

Temporal Magic Lens: Combined Spatial and Temporal Query and Presentation

Kathy Ryall, Qing Li, and Alan Esenther

TR2005-031 July 2005

Abstract

We introduce the concept of a temporal Magic Lens, a novel interaction technique that supports querying and browsing for video data. Video data is available from an increasingly number of sources, and yet analyzing and processing it is still often a manual, tedious task. A Temporal Magic Lens is an interactive tool that combines spatial and temporal components of video, creating a unified mechanism for analyzing video data; it can be used for viewing real-time video data, as well as for browsing and searching archival data. In this paper, we define the Temporal Magic Lens concept and identify its four key components. We present a sample implementation for each component, and then describe two usage scenarios for a prototype surveillance application.

Interact 2005

This work may not be copied or reproduced in whole or in part for any commercial purpose. Permission to copy in whole or in part without payment of fee is granted for nonprofit educational and research purposes provided that all such whole or partial copies include the following: a notice that such copying is by permission of Mitsubishi Electric Research Laboratories, Inc.; an acknowledgment of the authors and individual contributions to the work; and all applicable portions of the copyright notice. Copying, reproduction, or republishing for any other purpose shall require a license with payment of fee to Mitsubishi Electric Research Laboratories, Inc. All rights reserved.

Temporal Magic Lens: Combined Spatial and Temporal Query and Presentation

Kathy Ryall¹, Qing Li^{1,2}, and Alan Esenther¹

¹ MERL

201 Broadway, Cambridge, MA 02139 USA
{ryall, esenther}@merl.com

² Virginia Tech

Dept. of Computer Science, Blacksburg, VA 24061 USA
qili2@vt.edu

Abstract. We introduce the concept of a *Temporal Magic Lens*, a novel interaction technique that supports querying and browsing for video data. Video data is available from an increasing number of sources, and yet analyzing and processing it is still often a manual, tedious task. A Temporal Magic Lens is an interactive tool that combines spatial and temporal components of video, creating a unified mechanism for analyzing video data; it can be used for viewing real-time video data, as well as for browsing and searching archival data. In this paper, we define the Temporal Magic Lens concept and identify its four key components. We present a sample implementation for each component, and then describe two usage scenarios for a prototype surveillance application.

1 Introduction

Video data is available from an increasing number of sources, such as entertainment, web cams, home video, and surveillance systems. With more and more video data available, the task of understanding, analyzing, summarizing, or even finding an event of interest becomes a daunting task. Although there are a number of automatic techniques for event detection and object tracking, these do not solve the problem; while they may reduce the amount of data by converting a raw video stream into a set of abstract objects (e.g., events, people, and trajectories), there is still inherently a large amount of data for a person to deal with. Thus there is a need for interface and interaction support to deal with the abstract or meta-data extracted by the automatic techniques, as well as with the raw video data.

Analyzing and processing video data is still often a manual, tedious task. In particular, looking for a specific event in a large data stream can be very difficult. When dealing with “personal” video (i.e., content that you filmed, or a movie that you have seen before) it may be possible to exploit one’s own knowledge to help manage the search. In contrast, when dealing with arbitrary video, such as surveillance video or a movie that you have not seen before, finding an individual event may be like looking for a needle in a haystack. While methods exist to support querying by time (e.g., show me all the people that entered a building between 1:00 and 1:15) or by location (e.g., show me all the people that entered a particular room) few techniques attempt to

combine the spatial and temporal components inherent in video data into a unified query and presentation.

We introduce the concept of a *Temporal Magic Lens* to support a combined spatial and temporal query, enabling a novel presentation of the query results to aid people with their video data by serving as a “window in time” into their data. Users choose a *region* of interest and a *time period* of interest, and the computer provides a summary of that location in that particular time period. While a Temporal Magic Lens provides important dynamic information, and is intended to be used dynamically to query and browse streams of video data, it can also be used to provide static snapshots of different subsets of the data. The Temporal Magic Lens concept will be a powerful tool in video information exploration across many domains. For example, for any physical activity (e.g., sports, dance, T'ai Chi) users can quickly detect the minute difference between the body movements of coach's from time to time and thus learn it more quickly. In video editing, we can easily pick out frames in which big differences appear or detect small changes among a number of frames in which all objects seem to remain unchanged.

