Contextualized Videos: Combining Videos with Environment Models to Support Situational Understanding

Yi Wang, David M. Krum, *Member, IEEE*, Enylton M. Coelho, and Doug A. Bowman, *Member, IEEE*

Abstract— Multiple spatially-related videos are increasingly used in security, communication, and other applications. Since it can be difficult to understand the spatial relationships between multiple videos in complex environments (e.g. to predict a person's path through a building), some visualization techniques, such as video texture projection, have been used to aid spatial understanding. In this paper, we identify and begin to characterize an overall class of visualization techniques that combine video with 3D spatial context. This set of techniques, which we call contextualized videos, forms a design palette which must be well understood so that designers can select and use appropriate techniques that address the requirements of particular spatial video tasks. In this paper, we first identify user tasks in video surveillance that are likely to benefit from contextualized videos and discuss the video, model, and navigation related dimensions of the contextualized video design space. We then describe our contextualized video testbed which allows us to explore this design space and compose various video visualizations for evaluation. Finally, we describe the results of our process to identify promising design patterns through user selection of visualization features from the design space, followed by user interviews.

Index Terms— situational awareness, videos, virtual environment models, design space, testbed design and evaluation.

1 Introduction

Video cameras are widely used in many applications: factory monitoring, traffic and security surveillance, telerobotics, telemedicine, and teleconferencing [9, 17]. People observe videos to understand the situation, to make decisions, and to communicate with each other. Technological developments have made video cameras more affordable, allowing multiple cameras to be deployed to cover a larger space or to observe the space from multiple viewpoints. As the system scales, understanding the situation recorded by these videos can become very difficult, because observers often need to mentally reconstruct the spatial relationships between multiple cameras.

Building security surveillance is an illustrative example. In a standard security monitoring system that displays a number of video 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 [15, 19, 22]. A promising hypothesis is that *contextualized videos* – that is, a combination of videos with a model of the 3D environment – will allow observers to see the activities in the videos in their proper locations. In this case, spatial relations are presented in the visualization, allowing some cognitive work to be offloaded onto the perceptual system [21].

With the performance increase in graphics hardware, various contextualized video approaches are now technically feasible. For instance, Sawhney et al. [16] presented a system that projects a video stream onto a 3D model. Other techniques have also been proposed, such as video augmented virtual environments [18] and temporal video visualization [6]. Interesting research questions naturally follow: (1) Which tasks can benefit from contextualized videos? (2) What are the design possibilities to support these tasks? (3) Which specific technique should be used to support a specific type of user task?

- *Yi Wang and Doug A. Bowman are with the Center for HCI at Virginia Tech, E-Mail: {samywang, bowman}@vt.edu.*
- *David M. Krum and Enylton M. Coelho are with the Robert Bosch Research and Technology Center. E-Mail: {David.Krum, Enylton.Coelho}@us.bosch.com.*

Manuscript received 31 March 2007; accepted 1 August 2007; posted online 2 November 2007.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

To identify and clarify these research questions, we have been investigating different contextualized video designs, as well as 3D model visualization techniques, in the context of video surveillance of multi-story buildings. We employ a user-based approach to explore and define the design space for contextualized video and categorize previously suggested techniques. In our design space exploration, multi-story buildings are naturally of interest since modern surveillance systems must be able to scale up to support complex facilities. Furthermore, several visualization challenges, e.g. occlusion and display clutter, emerge if multiple floors must be monitored.

This paper summarizes our work to date on the design possibilities and the benefits of contextualized videos, based on the testbed design and evaluation method (Section 3). We first identify user tasks that are likely to benefit from contextualized videos (Section 4). We then propose a design space for visualizations with contextualized videos (Section 5). Finally, we describe our contextualized video testbed, which allows designers and users to explore a large part of the design space, and identify some promising design patterns through user interviews (Section 6).

Our research has the following contributions:

- We propose a data-based task classification that helps us understand when different contextualized video designs are appropriate.
- We identify the structure of the contextualized video design space and point out research opportunities.
- We explore an important subset of the design space and identify promising design patterns.

