization of their spatial context. In this way, the relationships between videos become obvious and the subtle cues conveyed by videos are preserved. This dissertation systematically investigates the design space of Contextualized Videos to provide design guidelines that can help future designers create designs optimized according to task characteristics and user characteristics. For this purpose, the dissertation addressed the five research questions proposed in Chapter 1 in detail:

The design space provides a way to categorize, to describe, and to analytically evaluate Contextualized Video designs from the designers' point of view. Chapter 4 described the four major design dimensions within the design space of Contextualized Videos, which addressed research question Q1 (What is the ontology of the design space of Contextualized Videos? What are the major design dimensions?). This chapter also formalized a subspace composed of the model visualization and the video-model layout dimension.

It is important to understand the realworld problem before investigating the effect of Contextualized Videos on general tasks. Chapter 3 surveyed the video surveillance domain. Chapter 5 designed and evaluated Contextualized Video designs for video surveillance domain activities along two design dimensions: model visualization and video-model layout. Therefore, these two chapters addressed research question Q2 (For a particular domain and activity, what are the usable Contextualized Video designs and their limitations?).

To provide design guidelines, we need not only a way to describe the design space, but also a way to describe the tasks. Based on the literature review in Chapter 2, the domain survey in Chapter 3, and the design and evaluation experience in Chapter 5, Section 4.5 further proposed a Contextualized Video task taxonomy, which addressed research question Q3 (What are the distinctive tasks in Contextualized Video interfaces, and how can we classify them in a way that is useful to designers?).

Using tasks selected from the taxonomy, Chapters 6 to 8 designed and evaluated Contextualized Video interfaces along three dimensions. They correspond to the three research cycles outlined in our research approach section (Section 1.7). Chapter 6 focused on the video-model layout dimension. Chapter 7 focused on the video processing dimension. Chapter 8 focused on the navigation dimension. The formal experiment findings addressed research question Q4 (What are the effects of various Contextualized Video designs on the performance of key Contextualized video tasks?).

During the formal experiments, we also used complex activities to check the external validity of the findings that were based on the relatively simple tasks selected from the taxonomy. Chapter 9 demonstrated how to use the design guidelines summarized in Chapters 5 to 8 to create a complex interfaces for a specific application. This piece of work addressed research Q5 (How beneficial are Contextualized Videos in complex activities?).

## **10.1 Design Guidelines**

Here we revisit this dissertation's major contributions, the Contextualized Video design guidelines, according to the task characteristic and the user characteristic. The context and limitations of the guidelines were also discussed to avoid misleading future designers.

The guidelines are mainly summarized from formal evaluation results on desktop computers with less than twenty videos and a single floor plan. There were at most 10 people walking around in the scene used for evaluation. Occlusion between people in the video is not a severe problem in any experiment in this dissertation. These are the preconditions of the guidelines.

### **10.1.1 Design According to Task Characteristics**

To benefit from these guidelines, designers should carefully analyze the tasks to be supported by the new interface.

#### **Model Visualization and Video Model Layout**

Guideline 1.1: Emphasize videos for tasks that mainly require video information. (Section 5.4)

Guideline 1.2: Emphasize model for tasks that mainly require contextual information. (Section 5.4)

Guideline 1.3: Maximize the videos and their nearby context for integrative tasks (involving information from both the model and the videos). (Section 5.4)

Guideline 1.4: Use embedded designs for time-critical tasks. (Section 5.4)

Embedded designs refer to camera-aligned and closely related videos in this guideline. We also assume that the embedded videos do not occlude the model features that are critical for the success of the task. If important features are occluded, please consider Guideline 2.1 for improvement suggestions.

Guideline 1.5: Use combined designs to allow flexible user strategy (Section 5.4)

Guideline 1.6: Use simple 2D maps and associated designs, or minimize the embedded videos, for spatial understanding tasks. (Section 5.4)

Guideline 1.7: Use 3D and embedded designs for local navigation tasks. (Section 5.4)

Guideline 2.1: If the task requires observation of the model without occlusion, then the videos can be placed farther away from the camera while keeping the orientation of the video aligned with camera. (Section 6.2)

Guideline 2.2: If the task requires precise spatial judgment and the model shows plenty of landmarks, try both user-aligned and camera-aligned video layout. (Section 6.2)

#### **Video Processing**

Guideline 3.1: If the information required by the task can be reliably extracted from the video, it should be visualized on the model. (Section 7.2)

Guideline 3.2: If the task requires detailed video information that cannot be fully extracted, raw videos are still needed. (Section 7.2)

Guideline 3.3: If the task is to learn how to carefully maneuver through the environment, show both the path and video on the interface. (Section 7.2)

## **Navigation**

Guideline 4.1: If the task focuses on video level tracking, and no spatial understanding is needed, then choose semi-automatic navigation in detailed view. (Section 8.4)

Peripheral videos and embedded videos were used in all the conditions in the formal evaluation. We chose to use this combined design according to Guideline 1.5. Guidelines 4.1, 4.2 and 4.3 may not be valid if the interface can not show both peripheral videos and embedded videos. For example, if peripheral videos can not be shown for some reason, then users of detailed view navigation may have more chance to lose track of the tracking subject.

The view-transition animation for semi-automatic navigation is 1.5-seconds long. Slower animation may cause a higher chance of missing important events in the video. Faster animation may cause more severe disorientation. Therefore, the animation length should be adjusted according to the task characteristics of the future application.

Guideline 4.2: If the realtime task requires precise spatial mapping between video and model, especially from the camera's point of view, then semi-automatic navigation is preferred. (Section 8.4)

Guideline 4.3: If the task is realtime route or procedural learning, then showing a consistent overview that contains the whole route or procedure is preferred. (Section 8.4)

## **10.1.2 Design According to User Characteristics**

Guideline 1.8: Use simple designs for short-term use and non-expert users. (Section 5.4)

Guideline 1.9: Provide rich context cues for users who are familiar with the environment. (Section 5.4)

Guideline 2.4: For low spatial ability users, use camera aligned design to make spatial mapping more intuitive. (Section 6.2)

Guideline 2.5: To support users of different preferences and working habits, both user-aligned video and camera-aligned video can coexist in a single user interface, to allow flexible user strategy. (Section 6.2)

Guideline 4.4: Between semi-automatic navigation and manual navigation, users may have quite different preferences. Support both techniques if possible. (Section 8.4)

Guideline 4.5: Pure manual navigation with more than two DOF is not recommended for non-expert users. (Section 8.4)

Guideline 4.6: If manual navigation is necessary and the user has no prior experience of the navigation technique, we recommend giving the user enough time to find a good usage strategy through both guided training and free practice, before evaluation of the technique. (Section 8.4)

## **10.1.3 Other Guidelines**

Guideline 1.10: Avoid conflicting and misleading visual cues. (Section 5.4)



Guideline 2.3: In a complex activity that contains tasks other than spatial mapping, consider other cues that can bypass spatial mapping between video and model. (Section 6.2)

Guideline 3.4: Provide cues to help users manage their attention, if both video and dynamic visualization are presented on the model. (Section 7.2)

Guideline 4.7: Peripheral videos are helpful for time-critical video-level tracking and video-model integration tasks. (Section 8.4)

Guideline 4.8: Avoid complex navigation containing both translation and rotation; particularly avoid such navigation if the landmarks cannot be seen in the intermediate view. This is true for both manual and semi-automatic navigation. (Section 8.4)

Guideline 4.9: Reduce the number of videos in the peripheral view when animating the videos. Users may lose track of the target shown in the peripheral view, because of severe visual clutter caused by many videos moving around. (Section 8.4)

These guidelines not only provide design recommendations for given tasks but also provide a framework that guides designers to systematically analyze the tasks and understand the tradeoffs between design choices. Application designers can find example designs by specifying the task characteristics, user characteristics and data characteristics of their application. These examples should be viewed as starting points of the design process instead of final designs. They should be viewed as visual components instead of a complete interface.

As demonstrated in Chapter 9, a real world application normally supports multiple task goals. Therefore, instead of directly using the designs evaluated in this dissertation, understand that real interfaces are likely to be more complex than these evaluated designs. Combinations and variations of designs might be used. Information other than video and environment models might be visualized. Highly innovative interfaces might be created. However, armed with the guidelines, future designers can explore the design possibilities in a systematic way.