In Section 2 we present related work that has motivated our research. In Section 3 we define the Temporal Magic Lens concept along with a sample implementation of each of its four key components. We present two sample usage scenarios for our system in Section 4, and conclude in Section 5 with a summary of our work and a discussion of open issues and future directions.

2 Related Work

We first review general video summarization techniques, followed by visualization techniques focusing on the issues of video query, browsing and presentation. Finally, we discuss previous work on magic lenses.

2.1 Video summarization techniques

Rendering animated pictures is both time and space consuming. In some cases, there is a risk of exhausting the whole system if appropriate resources are not available (i.e., not enough memory) or if resources are not appropriately managed (i.e., memory leaks). To solve this problem it is often highly desirable to present appropriate summarization or abstraction of the raw video data to users.

Video summarization, sometimes referred to as video parsing, is the first step in video information processing. In this step an index is created and important features of the video are extracted for later analysis. In general it includes two parts: temporal segmentation and key-frame abstraction. Temporal segmentation detects boundaries between consecutive camera shots or event sequences. Key-frame abstraction maps an entire segment to some small number of representative images, usually called key-frames – still images which best represent the video content in an abstracted manner. An index may be constructed from key-frames, and retrieval queries may be directed at key-frames, which can subsequently be displayed for browsing purposes.

A common approach is to integrate the key-frame extraction process with the processes of segmentation. When a new shot is identified, the key-frame extraction process is invoked, using parameters calculated during segmentation [15]. The challenge, which is also the focus of much previous research, is how to generate video summarization automatically based on context. Summarization algorithms can either rely on low-level image features such as color (brightness and dominant color etc.) and motion activity, geometry (i.e., size and location) or more advanced semantic analysis such as pattern detection [4,11]. While this technique is useful and powerful, the time it takes to process the raw images may make it undesirable for a system requiring real-time analysis. As described in Section 3, a Temporal Magic Lens can be used for real-time analysis, when no additional information (or meta-data) is available; it can also exploit the extra information when it is available.

2.2 Video Visualization

We started our work by seeking appropriate visualization techniques to present video data, combining both the raw images and any extracted information (i.e., meta-data). Meta-data is usually associated with individual frames, and may include object counts, object bounding boxes, motion vectors or other derived information. The algorithms used to create the meta-data often require multiple frames, and may not run in real-time. We quickly found it is possible to use our techniques to summarize the raw data directly and present summary views without the help of meta-data.

There is a long history of work in creating efficient methods to present and browse video data. In the DIVA system, for example, the spatial view is arranged in the center and temporal views appear on the sides, which simulates 3D visualization to give a sense of past and future [9]. The side view displays data streams (or summaries). While this approach is suitable for displaying many kinds of data streams in parallel, it requires large display space and meta-data to be available. It can also only be used to visualize data from a particular point in one direction (e.g., the past or the future).

In the Rapid Serial Visual Presentation Techniques (RSVP) project, Wittenburg *et al.* investigated various space layouts of video frames to help TV channel surfing and video summarization by frame selection [13]. Users can select a number of key frames by clicking the displayed frames. The system presents the video stream in temporal context in “time tunnels” and helps users to navigate to points of interest quickly and precisely. While RSVP provides convenient presentation and navigation methods for video streams, it does not support a direct query method in the same way that our Temporal Magic Lens does – RSVP is more concerned with navigation and browsing rather than posing or answering specific user-queries. We experimented with several RSVP techniques prior to developing our temporal magic lens concept.

Daniel and Chen presented a method for “summarizing” video sequences using opacity and color transfer functions in volume visualization [3]. In particular, they utilized color transfer functions to indicate different magnitudes of changes, or to remove parts of spatial object. This is one of the few examples that we have found that summarizes video data directly by solely utilizing visualization techniques. As with the DIVA system, this technique requires meta-data to be available.