2 Related Work

While the image processing and computer vision 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 [5, 6]. 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. Our major concern differs from their

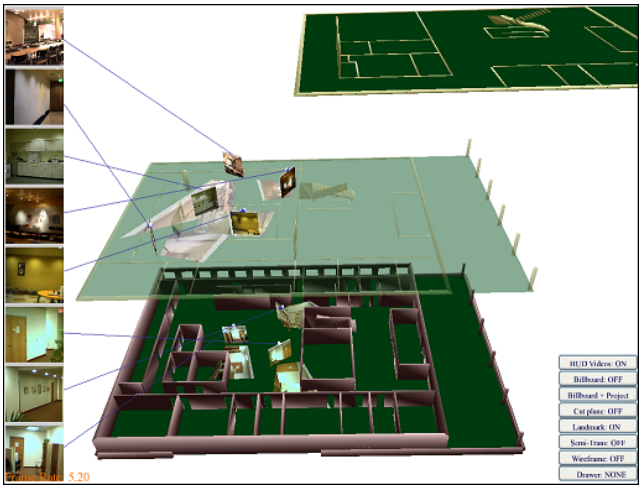

Fig. 1. An overview of the contextualized video testbed. Billboard video, video projection and associated video are shown together with different visualizations of the building model. The menus in the lower-right corner allow users to choose different video placement and visualization techniques on-line.

research goal and focuses on how to present multiple videos in such a way that users can offload the difficulty of spatial relationship reconstruction onto the display.

Sawhney et al. demonstrated the feasibility of projecting multiple videos onto a 3D environment model in their Video Flashlight work [16], while Sebe et al. presented an "Augmented Virtual Environment" system which integrated multiple videos into a 3D context model [18]. They detected moving objects inside the video and visualized them as textured dynamic rectangles moving around in the 3D model. Both of these papers demonstrated the technical feasibility of a particular video placement technique. Girgensohn et al. [8] proposed the Spatial Multi-Video (SMV) player, in which the videos are dynamically arranged on a head-up display according to the user's selection of interest and the spatial proximity of the cameras' field of view. SMV improved user performance in a suspect-tracking task. All of these techniques can be understood as points in the contextualized video design space. In this paper, we focus on exploring and defining this design space.

Several projects in teleconferencing and CSCW have placed live videos into collaborative virtual environments [9, 17]. In [17], the real-time 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.

Combining videos with environment models shares some characteristics with the use of multiple views for visualization tasks, because a video and the model show different aspects of the same data. Baldanodo et al. summarized eight guidelines for the design of multiple views for visualization [1]. These design rules could be used to analytically evaluate different contextualized video designs. For example, the rule of consistency between multiple views implies that we should try to render the landmarks inside the model in the same way as they appear in the video so the user can quickly match the landmarks. Tory et al. investigated how to combine 2D and 3D views for volume visualization and spatial relationship tasks [22, 23]. In [22], 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" [23], 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

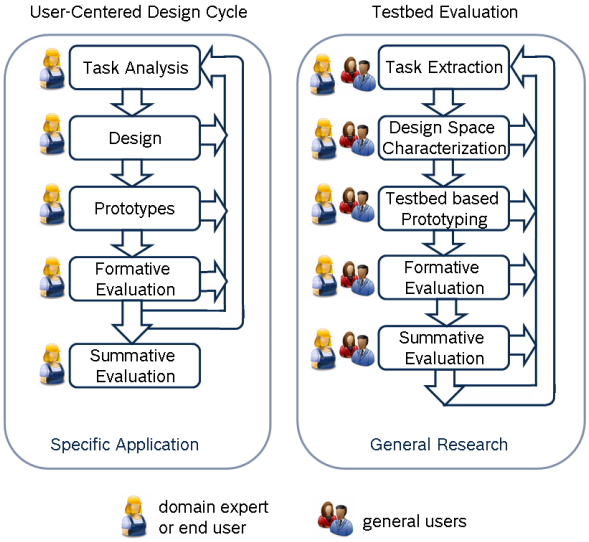

Fig. 2. Comparison of Testbed Evaluation and User-Centered Design.

several 3D views from multiple data sources for visualization is not a well explored problem.

3 Research Method

To help readers understand our work in a broader context, we give an overview of our research method in this section. The purpose of our research is not only to find individual effective visualization designs that prove effective, but also to develop general guidelines or theory that can help visualization designers to 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" [10].