## **10.2 Future Work**

Although we produced a rich set of design guidelines and demonstrated the application of these guidelines, these guidelines are by no means complete. There are numerous subspaces to be explored within the Contextualized Video design space. The design and evaluation within each subspace can possibly produce additional useful design guidelines. I list a few promising subspaces to be explored in the future:

- (1) The scalability of different Contextualized Video designs, especially in terms of the number of videos, the model's complexity and available display resolution. The findings in this dissertation are mainly based on desktop computers with less than twenty videos and a single floor plan. Although we prototyped and analytically evaluated multi-floor building visualization, formal experiments need to be performed in the future. Multi-view multi-scale interfaces can potentially handle more videos and larger models. An example interface is described in Chapter 9, yet the scalability of this interface still needs to be evaluated in the future.

There were at most 10 people walking around in the scenes used for evaluation. It is not clear whether the evaluation results carry well to a denser crowd. For example, if there are hundreds--even thousands--of people in the scene, the target might have a higher chance of being occluded by other people. Consequently, the user might have a higher chance of losing track of the target. In this case, transitional animation between views might no longer be appropriate. Although animation can help users understand the navigation process, it prohibits observation of the videos, which move around in the animation. Formal experiments need to be performed in the future to evaluate the scalability of different designs.

- (2) It is a promising direction to explore the designs that combine video processing with Contextualized Videos to create video analysis interfaces that involve humans into a deeper loop of video analysis. Video processing algorithms need not be perfect before being useful. Traditional video processing research targeted fully automatic tracking algorithms that do not need human interference. However, “Sight is the sense of choice” [Davies 2005]. Therefore, the intermediate result can be presented to human operators through Contextualized Video interfaces. Some subtle cues that are hard to identify by computers can be easily captured by humans, who can provide their feedback to the video processing algorithm through the interface. In this way, humans and computers can collaborate more closely to perform the real tasks through Contextualized Video interfaces.
- (3) We have compared manual and automatic navigation for various Contextualized Video tasks. It seems promising to explore navigation techniques that combine manual and semi-automatic navigation. Such combined navigation can use a semi-automatic model for global navigation and use manual mode for local navigation. In this way, a good balance of flexibility and user-control may be achieved.
- (4) Along the model visualization dimension, an interesting question is that, if the model becomes as “realistic” as videos, through advanced 3D scanning and image-based rendering, does that ease spatial mapping between model and video? Will the realistic models have a visual clutter problem? What are the tradeoffs?
- (5) We used the term Contextualized Videos to refer to visualizations that combine videos with their spatial context. However, the context can include time and social context as well. This is demonstrated in the complex interface prototyped in Chapter 10. Multiple recent works also included time and social context in the visualization [Girgensohn, Kimber et al. 2007] [Ivanov, Wren et al. 2007]. However, current research has been focusing on the design side. No framework has been proposed for the analysis and evaluation of such interfaces.

This dissertation focused on static cameras with fixed camera parameters. Pan-Tilt-Zoom (PTZ) cameras and moving cameras should be considered in the future. Some videos even contain several segments captured at different location and different time. The movement of the camera often makes it harder to understand the scene captured by the camera. These dynamic features of the camera may require new Contextualized Video designs.

As mentioned in Chapter 9, the design process is less organized than the process of extracting tasks from an application and finding design guidelines using tasks. The actual design step involves creativity, aesthetics, domain knowledge, experience and tacit knowledge that are not easily articulated. It contains many design-evaluation cycles. Often, we need to revisit the guidelines to remind us the features we would like to have. The guidelines may lead to quite different requirements for different tasks. Designers need to understand the high level application requirements as well as the workflow in order to make tradeoffs between those requirements. It is a very important future work to explore effective design procedures and formalize the design process.

## References

- Aretz, A. J. (1991). "The design of electronic map displays." Human Factors **33**.
- Axis. "Converting an Analog CCTV System to IP-Surveillance." from [http://www.axis.com/documentation/whitepaper/video/cctv\\_to\\_ipsurveillance.htm](http://www.axis.com/documentation/whitepaper/video/cctv_to_ipsurveillance.htm).
- Baddeley, A. D. (1986). Working Memory. Oxford, UK, Oxford University Press.
- Baddeley, A. D. (2000). "The episodic buffer: a new component of working memory?" Trends in Cognitive Science **4**: 417-423.
- Baldonado, M. Q. W., A. Woodruff and A. Kuchinsky (2000). Guidelines for using multiple views in information visualization. the Working Conference on Advanced Visual Interfaces, Palermo, Italy, ACM Press.
- BAPS. (2002). "Building Security & Control Systems." from <http://www.baps.co.nz/buildingsecuritysystems.html>.
- Bolter, J., L. F. Hodges, T. Meyer and A. Nichols (1995). "Integrating Perceptual and Symbolic Information in VR." IEEE Computer Graphics and Applications **15**(4): 8-11.
- Boukhelifa, N., J. C. Roberts and P. Rodgers (2003). A Coordination Model for Exploratory Multi-View Visualization. International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV 2003), IEEE.
- Bowman, D. (1999). Interaction Techniques for Common Tasks in Immersive Virtual Environments: Design, Evaluation, and Application. Computer Science, Georgia Tech. **Doctor of Philosophy**.
- Bowman, D. and L. Hodges (1999). "Formalizing the design, evaluation, and application of interaction techniques for immersive virtual environments. ." The Journal of Visual Languages and Computing **10**(1): 17.
- Bowman, D., L. Hodges and J. Bolter (1998). "The Virtual Venue: User-Computer Interaction in Information-Rich Virtual Environments. ." Presence: Teleoperators and Virtual Environments **7**(5): 478-493.
- Bowman, D., D. Johnson and L. Hodges (2001). "Testbed Evaluation of Virtual Environment Interaction Techniques." Presence: Teleoperators and Virtual Environments **10**(1): 21.
- Bowman, D., E. Kruijff, J. LaViola and I. Poupyrev (2004). 3D User Interfaces: Theory and Practice. Boston, Addison-Wesley.
- Bowman, D., C. North, J. Chen, N. Polys, P. Pyla and U. Yilmaz (2003). Information-Rich Virtual Environments: Theory, Tools, and Research Agenda. ACM Virtual Reality Software and Technology. .
- Bowman, D. W., L. Hodges and D. Allison (1999). "The Educational Value of an Information-Rich Virtual Environment " Presence: Teleoperators and Virtual Environments **8**(3): 317-331.
- Burgess, N. (2006). "Spatial memory: how egocentric and allocentric combine." Trends in Cognitive Sciences **10**(12): 551-557.
- Burgess, N., S. Becker, J. A. King and J. O'Keefe (2001). "Memory for events and their spatial context: models and experiments." Philosophical Transactions of the Royal Society B Biological Sciences **356**(1413): 1493-1503.