Freeman and Zhang [5] introduced “shape-time photography,” a novel method to describe shape relationships over time in a single photograph. Using data gathered from a stationary stereo camera, they compute a composite image that can be used to summarize events or for instructional materials (i.e., in lieu of illustrations). Their work shares many of the same goals as our current work, namely capturing changes in time in a single image, and determining which pixels to contribute to the final image. In our work, however, we focus on the interface and interaction technique that allow users to select the spatial and temporal regions of interest, and the larger context in which a compositing technique may be used. Freeman and Zhang focus on the compositing itself. While their compositing technique could be used as part of our temporal magic lens, it is not suitable for much of today’s video as it requires stereo images.

For practical video analysis, it is important to present information at different temporal scales in video editing and analysis. For example, users may only have two minutes to view a half-hour video clip. The Hierarchical Video Magnifier provides users with a detail+context view by using successive timelines [10]. Initially it uses a timeline to represent the total duration of the video clip. Users can select a portion of the timeline and expand it to a second timeline. The timelines create an explicit spatial hierarchical structure of the video source. The disadvantage is that it is hard to scale up. As with the RSVP approach, this work is more concerned with browsing and navigation support rather than explicit querying. Alternatively, Silver2 [8] system applies a spatial magic lens to the timeline to support semantic zooming in the timeline, and hence provides a “fisheye” view.

2.3 Magic Lens

We propose a novel magic lens technique that can be integrated into video analysis systems to support video navigation and exploration with or without meta-data. The magic lenses with which we are familiar to date are spatial in nature – our Temporal Magic Lens adds a new dimension (time) to determine which content should be displayed in the lens. It combines temporal and spatial aspects of a video-based query, and provides a unified presentation of the query results.

The “magic lens” concept was introduced by Bier et. al. as a see-through interface in 1993. A magic lens is a screen region that semantically transforms the content underneath it; users can move the lens to control what region is affected. In practice, a magic lens is a composable visual filter that may be used as an interactive visualization technique [2]. A lens may act as a magnifying glass, zooming in on the content on which it is placed. It may also function as an x-ray tool to reveal otherwise hidden information. Multiple lenses may be stacked on top of one another to provide a composition of the individual functionality. In some cases, different lens ordering will generate different result. A Magic Lens interface offers many advantages over traditional controls. It reduces dedicated screen space while providing the ability to view context and detail simultaneously. It enhances data of interest and suppresses distracting information, and thus reduces execution errors. The Magic Lens concept has been used in many applications, such as Pad++ [1], Debugging Lenses [7] and Document

Lens [12]. Our work builds upon this earlier work by providing access to temporal information, and is appropriate to help video querying, browsing and presentation.

3 Temporal Magic Lenses

A Temporal Magic Lens combines spatial and temporal query components into a single query. A user picks a region of interest and a time period of interest, and then the computer present a composite of some subset of the frames to give a summary of that location in that particular time period. The choice of parameter settings may be done manually (by the user) or more automatically (by the system). We believe a Temporal Magic Lens should also provide users with techniques and interactions to quickly understand and further analyze the data displayed within it. In this section, we first describe the concept of a Temporal Magic Lens by defining its four key components. We then describe a sample implementation of a Temporal Magic Lens for an example video surveillance application prototype.

3.1 Key Components

A Temporal Magic Lens has four key components: the spatial query, the temporal query, methods for rendering/compositing multiple video frames, and mechanisms to support drill-down in the data.

Spatial Query: The spatial query indicates the spatial region of interest. It is a contiguous set of pixels, and may be of any shape (e.g., rectangle, oval, blob, etc.); it defines the physical boundaries of the magic lens, and is also the area in which the query results are displayed. The user may select the region manually (i.e., by drawing directly over an image) or the system may automatically select the region (i.e., using an event or motion detection algorithm). To date we have only worked with fixed camera video data, and so the spatial query is specified relative to a fixed (geographical) window. The question of defining the spatial query for panning or mobile cameras is left for future work. We have found that while the initial region could be created using the same manual techniques described above, it is our experience that the region should track the camera's motion (or the motion of objects in the camera's view), and not remain fixed relative to the window frame. World-coordinate frames, if available, might also be useful.