We mainly followed Bowman et al.'s testbed evaluation method [2], which is targeted at inventing and evaluating generic designs instead of application-specific techniques. As a user-based research method, we tried to involve users in every step. However, since we targeted general research questions instead of specific applications, we did not constrain our users to be domain experts. Sometimes it is better to involve non-expert users to eliminate the effect of prior experience. The differences between testbed design and evaluation and user-centered design are discussed at length in [3]. We illustrate the major differences in Fig. 2.

Specifically, our research contains the following steps:

- Task Extraction: Our first step is to extract domain tasks from the applications. We performed a field study of the building security guards' work, exacted low-level tasks and identified some tasks that are likely to benefit from contextualized videos (Section 4).
- Design Space Characterization: We continuously refined our understanding of the contextualized video design space throughout the whole research cycle. The goals are to identify the primary design dimensions and create a framework to generate new designs. We summarize the design space in Section 5.
- Testbed based Prototyping: We started to explore the design space by prototyping a spectrum of video placement designs in a common testbed. We noticed that the occlusion between the model and the videos often caused serious problems for *embedded videos* (contextualized videos where the videos

are placed directly into the 3D model); we then investigated how to adopt scientific and engineering visualization techniques to manage occlusion. This led to a selected set of prototypes along another design dimension: model processing. The testbed allows us to combine the various techniques in many ways. The prototypes are described as examples of the design possibilities in Section 5.

- Formative Evaluation: Using the testbed, we performed a preliminary exploration of the design space with users in order to investigate the users' usage patterns (also known as usage model [13]) for two interacting design dimensions: the video placement dimension and the occlusion management dimension. We also surveyed users' preference over different video placement designs. Although the users' preference could not lead to conclusions, they provided reasonable indications. We describe the evaluation process and early findings in Section 6.
- Summative Evaluation: Based on hypotheses generated by the formative evaluation, we are currently planning a controlled experiment to summatively evaluate the choices along the primary design dimensions.

4 TASK EXTRACTION

In order to extract the real world tasks to motivate our contextualized video designs, we performed a field study, followed by a task categorization, which highlighted the need to combine videos with their environment model. However, we did not attempt to perform a complete survey of the video surveillance domain.

4.1 Field Study

We 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.

We 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, we 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, we 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 figure out the blind areas of the cameras and how to decide whether a person is suspicious or not.

We found that the security guards mainly perform two activities in the control room:

- Monitoring activity: The main responsibilities of security guards are to observe. They frequently scan the videos and try to discover suspicious persons or dangerous situations.
- Tracking activity: Security guards track persons as they move within the building and from video camera to video camera. This allows security guards to determine if people are acting suspiciously, for example, accessing restricted areas or moving equipment. This also allows suspicious individuals to be intercepted and questioned.

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 [24].

4.2 Task Classification

We extracted low-level tasks from the activities and classified the tasks according to users' information requirements: the video content, the environment context (i.e. the model), or the relationship between the videos and the environment.

Type 1 – Video intensive tasks (requiring information presented in videos):

- 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 relating the two videos in a global reference frame.

Type 2 – Model intensive tasks (requiring contextual information largely available from the 3D model):

- 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

Type 3 – Integrative tasks (involving information from both the model and the videos):

- 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. Appears in tracking activities.
- Landmark-based prediction - predict the future location of a person outside the video camera's range, based purely on landmarks. Appears in tracking activities.
- Multi-video registration - judge the spatial relationship between objects in two or more videos.

From the classification, we can see that Type 3 tasks are rooted in the relationship of the videos to the spatial context. When the required information is not present in the users' working memory, they either recall it from their long-term memory or recognize it from external displays. Because our visual system has a very high information bandwidth, recognition may sometimes be more efficient than recall [4, 21].

Novice users may prefer to use external displays. For instance, they may need to look at the 3D environment model in order to make an orientation-based prediction, while an expert user can do the same task on a 2D layout of videos. This is because after an extended period of time at the same site, a security guard can develop a mental model that establishes a correlation of the videos generated by the surveillance cameras and the physical location where the cameras are located. However, as the complexity of the environment scales, some technical assistance would be helpful in supporting the development of such a mental model as well as alleviating the cognitive load of maintaining and referring to such a model.