- Byrne, P., S. Becker and N. Burgess (2007). "Remembering the past and imagining the future: a neural model of spatial memory and imagery. ." Psychological Review **114**(2): 340-375.
- Chandler, D. (1997). Visual Perception, World Lecture Hall. .
- Chen, B., B. Neubert, E. Ofek, O. Deussen and M. F. Cohen (2009). Integrated videos and maps for driving directions. Symposium on User Interface Software and Technology ACM.
- Chen, J., M. A. Narayan and M. A. Perez-Quinones (2005). The Use of Hand-held Devices for Search Tasks in Virtual Environments. IEEE VR2005 workshop on New Directions in 3DUI, Germany, IEEE Press.
- Chen, J., P. Pyla and D. Bowman (2004). Testbed Evaluation of Navigation and Text Display Techniques in an Information-Rich Virtual Environment. . IEEE Virtual Reality, Chicago, IL.
- Chen, M., R. Botchen, R. Hashim and I. Thornton (2006). "Visual Signatures in Video Visualization." IEEE Transactions on Visualization and Computer Graphics **12**(5): 1093-1100.
- comScore (2010). Video Metrix.
- Dalton, R. C. (2003). "The secret is to follow your nose: Route path selection and angularity." Environment and Behavior **35**(1): 107-131.
- Daniel, G. and M. Chen (2003). Video Visualization. Proceedings of the 14th IEEE Visualization 2003 Washington, DC, USA, IEEE Computer Society.
- Darken, R. and H. Cevik (1999). Map usage in virtual environments: orientation issues. Proceedings of IEEE Virtual Reality, IEEE Press.
- Davies, E. R. (2005). Machine Vision: Theory, Algorithms, Practicalities. San Francisco, CA, Morgan Kaufmann.
- de Haan, G., J. Scheuer, R. d. Vries and F. H. Post (2009). Egocentric Navigation for Video Surveillance in 3D Virtual Environments. IEEE symposium on 3D user interfaces, IEEE.
- Dourish, P. (2001). Where the Action Is: The Foundations of Embodied Interaction., MIT Press,.
- Dykstra, P. (1994). "X11 in Virtual Environments: Combining Computer Interaction Methodologies." j-X-RESOURCE **9**(1): 195-204.
- Ekstrom, R. B., J. W. French and H. H. Harman (1976). Manual for kit of factor referenced cognitive tests, Educational Testing Service, Princeton, N.J.
- Elkind, J. I., S. K. Card, J. Hochberg and B. M. Huey (1990). Human Performance Models for Computer-Aided Engineering. Orlando, FL, Academic Press.
- Ellis, W. D. (1938). Source Book of Gestalt Psychology. New York, Harcourt, Brace and Co.
- Elmqvist, N. and P. Tsigas (2007). Taxonomy of 3D Occlusion Management Techniques. IEEE Virtual Reality.
- Etienne, A. S., R. Maurer and V. Seguinot (1996). "Path integration in mammals and its interaction with visual landmarks." Journal of Experimental Biology **199**(1): 201-209.
- Eysenck, M. W. (2001). Principles of Cognitive Psychology, Psychology Press.
- Faraday, P. and A. Sutcliffe (1996). An Empirical study of Attending and Comprehending Multimedia Presentations. ACM Multimedia, Boston, MA.

- Foreman, N. (2005). "Transfer of Spatial Knowledge to a Two-Level Shopping Mall in Older People, Following Virtual Exploration " Environment and Behavior **37**(2): 275-292.
- Forsyth, D. and J. Ponce (2004). Computer Vision: A Modern Approach. New Delhi, Prentice-Hall of India.
- Gailing, T., E. Lindberg and T. Mantyla (1983). "Orientation in buildings: Effects of familiarity, visual access, and orientation." Journal of Applied Psychology **68**(1): 177-186.
- Girgensohn, A., D. Kimber, J. Vaughan, T. Yang, F. Shipman, T. Turner, E. Rieffel, L. Wilcox, F. Chen and T. Dunnigan (2007). DOTS: support for effective video surveillance. International Multimedia Conference ACM.
- Girgensohn, A., F. Shipman, T. Turner and L. Wilcox (2007). Effects of Presenting Geographic Context on Tracking Activity between Cameras. SIGCHI, San Jose.
- Google. (2008c). "KML Documentation." <http://code.google.com/apis/kml/documentation/>.
- Gortler, S. J., R. Grzeszczuk, R. Szeliski and M. F. C. . (1996). The lumigraph. ACM SIGGRAPH.
- Han, J. and B. Smith (1996). CU-SeeMe VR immersive desktop teleconferencing. MULTIMEDIA '96: Proceedings of the fourth ACM international conference on Multimedia.
- Harris, D. H. and F. D. Chaney (1969). Human factors in quality assurance. New York, Wiley.
- Hochmain, H. and A. U. Frank (2002). "Influence of estimation errors on wayfinding decisions in unknown street networks -- analyzing the least angle strategy." Spatial Cognition and Computation **2**(4): 283-313.
- Howard, I. P. (1991). Spatial vision within egocentric and exocentric frames of reference. Pictorial Communication in Virtual and Real Environments. S. R. Ellis. Bristol, PA Ed. Taylor & Francis: 338-358.
- ipcamerasupply. "What is an IP camera?" from <http://www.ipcamerasupply.com/>.
- ITI Technologies, I. "CAMERA SURVEILLANCE SYSTEMS." from <http://www.ititek.com/camera.shtml>.
- Ivanov, Y., C. Wren, A. Sorokin and I. Kaur (2007). "Visualizing the History of Living Spaces." IEEE Transactions on Visualization and Computer Graphics: 6-13.
- Jiang, Y., I. R. Olson and M. M. Chun (2000). "Organization of Visual Short-Term Memory. ." Journal of Experimental Psychology: Learning, Memory, and Cognition **26**(3): 683-702.
- Jones, Q. (1999). The Transfer of Spatial Knowledge from Virtual to Natural Environments as a Factor of Map Representation and Exposure Duration, NAVAL POSTGRADUATE SCHOOL MONTEREY CA. **Master**.
- Krum, D. (2004). Wearable Computers and Spatial Cognition. College of Computing, Georgia Institute of Technology. **PhD**.
- Krum, D., W. Ribarsky, C. Shaw, L. Hodges and N. Faust (2001). Situational visualization. ACM Symposium on Virtual Reality Software and Technology.
- Lam, H. (2008). A Framework of Interaction Costs in Information Visualization. IEEE Information Visualization Conference, IEEE.

- Landy, M. S., L. T. Maloney, E. B. Johnston and M. Young (1995). "Measurement and Modeling of Depth Cue Combination: in Defense of Weak Fusion." Vision Research **35**: 389-412.
- Lawton, C. A. (1996). "Strategies or indoor wayfinding: The role of orientation." Journal of Environmental Psychology **16**(2): 137-145.
- Lehikoinen, J. and R. Suomela (2002). "Perspective map." Proceedings of the Sixth International Symposium on Wearable Computers (ISWC 2002): 171-178.
- Levoy, M. and P. Hanrahan (1996). Light field rendering. ACM SIGGRAPH.
- Logie, R. H. (1995). Visuo-spatial memory. Hove, UK, Erlbaum.
- Lynch, K. (1960). The Image of the City. Cambridge, Massachusetts, MIT Press.
- Mackworth, N. H. (1948). "The breakdown of vigilance during prolonged visual search." Quarterly Journal of Experimental Psychology **17**: 302-325.
- Mark, W., L. McMillan and G. Bishop (1997). Post-rendering 3d warping. Symposium on I3D Graphics.
- Martin, J. (1989). High Tech Illustration, Addison-Wesley.
- McCormick, E., C. D. Wickens, R. Banks and M. Yeh (1998). "Frame of reference effects on scientific visualization subtasks." Human Factors **40**: 443-451.
- McGuffin, M. J., L. Tancau and R. Balakrishnan (2003). Using deformations for browsing volumetric data. IEEE Visualization.
- McMillan, L. and G. Bishop (1995). Plenoptic modeling: an image-based rendering system. ACM SIGGRAPH.
- Merriam-Webster's (2000). Merriam-Webster's Collegiate Dictionary (Tenth Edition).
- Microsoft. (2008). "Virtual Earth." <http://maps.live.com/>.
- Miyake, A. and P. Shah (1999). Models of Working Memory: Mechanisms of Active Maintenance and Executive Control. New York, Cambridge University Press.
- Moeser, S. D. (1988). "Cognitive mapping in a complex building." Environment and Behavior **20**(1): 21-49.
- Mou, W., T. P. McNamara, C. M. Valiquette and B. R. B (2004). "Allocentric and Egocentric Updating of Spatial Memories." Journal of Experimental Psychology: Learning, Memory and Cognition **30**: 142-157.
- Munro, A., R. Breaux, J. Patrey and B. Sheldon (2002). Cognitive Aspects of Virtual Environment Design. Handbook of virtual environments: design, implementation, and applications. Mahwah, N.J., Lawrence Erlbaum Associates.
- NASA. (2006). "World Wind 1.4." <http://worldwind.arc.nasa.gov/>.
- Norman, D. and D. Bobrow (1975). "On data-limited and resource-limited processing." Cognitive Psychology **7**: 44-60.
- North, C. and B. Shneiderman (1997). A taxonomy of multiple window coordinations.
- Péruch, P., L. Belingard and C. Thinus-Blanc (2000). "Transfer of Spatial Knowledge from Virtual to Real Environments." Spatial Cognition II: 12.
- Phillips, W. A. (1974). "On the Distinction Between Sensory Storage and Short-Term Visual Memory." Perception & Psychophysics **16**: 283-290.
- Piaget, J. and B. Inhelder (1967). The Child's Conception of Space. New York, Norton.
- Pillay, H. K. (1994). "Cognitive Load and Mental Rotation: Structuring Orthographic Projection for Learning and Problem Solving." Instructional Science **22**: 23.