Temporal Query: The temporal query indicates the temporal region of interest. Specifying the start time and duration of the time window may also be done using manual or automated methods. For real-time data the temporal window can only extend into the past. For archival data, it may extend into the past, future, or some combination of the two. In addition to the depth of the temporal window, there is also a question of the granularity of the data within that window – how much data to display or summarize. Within the temporal window we must also specify or determine how many frames (or layers) should be included in the summary.

Rendering/Compositing: Once we have the frames that result from the temporal and spatial query, we must consider the method for compositing the frames for presentation to the user. In essence, we will blend multiple frames to create a single

frame, encapsulating or summarizing the query results. When compositing multiple images, the stacking order and the weighting of each frame controls the final appearance of the Temporal Magic Lens. Rather than composite the entire region of interest, it may also be advantageous to identify clipping regions within the larger spatial query. The compositing issue is addressed in more detail by Freeman and Zhang [5].

Drill-Down: The contents (i.e., composited image) of a Temporal Magic Lens contain data across a range of times. Users may want to distinguish which part of the frame comes from which point in time, and may want to further explore that time period in more detail. Thus the Temporal Magic Lens should provide easy interactions for a user to better understand and explore the summary data provided in the composited view, including methods to determine which frames objects/content came from, and easy visual indications of temporal distance. For example, in addition to the composited frame (in the Temporal Magic Lens area), the interface could provide a secondary region with thumbnails (one for each frame that contributed to the summary) or alternatively could provide some interactive feedback using mouse-overs.

3.2 Sample Temporal Magic Lens

We have implemented a sample Temporal Magic Lens for a video surveillance prototype. Here we describe implementations for each of the core components.

Spatial Query: We support manual specification of the spatial query, and use a rectangular lens shape. Users select the region of interest by drawing rectangular areas in the video player window. Its size and position may be adjusted at any time. A single Temporal Magic Lens is shown in the video player in Figure 1a.

Temporal Query: We have implemented a timeline slider widget (Figure 1b) to support specifying the temporal window along with the number of layers of interest within that window. We created our own specialized widget instead of using general sliders to enable pixel-based moving and selection. The vertical line represents the time pointer and will automatically move forward/backward in auto-play mode. It can also be moved freely along the timeline by the user, moving either forward or backward in time. This control is similar to existing widgets appearing in many popular video player applications with an additional function to set the temporal depth of the magic lens. In this implementation the current frame (longer vertical bar) serves as one end of the time interval while the black bar indicates the temporal depth – how far temporally the magic lens can go (the depth of layers). In an alternate version the time period can bracket the current frame, extending in both the past and the future directions. The slider also provides information such as the number of frames composited in the magic lens and their relative temporal distance and weightings.

In the example shown in Figures 1a-c, the temporal magic lens composites eight frames. Each frame is represented by a small rectangle between the temporal pointer and the temporal depth bar. The inter-bar spacing indicates relative temporal positioning, and the bars' gray-levels indicate their relative weighting for compositing. In this example the frames are evenly-spaced. The more “recent” frames are more heavily weighted, making them clearer in the final (composited) image displayed in the lens. The weighting, spacing and “recent” details are discussed in more detail below.

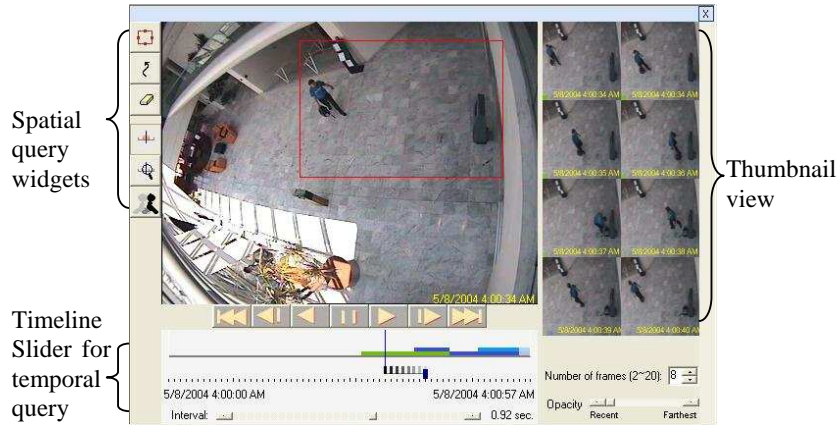


Fig. 1a. The Temporal Magic Lens includes four basic widgets: video player with spatial query widgets (upper left quadrant), timeline sliders to control the temporal query (bottom left corner), thumbnail views for drill down (right-hand side), and compositing/rendering techniques (inset of video player, and below in Figure 1c).