5 DESIGN SPACE

Contextualized video visualizations combine video and model data to help people understand complex situations. Naturally, layout is a key design dimension. Nonetheless, several other issues are also relevant and need to be addressed. The functional module diagram (Fig. 3) 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 is processed before combining with the model.
- Model Processing 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.

In the rest of this section, we mainly discuss two dimensions, video-model layout and model processing, which were explored using our testbed evaluation method. For the other two dimensions, we only mention some possible directions for further research.

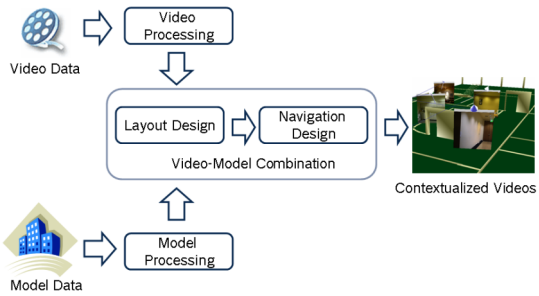

Fig. 3. The contextualized video visualization design framework

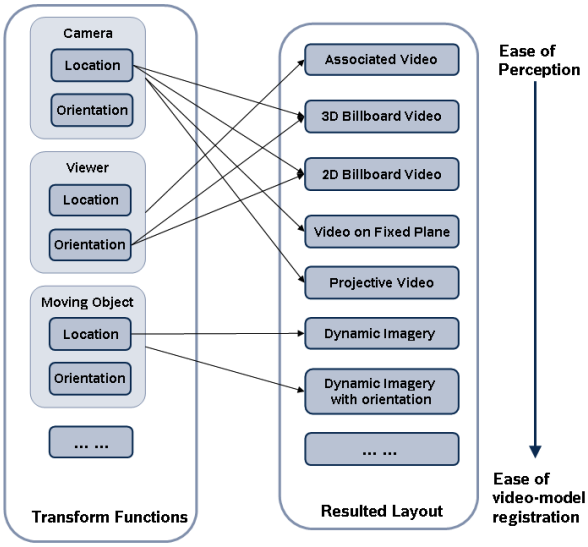

Fig. 4. Mapping between transform functions and the resulting video-model layout designs.

5.1 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.

The video-model layout design module shown in Fig. 4 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}.

$$V_{aug} = M_{layout}\,(V_{post}) \qquad (1)$$

M_{layou} 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. Fig. 5 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. We 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.

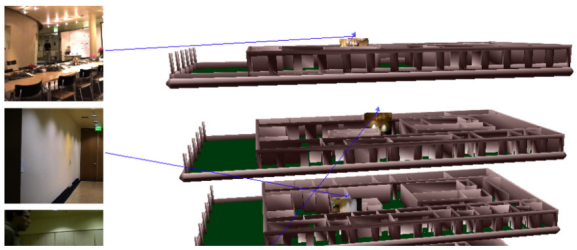

Fig. 5. Associated Video. Callout lines are used to show association.

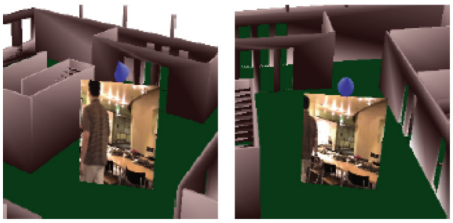

Fig. 6. 2D billboard Video.

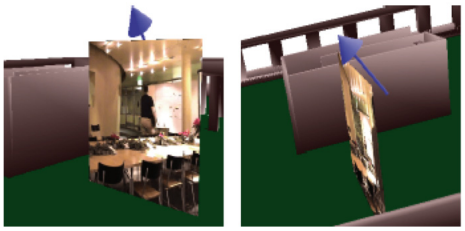

Fig. 7. Video on Fixed-planes. The video is hard to observe in (b).

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 Fig. 5) 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 compensate 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 (Fig. 6). 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 can not 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 (Fig. 7). 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.