- Plumlee, M. and C. Ware (2003). "Integrating multiple 3d views through frame-of-reference interaction." International Conference on Coordinated and Multiple Views in Exploratory Visualization.
- Polys, N. (2006). Display Techniques in Information-Rich Virtual Environments. Computer Science, Virginia Tech. Doctor of Philosophy.
- Psathas, G. (1995). Conversation Analysis: The Study of Talk-in-Interaction., Thousand Oaks, CA: SAGE Publications, Inc.
- Quek, F., R. Bryll, H. Arslan, C. Kirbas and D. McNeill (2002). "A multimedia database system for temporally situated perceptual psycholinguistic analysis." Multimedia Tools and Applications. Kluwer Academic Publishers. 18(2): 91-113.
- Quek, F., Y. Shi, C. Kirbas and S. Wu (2002). VisSTA: A Tool for Analyzing Multimodal Discourse Data. Seventh International Conference on Spoken Language Processing, Denver, CO.
- Roberts, J. C. (1999). On Encouraging Coupled Views for Visualization Exploration. Visual Data Exploration and Analysis VI, Proceedings of SPIE, IS&T and SPIE.
- Rosson, M. and J. Carroll (2002). Usability Engineering: Scenario-based development of human-computer interaction, Morgan Kaufmann Publishers.
- Ruddle, R., S. Payne and D. Jones (1998). "The effects of maps on navigation and search strategies in very-large-scale virtual environments." Journal of Experimental Psychology 5(1): 54-75.
- Sauers, R. v. (2009). "Car Quest Brake System Diagram." from <http://rogerdanielsalignment.com/brake-maint.php>.
- Sawhney, H. S., A. Arpa, R. Kumar, S. Samarasekera, M. Aggarwal, S.Hsu, D. Nister and K. Hanna (2002). Video Flashlights: Realtime Rendering of Multiple Videos for Immersive Model Visualization. Proc. 13th Eurographics workshop on Rendering.
- Scharl, A. and K. Tochtermann (2007). How Geobrowsers, Social Software and the Web 2.0 are Shaping the Network Society. London, Springer.
- Schnädelbach, H., A. Penn, P. Steadman, S. Benford, B. Koleva and T. Rodden (2006). Moving office: inhabiting a dynamic building. CSCW '06: Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work, Banff, Alberta, Canada.
- Sebe, I. O., J. Hu, S. You and U. Neumann (2003). 3D Video Surveillance with Augmented Virtual Environments. First ACM SIGMM International Workshop on Video Surveillance.
- Seitz, S. M. and C. M. Dyer (1996). View morphing. ACM SIGGRAPH.
- Shepard, R. N. and J. Metzler (1971). "Mental Rotation of Three-Dimensional Objects." Science(171): 701-703.
- Shi, Y., R. T. Rose and F. Quek (2004). A System for Situated Temporal Analysis of Multimodal Communication. 4th International Conference on Language Resources and Evaluation, Lisbon, Portugal.
- Shneiderman, B. (1996). The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. IEEE Symposium on Visual Languages.
- Shneiderman, B. and C. Plaisant (2005). Designing the User Interface: Strategies for effective human-computer interaction, Addison-Wesley.



- Shneidermann, B. (2003). "Visualization viewpoints." IEEE Computer Graphics and Applications **23**(4): 6.
- Shoemake, K. (1992). ARCBALL: a user interface for specifying three-dimensional orientation using a mouse." Graphics Interface.
- Sholl, M. J. and T. L. Nolin (1997). "Orientation specificity in representations of place." Journal of Experimental Psychology: Learning, Memory and Cognition **23**: 1496-1503.
- Shum, H.-Y. and S. B. Kangues (2000). "A Review of Image-based Rendering Technique." IEEE/SPIE Visual Communications and Image Processing (VCIP): 2-13.
- Soh, B. K. and T. L. Smith-Jackson (2004). "Influence of Map Design, Individual Differences, and Environmental Cues on Wayfinding Performance." SPATIAL COGNITION AND COMPUTATION **4**(2): 30.
- State of California Employment Development Department. (2002). "Labor market information: Security guards. [Online]." from <http://www.calmis.cahwnet.gov/file/occguide/SECURGRD.HTM>.
- Stevens, A. and P. Coupe (1978). "Distortions in Judged Spatial Relations." Cognitive Psychology **10**: 422-437.
- Stoakley, R., M. J. Conway and R. Pausch (1995). Virtual reality on a WIM: interactive worlds in miniature. Proceedings of the SIGCHI conference on Human factors in computing systems. Denver, Colorado, United States, ACM Press/Addison-Wesley Publishing Co.
- Sutcliffe, A. and P. Faraday (1994). "Designing Presentation in Multimedia Interfaces." ACM CHI.
- Tan, D. S., G. G. Robertson and M. Czerwinski (2001). Exploring 3D Navigation: Combining Speed-coupled Flying with Orbiting. ACM SIGCHI.
- Thomas, J. J. and K. A. Cook (2005). Illuminating the Path: The research and development agenda for visual analytics, IEEE Press.
- Thorndyke, P. W. and B. Hayes-Roth (1978). Spatial knowledge acquisition from maps and navigation. Paper presented at the meetings of the Psychonomic Society. San Antonio, TX.
- Tory, M. (2003). Mental Registration of 2D and 3D Visualizations (An Empirical Study). Proc. IEEE Visualization.
- Tory, M., A. E. Kirkpatrick, M. S. Atkins and T. Moller (2006). "Visualization Task Performance with 2D, 3D, and Combination Displays." IEEE Transactions on Visualization and Computer Graphics: 2-13.
- Tory, M. and T. Moller (2003). Rethinking Visualization: A High-Level Taxonomy. IEEE Symposium on Information Visualization.
- Tory, M., Moller, T., Atkins, M. S., and Kirkpatrick, A. E. (2004). Combining 2D and 3D views for orientation and relative position tasks. Proc. SIGCHI.
- Velez, M. C., D. Silver and M. Tremaine (2005). Understanding Visualization through Spatial Ability Differences. IEEE Visualization.
- Viktor, S., W. Eide, F. Eliassen, O.-C. Granmo and O. Lysne (2003). "Supporting timeliness and accuracy in distributed real-time content-based video analysis " MULTIMEDIA '03: Proceedings of the eleventh ACM international conference on Multimedia

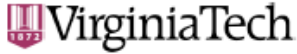
- Vogel, E. K., G. F. Woodman and S. J. Luck (2001). "Storage of Features, Conjunctions, and Objects in Visual Working Memory." Journal of Experimental Psychology: Human Perception and Performance **27**(1): 92-114.
- Wakefield, J. (2002). Watching your every move. BBC News Online (7 February). <http://news.bbc.co.uk>.
- Waller, D. (2006). "Egocentric and nonegocentric coding in memory for spatial layout: Evidence from scene recognition." Mem Cogni **34**(3): 14.
- Waller, D., E. Hunt and D. Knapp (1998). "The Transfer of Spatial Knowledge in Virtual Environment Training." Presence: Teleoperators and Virtual Environments **7**(2): 129-143.
- Wang, R. F. and J. R. Brockmole (2003). "Simultaneous spatial updating in nested environments." Psychonomic bulletin & review **10**: 981-986.
- Wang, R. F. and E. S. Spelke (2002). "Human spatial representation: insights from animals." Trends in Cognitive Sciences: 376-382.
- Wang, Y., D. M. Krum, E. M. Coelho and D. A. Bowman (2007). "Contextualized Videos: Combining Videos with Environment Models to Support Situational Understanding." IEEE Transactions on Visualization and Computer Graphics **13**(6): 1568-1575.
- Wang, Y., Y. Wang, S. Patel and D. Patel (2006). "A Layered Reference Model of the Brain (LRMB)." IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS **36**(2): 124-133.
- Ware, C. (2000). Information Visualization: Perception for Design. New York, Morgan Kaufman.
- Wertheimer, M. (1923). "Laws of organization in perceptual forms." Psychologische Forschung **4**: 301-350.
- Wickens, C. and J. Hollands (2000). Engineering Psychology and Human Performance, 3rd Edition, Prentice-Hall, Inc.
- Wickens, C. D., and Carswell, C.M. (1995). "The proximity compatibility principle: its psychological foundation and relevance to display design." Human Factors(37): 473-494.
- Wickens, C. D., M. A. Vincow, A. W. Schopper and J. E. Lincoln (1997). "Computational Models of Human Performance in the Design and Layout of Controls and Displays. ." Wright-Patterson AFB, Crew Systems Ergonomics Information Analysis Center: Dayton, OH.
- Wijk, J. v. and W. Nuij (2003). Smooth and efficient zooming and panning. IEEE Symposium on Information Visualization.
- Wikipedia. "Google Street View Wiki." [http://en.wikipedia.org/wiki/Google\\_Street\\_View](http://en.wikipedia.org/wiki/Google_Street_View).
- Williams, H. P., S. Hutchinson and C. D. Wickens (1996). "A comparison of methods for promoting geographic knowledge in simulated aircraft navigation." Human factors **38**(1): 50-64.
- Xu, L. (2007). Issues in video analytics and surveillance systems: Research / prototyping vs. applications / user requirements. IEEE Conference on Advanced Video and Signal Based Surveillance: 10-14.
- Yost, B. (2006a). The Visual Scalability of Integrated and Multiple Views for High Resolution Displays. Computer Science. Blacksburg, VA, Virginia Tech. **PhD**.