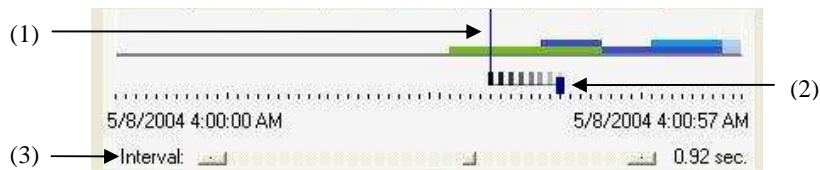


Fig. 1b. A closer look at the Timeline Slider from Fig. 2a: Our composite widget includes (1) Time Pointer, (2) Temporal Depth Bar, and 3) Interval Scrollbar.



Fig. 1c. Temporal Magic Lens detail from Fig 1a.

Left: Defining the temporal magic lens. Right: Seeing through the temporal magic lens; the ghosting gives a preview of “future” events. The effect is more evident from the actual motion of the video in the prototype system when viewed interactively, than in the static images shown here.

Rendering/Compositing: Choosing appropriate rendering opacity level is rather tricky when many images are overlaid together. Traditional methods primarily ad-

dress compositing fixed, unrelated images [6, 14]. Their goal is to make either foreground or background more clear while maintaining the visual cues in the whole context. For video data, pictures taken over a time period are normally played sequentially, one by one. Compositing a series of pictures from video data which include moving objects requires us to not only be able to make a particular frame of interest the clearest (i.e., the most recent or the farthest away temporally), but also clearly show the continuous moving sequence as well.

We use an interpolation approach to composite the view of the magic lens. Given the starting and ending pictures (a time period) and the number of pictures to composite, we calculate the time interval between consecutive composite frames and select them from the disk. The selected images are further filtered by the number of objects detected in the frame. Currently frames containing no object (as determined by meta-data, if available) are discarded as they are essentially background images and do not contribute any new information from the user's perspective. Their exclusion brings us a clearer view by minimizing any blurring or fading of the context. We also tried to only composite regions with objects (local compositing) in hope of further reducing the blurs. However, the result is not as good as expected due to the roughness of the meta-data. The detected object rectangles are so large that they overlap each other. As a result, local compositing does not make any big difference from global compositing and it brings even more confusion with sharp edges around each object rectangle.

We evaluated three techniques for creating an appropriate transparency effect when compositing multiple frames; ideally the temporal lens should enable users to see all composited frames as clearly as possible or with emphasis on a subset. Based on our experiment, we determined that simply applying a fixed opacity value is unacceptable. For a fixed opacity value, we observed that images will be perceived differently – images above will always be clearer than those below. This method can either show the most recent images or the farthest ones more distinctively than others, but showing all images with the same perceived opacity level, as might be desirable when users are interested in observing the overall motion sequence, is impossible.

The second method we tried composites images according to different orders to emphasize a subset of images (order switch method). We let the *Opacity* scrollbar control to the number of composited frames and assign a fixed alpha value to each image (0.5). When users want to see the farthest phantom (i.e., the least recent temporally), the system will draw image with the most recent timestamp first (i.e., closest to the time pointer), and less recent one next. As a result, the farthest image (which was drawn last) will be the clearest one since it is rendered on top of the others. For example, given a total of six overlaid frames, if the last frame is the most interesting frame, the overlaying order will be 1, 2, 3, 4, 5, 6 (from bottom to top); if the fifth one is the most interesting frame, the order should be 1, 2, 3, 6, 4, 5 and so on. This method makes the interesting frame and its peripherals distinct from others. The disadvantage is that it still cannot provide a balanced transparency view; even the neighborhood can be blurred and images far away will completely vanished. Users perceive fewer phantoms than the real number being composited.