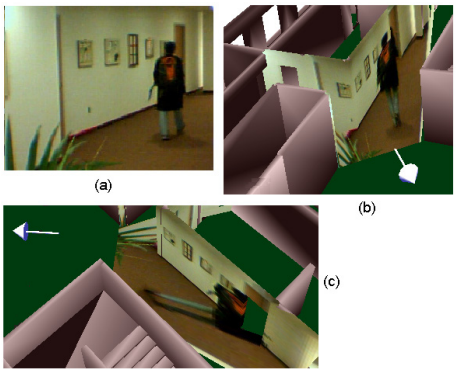

Fig. 8. Video Projection: (a) original video, (b) viewpoint approximately follows the video camera, (c) viewpoint far away from video camera, severe distortion and image fragmentary due to the missing door of the model.

Video Projection – This embedded video design projects videos onto the 3D model in the same way as a projector would (Fig. 8). 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. Fig. 9 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 [16].

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 [18].

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. Fig. 9 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 effort to align with the model. On the other end, there are dynamic imagery and video projections, which are harder to examine but easier 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.

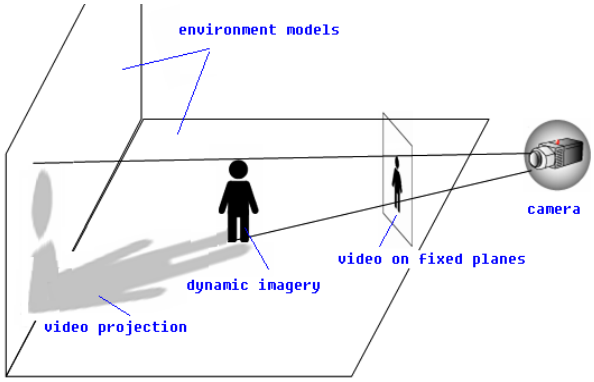

Fig. 9. Video projection, video on fixed planes and dynamic imagery.

5.2 Model Processing

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 processing techniques.

Various techniques to deal with occlusion and clutter have been investigated in areas such as scientific visualization [7, 12] and engineering illustration [11]. 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 (Fig. 10) and rotate-and-shear (Fig. 11). 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 (Fig. 12), wireframe and semitransparency (Fig. 13). 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.

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.

5.3 Video Processing

Video processing and computer vision are both well-developed research areas with numerous research results, many of which can be adopted to create innovative contextualized video designs. Sebe et al.'s "Augmented Virtual Environment" system [18] 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 and recognize objects like humans inside the 3D model. For instance, visualizing the video signatures [5] 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.

5.4 Navigation Design

Interaction, particularly navigation, is a primary component of contextualized video interfaces. Navigation allows the user to select the appropriate viewpoints for examining multiple videos in a single view, minimize image distortion, gain an understanding of the building structure, and find uncluttered, unobstructed views.

Depending on how much user intervention is needed, navigation techniques can be generally categorized into passive (automatic), active (user-controlled) and hybrid [7]. The proper navigation technique for contextualized videos is likely to depend on the possible views that the user will choose when working with the visualization. We identified six different types of navigations for contextualized video visualization:

- Navigation within one video (zoom and pan on video)
- Navigation from one video to another (shift)
- Navigation between two representations of the same video, e.g. the video in the associated view and the same video in the embedded view (as in [1])
- Navigation between a video and a larger view of its nearby context in the model (focus and contextualize)
- Navigation between an detailed view of the model and an overview (zoom on the model)
- Navigation between different parts of the model (travel)

Some navigation modes, e.g. passive navigation between two representations of the same video, have not been fully explored. In our user-based exploration (Section 6), we observed which viewpoints users chose. We plan to further explore this direction in the future.

6 Subspace Exploration

Since we are in the first research cycle as illustrated in Fig. 2, instead of trying to cover the whole design space, we focused on an important subspace which is composed of two major design dimensions: the video-model layout dimension and the model processing dimension. Since there are many possible visualizations that can be created by combining designs from the two dimensions, we asked users to look for potentially useful visualizations while considering realistic tracking tasks. To allow users to freely select viewpoints in the testbed, we implemented a trackball-like navigation technique (similar to [20]) allowing users to rotate the model and zoom in or out. With proper model processing and proper viewpoint, the occlusion effect can be reduced and the advantages of different video-model layout methods can be demonstrated.