Yost, B., H. Hacıahmetoglu and C. North (2007). Beyond Visual Acuity: The Perceptual Scalability of Information Visualizations for Large Displays. ACM Conference on Human Factors in Computer Systems (CHI).

Yost, B. and C. North (2006b). The Perceptual Scalability of Visualization. IEEE Symposium on Information Visualization, Baltimore Maryland, IEEE Press.

# Appendix A: Experiment 1 (Model Visualization and Video-Model Layout)

## IRB Approval:



Office of Research Compliance  
Institutional Review Board  
2000 Kraft Drive, Suite 2000 (0497)  
Blacksburg, Virginia 24061  
540/231-4991 Fax 540/231-0959  
e-mail [moored@vt.edu](mailto:moored@vt.edu)  
[www.irb.vt.edu](http://www.irb.vt.edu)


FWA000005721 expires 1/20/2010  
IRB # is IRB00000667

DATE: October 30, 2007

### MEMORANDUM

TO: Doug A. Bowman  
Tonya L. Smith-Jackson  
Yi Wang

Approval date: 10/29/2007  
Continuing Review Due Date: 10/14/2008  
Expiration Date: 10/28/2008

FROM: David M. Moore 

SUBJECT: **IRB Expedited Approval:** "Comparing Contextualized Video Visualization Designs Using Tracking Tasks", IRB # 07-542

This memo is regarding the above-mentioned protocol. The proposed research is eligible for expedited review according to the specifications authorized by 45 CFR 46.110 and 21 CFR 56.110. As Chair of the Virginia Tech Institutional Review Board, I have granted approval to the study for a period of 12 months, effective October 29, 2007.

As an investigator of human subjects, your responsibilities include the following:

1. Report promptly proposed changes in previously approved human subject research activities to the IRB, including changes to your study forms, procedures and investigators, regardless of how minor. The proposed changes must not be initiated without IRB review and approval, except where necessary to eliminate apparent immediate hazards to the subjects.
2. Report promptly to the IRB any injuries or other unanticipated or adverse events involving risks or harms to human research subjects or others.
3. Report promptly to the IRB of the study's closing (i.e., data collecting and data analysis complete at Virginia Tech). If the study is to continue past the expiration date (listed above), investigators must submit a request for continuing review prior to the continuing review due date (listed above). It is the researcher's responsibility to obtain re-approval from the IRB before the study's expiration date.
4. If re-approval is not obtained (unless the study has been reported to the IRB as closed) prior to the expiration date, all activities involving human subjects and data analysis must cease immediately, except where necessary to eliminate apparent immediate hazards to the subjects.

### Important:

If you are conducting federally funded non-exempt research, this approval letter must state that the IRB has compared the OSP grant application and IRB application and found the documents to be consistent. Otherwise, this approval letter is invalid for OSP to release funds. Visit our website at <http://www.irb.vt.edu/pages/newstudy.htm#OSP> for further information.

cc: File

*Invent the Future*

VIRGINIA POLYTECHNIC INSTITUTE UNIVERSITY AND STATE UNIVERSITY  
*An equal opportunity, affirmative action institution*

### **Post Task Questionnaire for contextualized video test bed**

Circle the one that matches your opinion the best:

1. The visualization design was easy to learn.
  - a. Disagree strongly
  - b. Disagree somewhat
  - c. Undecided
  - d. Agree somewhat
  - e. Agree strongly
  
2. The visualization design was easy to use.
  - a. Disagree strongly
  - b. Disagree somewhat
  - c. Undecided
  - d. Agree somewhat
  - e. Agree strongly
  
3. The visualization design was a joy to use.
  - a. Disagree strongly
  - b. Disagree somewhat
  - c. Undecided
  - d. Agree somewhat
  - e. Agree strongly
  
4. If my job was security surveillance, I would use this visualization design.
  - a. Disagree strongly
  - b. Disagree somewhat
  - c. Undecided
  - d. Agree somewhat
  - e. Agree strongly
  
5. This visualization design is useful for tracking and surveillance tasks.
  - a. Disagree strongly
  - b. Disagree somewhat
  - c. Undecided
  - d. Agree somewhat
  - e. Agree strongly



**Result Data:**

User ID	Knowledge	Gender	Model/Visu	Video/Placement	Overall Time	Time (sec)	Time (sec)	Time (sec)	Time (sec)	Time (sec)	Overall Replay	Replay	Replay	Replay	Replay	Non-effective time	Non-effective time	Non-effective time	Non-effective time
1	Explicit	M	2D	Associated	127.5	102	131	97	180	2	2	2	2	2	2	19	21	4	30
11	Explicit	M	2D	Associated	125.25	103	115	180	103	2.25	2	2	2	4	1	19	21	4	30
15	Explicit	F	2D	Associated	121.5	121	97	106	162	1.75	2	1	1	2	2	19	21	4	30
2	Explicit	M	2D	Embedded	123	160	137	45	150	2	3	2	1	2	2	19	21	4	30
7	Explicit	M	2D	Embedded	110.25	103	114	44	180	1.75	2	2	2	1	2	19	21	4	30
12	Explicit	M	2D	Embedded	128.5	105	140	117	152	1.75	1	2	2	2	2	19	21	4	30
3	Explicit	M	3D	Associated	173	167	180	180	165	2.75	3	3	3	3	2	19	21	4	30
8	Explicit	M	3D	Associated	147.25	154	165	90	180	2.75	3	3	3	2	2	19	21	4	30
13	Explicit	F	3D	Associated	137.75	118	178	96	159	2.25	2	3	3	2	2	19	21	4	30
4	Explicit	M	3D	Embedded	118.25	114	114	83	162	2	2	2	2	2	2	19	21	4	30
10	Explicit	M	3D	Embedded	88.25	54	76	48	175	1.25	1	1	1	1	2	19	21	4	30
14	Explicit	F	3D	Embedded	159.25	160	180	133	164	2.75	3	3	3	3	2	19	21	4	30
5	Explicit	M	3D	Combined	109.5	108	141	91	98	1.25	1	2	1	1	1	19	21	4	30
9	Explicit	M	3D	Combined	98.25	119	83	88	103	1	1	1	1	1	1	19	21	4	30
20	Explicit	F	3D	Combined	139	126	162	117	151	2.25	2	3	3	2	2	19	21	4	30
16	Explicit	M	2D	Combined	130	153	146	63	158	1.75	2	2	2	1	2	19	21	4	30
17	Explicit	M	2D	Combined	125.25	173	116	48	164	2	3	2	2	1	2	19	21	4	30
21	Explicit	F	2D	Combined	120.75	129	136	77	141	1.75	2	2	2	1	2	19	21	4	30
1	Tacit	M	2D	Associated	113.5	95	110	96	153	1	1	1	1	1	1	19	21	4	30
6	Tacit	M	2D	Associated	155	138	171	131	180	2.25	2	3	2	2	2	19	21	4	30
10	Tacit	F	2D	Associated	140	169	119	105	167	2.25	3	2	2	2	2	19	21	4	30
2	Tacit	M	2D	Embedded	100.75	68	136	48	151	1.5	1	2	1	1	2	19	21	4	30
7	Tacit	F	2D	Embedded	106.75	68	125	54	180	1.5	1	2	2	1	2	19	21	4	30
16	Tacit	M	2D	Embedded	122	115	131	97	145	2	2	2	2	2	2	19	21	4	30
3	Tacit	F	3D	Associated	113.75	91	119	94	151	1.75	1	2	2	2	2	19	21	4	30
9	Tacit	F	3D	Associated	107.25	89	134	78	128	1.25	1	2	2	1	1	19	21	4	30
14	Tacit	M	3D	Associated	135.5	154	126	105	157	2.25	3	2	2	2	2	19	21	4	30
4	Tacit	F	3D	Embedded	136.25	165	153	79	148	2.25	3	2	2	2	2	19	21	4	30
11	Tacit	F	3D	Embedded	121	89	129	111	155	1.75	1	2	2	2	2	19	21	4	30
13	Tacit	M	3D	Embedded	103	112	122	71	107	1.5	2	2	2	1	1	19	21	4	30
5	Tacit	M	3D	Combined	78.75	106	67	60	82	1.25	2	1	1	1	1	19	21	4	30
8	Tacit	F	3D	Combined	137	180	125	90	153	2.25	3	2	2	2	2	19	21	4	30
17	Tacit	M	3D	Combined	102.25	103	101	77	128	1.75	2	2	2	1	2	19	21	4	30
12	Tacit	F	2D	Combined	94.75	63	122	43	151	1.5	1	2	1	1	2	19	21	4	30
15	Tacit	M	2D	Combined	126	56	180	88	180	1.75	1	3	3	2	1	19	21	4	30
18	Tacit	M	2D	Combined	110.5	79	140	57	166	1.5	1	2	2	1	2	19	21	4	30