In the third method, we used a heuristic equation to control the opacity level for each composited frame according to their overlaying order, so that all images can be perceived as with the same opacity value (Equation 1, 2).

$$\sum_{i=1}^n \alpha_i \times \frac{i}{n} = 1 \quad (1)$$

$$\alpha_i \times \frac{i}{n} = \alpha_j \times \frac{j}{n} \quad (i \neq j) \quad (2)$$

α_i represents the alpha value of a given pixel on the i^{th} layer, n represents the number of total composited images. i/n stands for the weight of α_i and the multiplication stands for the heuristic value (user perceived value) on a certain layer. The condition when users perceived all composite images as with the same opacity level is defined as a balanced view. Thus we solve $\alpha_i = 1/i$. The equation maintains a continuous transparency spectrum, from the least recent (i.e., oldest) phantom being the clearest to a balanced view, and to the most recent image (i.e., newest) becoming most distinctive as shown in Figure 2.



Fig. 2a. The four single video frames to be composited into a single image.



Fig. 2b. The transparency spectrum illustrating the effect of different weightings. The frame on the left emphasizes the (temporally) most distant frame; the frame on the right emphasizes the most recent. The balanced view falls in the center. The five sample frames from different points along the spectrum *do not* represent a video sequence; each frame is a composite image showing the different weighting schemes. Only one frame would be displayed within the Temporal Magic Lens.

For example, when compositing two layers, both images are seen as clear as possible when the alpha value for the bottom image is equal to 100%, and the alpha value for the top image is equal to 50%, both images are perceived as having $\alpha = 50\%$. In an informal study, given three images, people perceived all three images to be equally weighted when in fact they are weighted with $\alpha_1 = 100\%$ (bottom), $\alpha_2 = 50\%$ (middle) and $\alpha_3 = 33\%$ (top). Starting from the balanced view, users can decide which side (most recently or farthest) should be more distinguished from others by dragging the Opacity scrollbar, located below the thumbnail view in Figure 1a.

Drill-Down: We provide thumbnails next to the video player to display frames contributing to the composited image in the Temporal Lens (Figure 1a) to enable

users to interactively explore the data. Dynamically mousing over a thumbnail allows users to better understand the temporal nature of the data; the system's behavior depends upon the availability of meta-data. When no meta-data is available, it can reorder the layers of the composited image, moving that (active) thumbnail to the top so that it becomes more distinct. When meta-data is available, the system will only show the objects within that (active) thumbnail's frame in the video player. Users can observe the change sequence by moving the cursor over a series of thumbnails and hence get better understanding of the compositing process. We also draw object indicators on each thumbnail to show detailed information about the detected objects.

4 Example System

We applied the temporal magic lens concept in a surveillance system and the result is very promising. There are a number of tasks that any surveillance system should provide for its users: helping users to quickly search video frames (i.e. go to the pictures taken in a certain time period), detecting outliers quickly and precisely (i.e. decide whether an outlier exists, and if so when and where it happens), and identifying suspect objects easily.

To support these tasks, we extended the temporal magic lens by exploring alternative implementations of three of its components. Our investigation into compositing techniques was described in the previous section.

Spatial Query: We augmented the spatial query with a *trajectory query*, which defines a spatial region of interest, and a filter to further constrain the query. This tool allows users to define a pattern of movement (the path and the direction) so that only trajectories with similar patterns will be displayed in the magic lens. A trajectory query is specified by drawing a sample trajectory, a straight line (rather than a rectangle) in the video player. Future work will explore alternate methods for specifying arbitrary trajectories, including non-linear paths and changes in speed. Trajectory queries also require meta-data to be available.