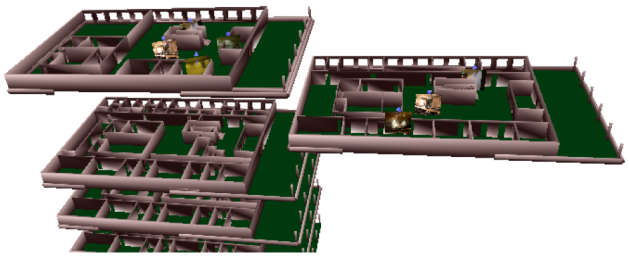

Fig. 10. Drawer.

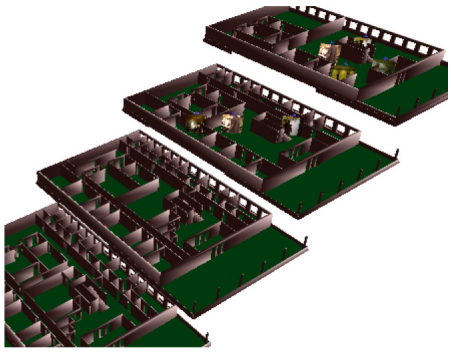

Fig. 11. Rotate-and-Shear.

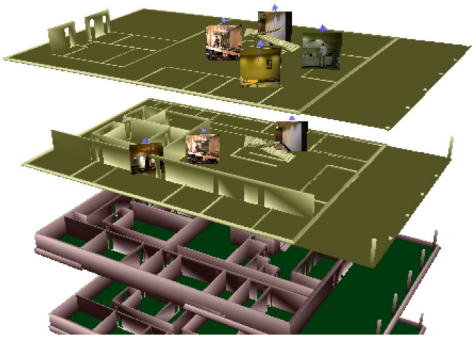

Fig. 12. Landmark.

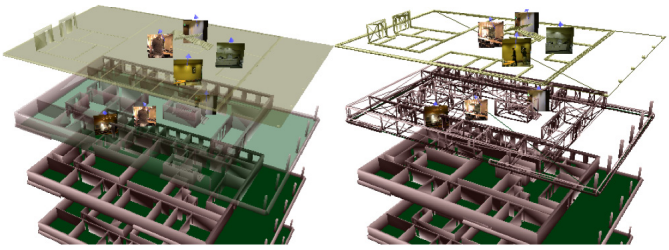

Fig. 13. Semitransparency and wireframe.

6.1 Testbed based Design and Exploration

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. Fig. 1 shows an overview of the testbed. The testbed is mainly implemented using OpenSceneGraph [14]. 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.

The prototyped techniques are described and analyzed in Section 5. 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.

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 an informal formative evaluation, 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.

Eleven users completed the study. We sampled one task from each task category described in Section 4. For each task, the users created a variety of interesting and reasonable visualizations.

Analyzing these visualizations, we discovered some commonly used usage patterns, some of which involved two or more design dimensions, indicating that in some cases a video-model layout method needs proper model processing support to show its advantages. These usage patterns helped us identify a limited number of promising designs to evaluate in a future experiment.

6.2 Promising Usage Patterns

6.2.1 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 can monitor the whole situation of the building.

The following patterns were found:

Pattern 1: Associated videos only

Pattern 2: 2D Billboard + semitransparency / landmark

Not surprisingly, associated videos received higher preference than embedded videos among most users. However, two users preferred embedded videos to associated videos (Fig. 14 (a) and (b)). Both users used 2D billboard. The common reason they gave were 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' orientation was fixed in object space and the users could not find a single view to see all the videos clearly.

6.2.2 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 Pattern 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 (Fig. 14 (c)) and 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 (Fig. 14 (d)). Users preferred to see the 3D walls, which were used to do 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.