User ID	Normali ze Factor	Normali ze Factor	Normali ze Factor	Normali ze Factor	Overall Effective Time	Effective Time	Effective Time	Effective Time	Effective Time	Effective Time	Overall Normalized Time	Normalized Time	Normalized Time	Normalized Time	Normalized Time
1	0.94084	0.83881	1.2425	0.7294	90.5	64	89	89	89	120	114.4778315	95.9657876	109.8835	120.52714	131.294898
11	0.94084	0.83881	1.2425	0.7294	93.75	65	73	164	73	73	123.0394284	96.90662865	96.462614	223.65861	75.1298582
15	0.94084	0.83881	1.2425	0.7294	89.75	83	76	98	102	102	111.2703414	113.8417676	81.364118	131.71007	118.165408
2	0.94084	0.83881	1.2425	0.7294	82.25	103	95	41	90	90	107.694492	150.5345688	114.91633	55.914663	109.942415
7	0.94084	0.83881	1.2425	0.7294	74.25	65	72	40	120	120	94.62436009	96.90662865	95.623809	54.672105	131.294898
12	0.94084	0.83881	1.2425	0.7294	96.25	86	98	109	92	92	118.11776009	98.78831076	117.43275	145.3781	110.871247
3	0.94084	0.83881	1.2425	0.7294	125	110	117	168	105	105	163.0294216	157.1204562	150.98496	223.65861	120.363656
8	0.94084	0.83881	1.2425	0.7294	92.75	97	102	82	90	90	131.6041519	144.8895225	138.40288	111.82931	131.294898
13	0.94084	0.83881	1.2425	0.7294	95.5	80	115	88	99	99	123.897087	111.0192445	149.30735	119.28459	115.97716
4	0.94084	0.83881	1.2425	0.7294	81.25	76	72	75	102	102	106.0441421	107.2558803	95.623809	103.13147	118.165408
10	0.94084	0.83881	1.2425	0.7294	62.25	35	55	44	115	115	75.46118419	50.80541696	63.749206	59.642297	127.647817
14	0.94084	0.83881	1.2425	0.7294	111.25	103	117	121	104	104	146.6006585	150.5345688	150.98496	165.25886	119.62424
5	0.94084	0.83881	1.2425	0.7294	85.75	89	99	87	68	68	101.1092547	101.6108339	118.27155	113.07185	71.4827777
9	0.94084	0.83881	1.2425	0.7294	79.75	100	62	84	73	73	91.5137494	111.9600855	69.620843	109.34421	75.1298582
20	0.94084	0.83881	1.2425	0.7294	96.75	88	99	109	91	91	127.4880918	118.5459729	135.88646	145.3781	110.141831
16	0.94084	0.83881	1.2425	0.7294	94	115	104	59	98	98	114.9866297	143.9486814	122.46558	78.260515	115.247744
17	0.94084	0.83881	1.2425	0.7294	84.5	116	74	44	104	104	109.8333647	162.7655025	97.301419	59.642297	119.62424
21	0.94084	0.83881	1.2425	0.7294	84.75	91	94	73	81	81	108.4924691	121.3684961	114.07753	95.676184	102.84767
1	0.94084	0.83881	1.2425	0.7294	95	76	89	92	123	123	103.133436	89.37990021	92.268587	119.28459	111.600663
6	0.94084	0.83881	1.2425	0.7294	112.75	100	108	123	120	120	141.8361111	129.8360666	143.43571	162.77377	131.294898
10	0.94084	0.83881	1.2425	0.7294	98.25	112	77	97	107	107	127.7749966	159.0021383	99.817835	130.46752	121.812488
7	0.94084	0.83881	1.2425	0.7294	69.5	49	94	44	91	91	86.95971135	63.97719173	114.07753	59.642297	110.141831
2	0.94084	0.83881	1.2425	0.7294	75.5	49	83	50	120	120	91.80508517	63.97719173	104.85067	67.097584	131.294898
16	0.94084	0.83881	1.2425	0.7294	85	77	89	89	85	85	111.0931741	108.1967213	109.8835	120.52714	105.765334
3	0.94084	0.83881	1.2425	0.7294	81.5	72	77	86	91	91	103.093925	85.61653599	99.817835	116.7995	110.141831
9	0.94084	0.83881	1.2425	0.7294	83.5	70	92	74	98	98	96.60469054	83.73485388	112.39992	96.918732	93.3652606
14	0.94084	0.83881	1.2425	0.7294	93.75	97	84	97	97	97	123.8912117	144.8895225	105.68947	130.46752	114.518328
4	0.94084	0.83881	1.2425	0.7294	94.5	108	111	71	88	88	122.4227134	155.2387741	128.33722	98.16128	107.963583
11	0.94084	0.83881	1.2425	0.7294	88.75	70	87	103	95	95	110.7307623	83.73485388	108.20589	137.92281	113.059495
13	0.94084	0.83881	1.2425	0.7294	74.5	74	80	67	77	77	93.49421733	105.3741981	102.33425	88.220897	78.0475226
5	0.94084	0.83881	1.2425	0.7294	55.5	68	46	56	52	52	72.57562515	99.72915182	56.199968	74.552871	59.8121201
8	0.94084	0.83881	1.2425	0.7294	95.25	123	83	82	93	93	124.4080067	169.3613899	104.85067	111.82931	111.600663
17	0.94084	0.83881	1.2425	0.7294	66.25	65	59	73	68	68	92.66685323	96.90662865	84.719339	95.676184	93.3652606
12	0.94084	0.83881	1.2425	0.7294	63.5	44	80	39	91	91	81.29465654	59.27298646	102.33425	53.429558	110.141831
15	0.94084	0.83881	1.2425	0.7294	96	37	117	80	150	150	111.0777921	52.68709907	150.98496	109.34421	131.294898
18	0.94084	0.83881	1.2425	0.7294	79.25	60	98	53	106	106	95.91687263	74.32644334	117.43275	70.825227	121.083072



# Appendix B: Experiment 2 and 3 (Video-Model Layout and Video Processing)

## IRB approval:




Office of Research Compliance  
Institutional Review Board  
2000 Kraft Drive, Suite 2000 (0497)  
Blacksburg, Virginia 24061  
540/231-4991 Fax 540/231-0959  
e-mail moored@vt.edu  
www.irb.vt.edu

FWA00000572( expires 1/20/2010)  
IRB # is IRB00000667

DATE: March 20, 2009

### MEMORANDUM

TO: Doug A. Bowman  
Yi Wang

FROM: David M. Moore 

Approval date: 3/20/2009  
Continuing Review Due Date: 3/5/2010  
Expiration Date: 3/19/2010

SUBJECT: **IRB Expedited Approval:** "Evaluating the Effect of Video Orientation on Event Understanding and Reconstructing Tasks", IRB # 09-260

This memo is regarding the above-mentioned protocol. The proposed research is eligible for expedited review according to the specifications authorized by 45 CFR 46.110 and 21 CFR 56.110. As Chair of the Virginia Tech Institutional Review Board, I have granted approval to the study for a period of 12 months, effective March 20, 2009.