Temporal Query: Two functions were added to support temporal queries. First, a *time compression control* enables users to see the video clip in different detail levels. This tool allows users to decide the time interval by scrolling the *Interval* scrollbar at the bottom of the timeline slider (Figure 1b). For example, a user is able to view a one-hour video clip in just three minutes by changing the time interval, in essence using "fast-forwarding." Instead of playing the frames one by one (taken every 0.05 second), the system plays every twentieth frame with a one second interval. Second we added *temporal zooming* support -- users can zoom into a given time period to see more detailed information. In 3a, the "zoom in" area is shown as a pink-shaded rectangle above the timeline slider. The time interval in the "zoom in" area is controlled by a separate interval scrollbar at the bottom of the object histogram.

Drill-Down: Due to the special characteristics of the surveillance system, we also incorporated an *object query* function, by modifying and extending three components. First, we added an object histogram as a background in the timeline, with the height representing the number of objects and colors distinguishing them. From it users can easily tell how many objects are in a given time period, when the object appears, and

how long it lasts. Second, we added an interactive object histogram component (the lower right portion of Figure 3a). Once users select a particular time period in the timeline slider, the corresponding histograms are shown as an interactive histogram. Each object is assigned a distinct color. Users can select different bars in the histogram and a path will appear in the video player indicating its trajectory. Third, we added colorful dots in the lower left corner of each thumbnail to indicate how many objects are detected in the frame. Every object is assigned a unique color consistent with the object histogram. Mousing over a thumbnail will display a rectangle representing the locations and sizes of those objects in the video player. The object query itself merits a thorough study, and is beyond the scope of this paper.

4.1 Usage Scenario One

For the example shown in Figure 1, the video sequence records the event that a man left his bag in the courtyard. It is a one-minute video with pictures taken 25fps. A quick look at the histogram in the timeline slider tells us the activity happens in the later half of the video clip so it is reasonable for us to slide the time pointer directly to the interesting time period. Our task is to detect the actual movement of the man. We first defined the magic lens to let it only cover the center region of the picture. The default number of frames to composite with the magic lens is six with equal perceived opacity. Figure 1c displays the result; we can easily identify the path along which the man walked, and the location in which the man dropped his bag.

4.2 Usage Scenario Two

Another scenario is visualizing the pictures obtained from a highway surveillance system. This video clip covers footage taken for five minutes along a highway. Suppose we only want to take 6 seconds to look at the whole movie. We need to specify a longer time interval (play every 50th frame) using the time interval scrollbar. Figure 3a-b shows the initial interval scrollbar. In Figure 3c we have dragged the interval scrollbar to the right. The summarized result is much clearer than the initial data.

From the long pink bar in Figure 3c we detect that there is an outlier in the last quarter of the video; some object stayed for a long time in the highway. Highlighting that time period by selecting it with the mouse provides an object histogram. Mousing over the bar of interest shows the trajectory of that object (Figure 3a). We can see from the thumbnail in the upper left corner that it is a red car and it stopped on the right curb for quite a long time. In Figure 3a, the diameter of the circle indicates how long the object stays in a place.

Compositing many frames may compromise the clarity of the view. In Figure 4 we defined a rectangular magic lens and the task is to compare the current frame with one of the future frames to see the difference. We can put cursor over the given frame on the right side and objects detected in that frame will be displayed as a rectangle with unique colors in the video player (on the left side). The current frame contains two vehicles, but there will be four vehicles twenty-one seconds later. The red car will still be visible and has driven back a few yards.



Fig. 3a. Object query - putting the cursor onto the object histogram shows the trajectory of the corresponding object in the video player. The object histogram spans the bottom of the application window. The long cyan bar at the bottom of the object histogram indicates an object present for a long period of time. Mousing over that bar brings up a thumbnail of that object (a red car), shown in the upper left-hand corner.



Fig. 3b. Larger view of the timeline slider with object histogram from Fig. 3(a).

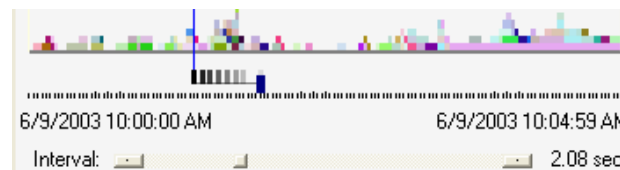


Fig. 3c. Timeline slider with object histogram after summarization (see Section 4.2). The interval scrollbar has been dragged to the right, indicating a longer time interval between frames (0.04 seconds in (a), 2.08 seconds in (b)). We can now see (from the long pink bar in the histogram) that some object lingered on the highway for a prolonged period of time.