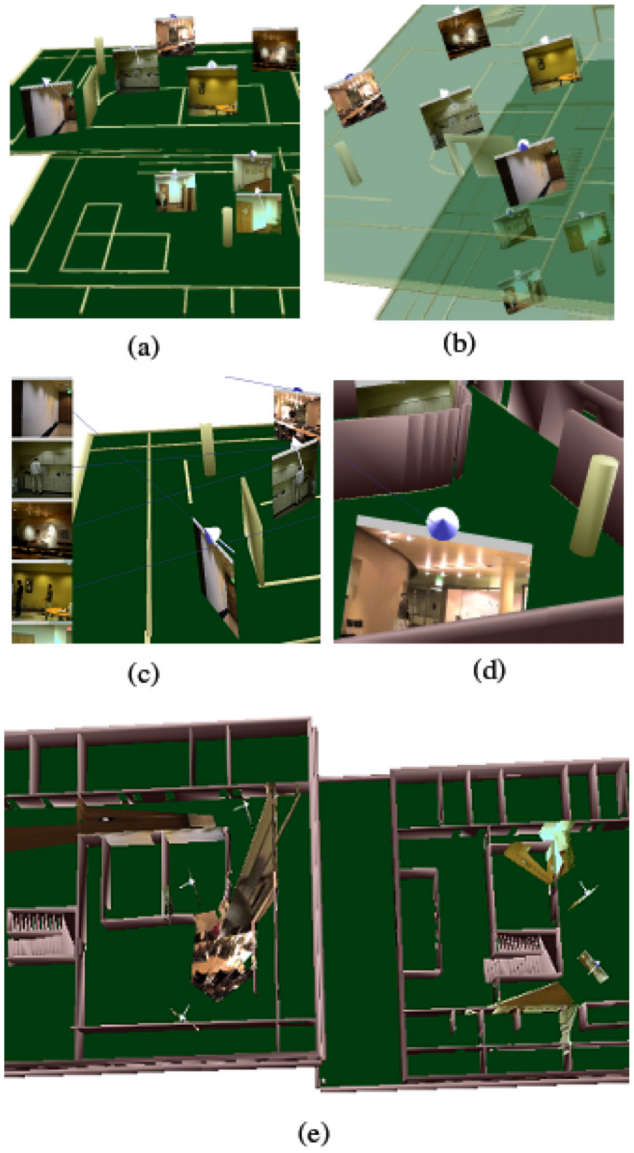

Fig. 14: Usage Patterns: (a) Pattern 2, 2D billboard + landmark view + explosion. (b) Pattern 2, Billboard + semi transparency. (c) Pattern 3, associated video + fixed plane video + landmark. (d) A view behind the camera used in Pattern 5. (e) Pattern 6, video projection + walls.

6.2.3 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 / 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 to see the 3D floors with walls on (Fig. 14 (e)); while others would rather 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.

6.3 Discussion

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 [16] and dynamic imagery [18] were useful for some tasks, they may not be an ideal solution for all tasks.
- Many people used the 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 she detects a suspicious person, the user can track the suspicious person using the embedded videos.

7 Conclusions and Future Work

This paper reports on our exploration of the design space of contextualized video visualization using a testbed design and evaluation approach. By identifying and describing the set of design possibilities, the set of all useful techniques can be identified and characterized. This set of techniques forms a design palette from which the application designer can knowledgably select techniques to match the needs of a particular application and its users.

We have proposed a data-based task classification and identified the tasks that are likely to benefit from contextualized videos. We proposed a design space for visualizations with contextualized videos. We then analyzed the video-model layout problem and the model processing problem in detail. Based on the testbed, we identified some promising design patterns through user interviews.

Our research 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 will be tested in our follow-on summative 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.

We have explored two of the four major design dimensions, but many research opportunities remain. We plan to prototype several designs from the video processing dimension and the navigation dimension as well. We expect to find more interesting results as the design complexity increases. For example, if we extract moving objects from videos, what are the possible designs to embed the extracted imagery into the environment model?

Navigation is also of great importance for contextualized video design. Without usable interaction, the visualization cannot support the users' entire working process. A promising direction is to investigate automatic navigation between multiple views. For example, if the user clicks a video in the associated view, an effective embedded view of the video could be shown automatically.

Another area of study would be the scalability characteristics of different contextualized video designs, especially in terms of number of videos, display size, or model complexity. Our design space creates a framework for future study of these design issues.

Acknowledgements

This work was supported by a grant from Bosch.

References

[1] M.Q.W. Baldonado, A. Woodruff,, and A. Kuchinsky, "Guidelines for Using Multiple Views in Information Visualization," *Proc. Working Conference on Advanced Visual Interfaces*, pp. 110-119, 2000.