As an investigator of human subjects, your responsibilities include the following:

1. Report promptly proposed changes in previously approved human subject research activities to the IRB, including changes to your study forms, procedures and investigators, regardless of how minor. The proposed changes must not be initiated without IRB review and approval, except where necessary to eliminate apparent immediate hazards to the subjects.
2. Report promptly to the IRB any injuries or other unanticipated or adverse events involving risks or harms to human research subjects or others.
3. Report promptly to the IRB of the study's closing (i.e., data collecting and data analysis complete at Virginia Tech). If the study is to continue past the expiration date (listed above), investigators must submit a request for continuing review prior to the continuing review due date (listed above). It is the researcher's responsibility to obtain re-approval from the IRB before the study's expiration date.
4. If re-approval is not obtained (unless the study has been reported to the IRB as closed) prior to the expiration date, all activities involving human subjects and data analysis must cease immediately, except where necessary to eliminate apparent immediate hazards to the subjects.

### Important:

If you are conducting **federally funded non-exempt research**, please send the applicable OSP/grant proposal to the IRB office, once available. OSP funds may not be released until the IRB has compared and found consistent the proposal and related IRB application.

cc: File

*Invent the Future*

VIRGINIA POLYTECHNIC INSTITUTE UNIVERSITY AND STATE UNIVERSITY  
*An equal opportunity, affirmative action institution*

**Post Task Questionnaire for Contextualized Video test bed:**

6. Please rank the following designs by numbers in terms of “easiness to learn” (1 is the best and 4 is the worst, equal rank is allowed).

Closely associated and camera-oriented video  
Closely associated and user-oriented video  
Remotely associated and camera-oriented video  
Remotely associated and user-oriented video

7. Please rank the following designs by numbers in terms of “easiness to use” (1 is the best and 4 is the worst, equal rank is allowed).

Closely associated and camera-oriented video  
Closely associated and user-oriented video  
Remotely associated and camera-oriented video  
Remotely associated and user-oriented video

8. Please rank the following designs by numbers in terms of “usefulness” (1 is the best and 4 is the worst, equal rank is allowed).

Closely associated and camera-oriented video  
Closely associated and user-oriented video  
Remotely associated and camera-oriented video  
Remotely associated and user-oriented video

**Result Data:**

UserID	Distance	Alignment	TimeAve
1	0	0	3.5375
1	0	1	4.6
1	1	0	3.675
1	1	1	5.1
2	0	0	4.8875
2	0	1	5.775
2	1	0	4.2875
2	1	1	3.9
3	0	0	2.775
3	0	1	4.3125
3	1	0	2.65
3	1	1	2.3125
4	0	0	1.8625
4	0	1	6.1375
4	1	0	3.375
4	1	1	3.8625
5	0	0	5.1

5	0	1	2.5
5	1	0	4.35
5	1	1	4.1
6	0	0	1.45
6	0	1	4.9
6	1	0	1.6875
6	1	1	2.6625
7	0	0	4.2
7	0	1	5.0825
7	1	0	4.15
7	1	1	4.625
8	0	0	1.5375
8	0	1	4.8875
8	1	0	1.9625
8	1	1	4.6
9	0	0	5.6
9	0	1	3.675
9	1	0	3.8
9	1	1	2.2875
10	0	0	8.7
10	0	1	8.0625
10	1	0	7.3375
10	1	1	7.2125
11	0	0	1.4
11	0	1	1.575
11	1	0	2.7375
11	1	1	2.5
12	0	0	5.3125
12	0	1	5.7125
12	1	0	5.575
12	1	1	7.2875
13	0	0	5.5625
13	0	1	7.2625
13	1	0	4.075
13	1	1	6.6125
14	0	0	3.375
14	0	1	3.5125
14	1	0	1.8375
14	1	1	1.7125
15	0	0	2.4
15	0	1	4.7
15	1	0	3.9375
15	1	1	2.0875
16	0	0	2.9875
16	0	1	2.475
16	1	0	3.9125
16	1	1	4
17	0	0	2.3875
17	0	1	5.1
17	1	0	3.4375

17	1	1	1.5875
18	0	0	4.45
18	0	1	5.3
18	1	0	4.25
18	1	1	4.7125
19	0	0	2.75
19	0	1	3.45
19	1	0	2.0625
19	1	1	7.4125
20	0	0	6.7
20	0	1	6.8875
20	1	0	5.7125
20	1	1	6.6
21	0	0	4.125
21	0	1	7.525
21	1	0	4.9375
21	1	1	7.6375
22	0	0	3.6625
22	0	1	5.65
22	1	0	3.3125
22	1	1	3.5875
23	0	0	2.4125
23	0	1	3.525
23	1	0	3.15
23	1	1	1.4375
24	0	0	5.05
24	0	1	2.675
24	1	0	4.375
24	1	1	3.875
25	0	0	4.1
25	0	1	6.8
25	1	0	5.15
25	1	1	4.7
26	0	0	4.4
26	0	1	3.6
26	1	0	3.625
26	1	1	3.1
27	0	0	3.325
27	0	1	4.225
27	1	0	3.9875
27	1	1	6.3
28	0	0	5.2
28	0	1	4.9375
28	1	0	4.5125
28	1	1	5.5
UserID	Distance	Alignment	Score
1	0	0	6
1	0	1	6
1	1	0	7
1	1	1	8

2	0	0	6
2	0	1	5
2	1	0	6
2	1	1	7
3	0	0	8
3	0	1	8
3	1	0	7
3	1	1	6
4	0	0	7
4	0	1	6
4	1	0	8
4	1	1	8
5	0	0	8
5	0	1	6
5	1	0	7
5	1	1	6
6	0	0	8
6	0	1	7
6	1	0	8
6	1	1	7
7	0	0	8
7	0	1	7
7	1	0	6
7	1	1	4
8	0	0	8
8	0	1	8
8	1	0	8
8	1	1	6
9	0	0	8
9	0	1	8
9	1	0	6
9	1	1	8
10	0	0	6
10	0	1	2
10	1	0	6
10	1	1	3
11	0	0	8
11	0	1	8
11	1	0	7
11	1	1	8
12	0	0	8
12	0	1	7
12	1	0	8
12	1	1	7
13	0	0	8
13	0	1	5
13	1	0	6
13	1	1	2
14	0	0	7
14	0	1	8

14	1	0	6
14	1	1	8
15	0	0	8
15	0	1	4
15	1	0	7
15	1	1	4
16	0	0	7
16	0	1	8
16	1	0	8
16	1	1	6
17	0	0	6
17	0	1	5
17	1	0	7
17	1	1	4
18	0	0	6
18	0	1	4
18	1	0	5
18	1	1	2
19	0	0	6
19	0	1	5
19	1	0	7
19	1	1	3
20	0	0	6
20	0	1	7
20	1	0	8
20	1	1	7
21	0	0	3
21	0	1	4
21	1	0	6
21	1	1	4
22	0	0	8
22	0	1	6
22	1	0	8
22	1	1	7
23	0	0	8
23	0	1	7
23	1	0	7
23	1	1	8
24	0	0	7
24	0	1	7
24	1	0	6
24	1	1	7
25	0	0	7
25	0	1	6
25	1	0	8
25	1	1	7
26	0	0	7
26	0	1	6
26	1	0	6
26	1	1	6

27	0	0	6
27	0	1	6
27	1	0	7
27	1	1	5
28	0	0	7
28	0	1	5
28	1	0	6
28	1	1	5

# Appendix C: Experiment 4 (Navigation)

## IRB Approval:



VirginiaTech

Office of Research Compliance  
Institutional Review Board  
2000 Kraft Drive, Suite 2000 (0497)  
Blacksburg, Virginia 24060  
540/231-4606 Fax 540/231-0959  
e-mail [irb@vt.edu](mailto:irb@vt.edu)  
Website: [www.irb.vt.edu](http://www.irb.vt.edu)

### MEMORANDUM

DATE: June 9, 2010

TO: Doug A. Bowman, Yi Wang

FROM: Virginia Tech Institutional Review Board (FWA00000572, expires June 13, 2011)

PROTOCOL TITLE: Evaluate the Effect of Navigation Strategy in Contextualized Videos Task Performance

IRB NUMBER: 10-511

Effective June 9, 2010, the Virginia Tech IRB Chair, Dr. David M. Moore, approved the new protocol for the above-mentioned research protocol.

This approval provides permission to begin the human subject activities outlined in the IRB-approved protocol and supporting documents.

Plans to deviate from the approved protocol and/or supporting documents must be submitted to the IRB as an amendment request and approved by the IRB prior to the implementation of any changes, regardless of how minor, except where necessary to eliminate apparent immediate hazards to the subjects. Report promptly to the IRB any injuries or other unanticipated or adverse events involving risks or harms to human research subjects or others.

All investigators (listed above) are required to comply with the researcher requirements outlined at <http://www.irb.vt.edu/pages/responsibilities.htm> (please review before the commencement of your research).

### PROTOCOL INFORMATION:

Approved as: **Expedited, under 45 CFR 46.110 category(ies) 7**

Protocol Approval Date: 6/9/2010

Protocol Expiration Date: 6/8/2011

Continuing Review Due Date\*: 5/25/2011

\*Date a Continuing Review application is due to the IRB office if human subject activities covered under this protocol, including data analysis, are to continue beyond the Protocol Expiration Date.

### FEDERALLY FUNDED RESEARCH REQUIREMENTS:

Per federal regulations, 45 CFR 46.103(f), the IRB is required to compare all federally funded grant proposals / work statements to the IRB protocol(s) which cover the human research activities included in the proposal / work statement before funds are released. Note that this requirement does not apply to Exempt and Interim IRB protocols, or grants for which VT is not the primary awardee.

The table on the following page indicates whether grant proposals are related to this IRB protocol, and which of the listed proposals, if any, have been compared to this IRB protocol, if required.



**Post Task Questionnaire for Contextualized Video test bed:**

9. Please rank the following designs by numbers in terms of “**easiness to learn**” (1 is the fastest to learn and 4 is the slowest to learn, equal rank is allowed).

Auto Navigation in Detailed View  
Auto Navigation in Overview  
Manual Navigation in Detailed View  
Manual Navigation in Overview

10. Please rank the following designs by numbers in terms of “**easiness to use**” (1 is the easiest to use and 4 is the worst, equal rank is allowed).

Auto Navigation in Detailed View  
Auto Navigation in Overview  
Manual Navigation in Detailed View  
Manual Navigation in Overview

11. Please rank the following designs by numbers in terms of “**usefulness**” (1 is the most useful in helping you do the tasks and 4 is the worst, equal rank is allowed).

Auto Navigation in Detailed View  
Auto Navigation in Overview  
Manual Navigation in Detailed View  
Manual Navigation in Overview

12. Please rate these techniques according to your **preference** to use (1 is the most desired and 4 is the least, equal rank is allowed):

Auto Navigation in Detailed View  
Auto Navigation in Overview  
Manual Navigation in Detailed View  
Manual Navigation in Overview

**Results:**

User	Score	Game	Nav	Video	NavMode	NavContext	Task1	Task2	Task3
1	19	3	4	1	1	1	6	8	4
1	19	3	4	1	2	1	3.5	4.5	4.5
1	19	3	4	1	1	2	5	8	4.5
1	19	3	4	1	2	2	5	8	4.5
2	14	0	1	1	1	1	5.5	7.5	1.0
2	14	0	1	1	2	1	3.5	1.5	1.5

2	14	0	1	1	1	2	5.5	6.5	2.5
2	14	0	1	1	2	2	4.5	6	3
3	6	1	3	1	1	1	6	6.5	2
3	6	1	3	1	2	1	4	3	1.5
3	6	1	3	1	1	2	4.5	5	4
3	6	1	3	1	2	2	5.5	5.5	3.5
4	13	3	3	1	1	1	6	6.5	4
4	13	3	3	1	2	1	3.5	6	4
4	13	3	3	1	1	2	6	7.5	4.5
4	13	3	3	1	2	2	5.5	6	5
5	20	3	3	1	1	1	6	7.5	2
5	20	3	3	1	2	1	4	3	1
5	20	3	3	1	1	2	4.5	6.5	3
5	20	3	3	1	2	2	5.5	6	2.5
6	12	1	2	1	1	1	5	7.5	3
6	12	1	2	1	2	1	6	5.5	3
6	12	1	2	1	1	2	6	6.6	3.5
6	12	1	2	1	2	2	5.5	6	3.5
7	9	4	4	1	1	1	6	7.5	3.5
7	9	4	4	1	2	1	3	6	4
7	9	4	4	1	1	2	5.5	6	3.5
7	9	4	4	1	2	2	6	7	3.5
8	7	1	4	1	1	1	5	5	3.5
8	7	1	4	1	2	1	3.5	4	2.5
8	7	1	4	1	1	2	6	6.5	3.5
8	7	1	4	1	2	2	6	5.5	4
9	10	1	3	1	1	1	4	5.5	1
9	10	1	3	1	2	1	1	2.5	3
9	10	1	3	1	1	2	4.5	6.5	4
9	10	1	3	1	2	2	5.5	5.5	3.5
10	18	4	3	1	1	1	6	6	3
10	18	4	3	1	2	1	4	4	3
10	18	4	3	1	1	2	6	8	3.5
10	18	4	3	1	2	2	5.5	4.5	4
11	19	5	5	4	1	1	5.5	7.5	4.5
11	19	5	5	4	2	1	6	5	5
11	19	5	5	4	1	2	6	7	3.5
11	19	5	5	4	2	2	6	7.5	4
12	9	3	4	1	1	1	6	7	3.0
12	9	3	4	1	2	1	5.5	3	4
12	9	3	4	1	1	2	5.5	7.5	5

12	9	3	4	1	2	2	5.5	7.5	4
13	12	3	3	1	1	1	6	6	3.5
13	12	3	3	1	2	1	3.5	2.5	3.5
13	12	3	3	1	1	2	4	6	3.5
13	12	3	3	1	2	2	4.5	6	3.5
14	5	1	3	1	1	1	6	5.5	4.5
14	5	1	3	1	2	1	4.5	5	3
14	5	1	3	1	1	2	4	6	4.0
14	5	1	3	1	2	2	4.5	5.5	4.5
15	17	3	3	1	1	1	4.5	8	4.5
15	17	3	3	1	2	1	4.5	3	4
15	17	3	3	1	1	2	4.5	7.5	4.5
15	17	3	3	1	2	2	6	5.5	3.5
16	20	4	3	2	1	1	6	7.5	4
16	20	4	3	2	2	1	5.5	6	4.5
16	20	4	3	2	1	2	5.5	6	4.5
16	20	4	3	2	2	2	4.5	8	4

Recorded Interface clicking patterns:

Task	Tech	Video Bank Pick	Context View Pick	Peripheral View Pick	Central Video Mark	Context Video Mark	Peripheral Video Mark	Total Picks
A11	AD	100	78	323	88	38	70	501
A11	MD	148	138	178	129	24	50	464
A11	AO	117	78	339	136	2	89	534
A11	MO	113	79	296	172	0	59	488
1	AD	35	27	113	88	38	70	175
1	MD	69	67	64	129	24	50	200
1	AO	50	34	163	136	2	89	247
1	MO	55	42	130	172	0	59	227
2	AD	34	32	103				169
2	MD	26	39	67				132
2	AO	31	26	98				155
2	MO	29	20	97				146
3	AD	31	19	107				157
3	MD	53	32	47				132
3	AO	36	18	78				132
3	MO	29	17	69				115