[2] D.A. Bowman, D. Johnson, and L. Hodges, "Testbed Evaluation of Virtual Environment Interaction Techniques," *Presence: Teleoperators and Virtual Environments*, vol. 10, no. 1, pp. 75-95, 2001.

[3] D.A. Bowman, J. Gabbard, and D. Hix, "A Survey of Usability Evaluation in Virtual Environments: Classification and Comparison of Methods," *Presence: Teleoperators and Virtual Environments*, vol. 11, no. 4, pp. 404-424, 2002.

[4] S.K. Card, J.D. Mackinlay, et al. *Readings in Information Visualization: Using Vision to Think,* San Francisco, Morgan Kaufmann, 1999.

[5] M. Chen, R. Botchen, R. Hashim, and I. Thornton, "Visual Signatures in Video Visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 5, pp. 1093-1100, 2006.

[6] G. Daniel and M. Chen, "Video Visualization," *Proc. IEEE Visualization*, pp. 409-416, 2003.

[7] N. Elmqvist and P. Tsigas, "A Taxonomy of 3D Occlusion Management Techniques," *Proc. IEEE Virtual Reality*, 2007.

[8] A. Girgensohn, F. Shipman, T. Turner, and L. Wilcox, "Effects of presenting geographic context on tracking activity between cameras," In *Proc. SIGCHI*, pp. 1167-1176, 2007.

[9] J. Han and B. Smith, "CU-SeeMe VR immersive desktop teleconferencing," *Proc. 4th ACM International Conference on Multimedia*, pp. 199-207, 1996.

[10] D. House, A. Bair, and C. Ware, "On the Optimization of Visualizations of Complex Phenomena," *Proc. IEEE Visualization*, pp. 1-21, 2005.

[11] J. Martin., "High Tech Illustration," Addison-Wesley, 1989.

[12] M.J. McGuffin, L. Tancau, and R. Balakrishnan, "Using deformations for browsing volumetric data," *Proc. IEEE Visualization*, pp. 53, 2003.

[13] D.A. Norman, *The Design of Everyday Things*. New York, Doubleday, 1990.

[14] Open Scene Graph. http://www.openscenegraph.org.

[15] H.K. Pillay. "Cognitive Load and Mental Rotation: Structuring Orthographic Projection for Learning and Problem Solving," *Instructional Science*, vol. 22, pp. 91-113, 1994.

[16] H.S. Sawhney, A. Arpa, R. Kumar, S. Samarasekera, M. Aggarwal, S. Hsu, D. Nister, and K. Hanna, "Video Flashlights: Real Time Rendering of Multiple Videos for Immersive Model Visualization", *Proc. 13th Eurographics workshop on Rendering,* pp. 157-168, 2002.

[17] H. Schnädelbach, A. Penn, P. Steadman, S. Benford, B. Koleva and T. Rodden, "Moving Office: Inhabiting a Dynamic Building," *Proc. ACM Conference on Computer Supported Cooperative Work*, pp. 313-322, 2006.

[18] I.O. Sebe, J. Hu, S. You, and U. Neumann, "3D Video Surveillance with Augmented Virtual Environments," *First ACM SIGMM International Workshop on Video Surveillance*, pp. 107-112, 2003.

[19] R.N. Shepard and J. Metzler. "Mental Rotation of Three-Dimensional Objects," *Science*, vol. 171, pp. 701-703, 1971.

[20] K. Shoemake, "ARCBALL: a user interface for specifying three-dimensional orientation using a mouse." In *Proc. Graphics Interface*, pp. 151-156, 1992.

[21] J.J. Thomas and K.A. Cook, *Illuminating the Path: The Research and Development Agenda for Visual Analytics*, IEEE Press, 2005.

[22] M. Tory, "Mental Registration of 2D and 3D Visualizations (an Empirical Study)," *Proc. IEEE Visualization*, pp. 49, 2003.

[23] M. Tory, A. E. Kirkpatrick, M.S. Atkins, T. Moller, "Visualization Task Performance with 2D, 3D, and Combination Displays," *IEEE Transactions on Visualization and Computer Graphics*, pp. 2-13, 2006.

[24] M.C. Velez, D. Silver and M. Tremaine. "Understanding Visualization through Spatial Ability Differences," *Proc. IEEE Visualization*, 2005.