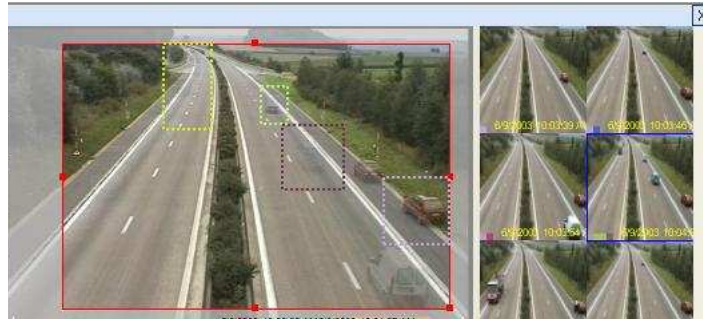


Fig. 4. Object query. Moving cursor over a thumbnail will show the detected objects in the particular frame in the video player.

5 Conclusion

In this paper we have introduced the concept of a Temporal Magic Lens, a novel interaction technique that supports querying and browsing for video. It can be used to search and browse video surveillance data, home video, or any other recorded digital content. By combining spatial and temporal aspects of the data, it creates a unified display that supports people's browsing and querying, and provides both dynamic and static views into their video data. We have described its four key components, along with our first investigation into implementations for each. The two scenarios presented in this paper illustrate how to apply the Temporal Magic Lens concept in a surveillance system. Although our initial results are limited due to the roughness of the available meta-data, they suggest better results will follow with more accurate meta-data (e.g., enabling local instead of global compositing). More importantly, they illustrate that the Temporal Magic Lens may be used even in cases where no meta-data is available.

Our current prototype is an initial exploration of the Temporal Magic Lens concept; at present it is a vehicle for our own research rather than an end-user's tool (e.g., our compositing techniques still have many degrees of freedom). Additional study is needed to determine and evaluate the best implementations for each component. In the cases where multiple techniques should be supported (e.g., image compositing), easier-to-understand (and perhaps more intuitive) controls need to be developed. Moreover, by defining the Temporal Magic Lens concept and clearly delineating its four components, we provide a foundation for other researchers to apply their technology (e.g., meta-data extraction, object recognition, compositing techniques, and interactive timelines) to the domain of video.

There are a number of future directions to explore. We would like to support defining and viewing multiple temporal magic lenses simultaneously. In addition, integrating dynamic interaction to the Temporal Magic Lens concept will increase its query power; its filtering function should not only consider the existence of objects, but also the dwell time and other users-specific interests. Finally, to date we have only applied the tool to analyze data from a single, fixed camera. Compositing frames from multiple cameras may generate more interesting results. As previously noted,

supporting non-fixed cameras (i.e., pan-tilt-zoom cameras) will require more sophisticated compositing techniques as well as additional meta-data. The Temporal Magic Lens provides a framework for video query and presentation for us and others to build upon, and for designing next-generation video information exploration systems.

Acknowledgements

We thank Tom Lanning and Kent Wittengburg for their suggestions and discussion on this project, Fatih Porikli and Ajay Divakaran for their data and input, Bill Buxton for his comments on this work, and Joe Marks for feedback on this paper.

References

1. Bederson, B. B. and Hollan, J. "Pad++: a zooming graphical interface for exploring alternate interface physics," *Proceedings of UIST '94*, pp. 17-26, 1994.
2. Bier, E. A., Fishkin, K., Pier, K. and Stone, M. C. "Toolglass and magic lenses: the see-through interface," *Proceedings of SIGGRAPH'93*, pp. 73-80, 1993.
3. Daniel, G. and Chen, M. "Video visualization," *Proceedings IEEE Visualization 2003*, pp. 409-416.
4. Divakaran, A., Pekar, K. A., Radharkishnan, R., Xiong, Z. and Cabasson, R. "Video Summarization Using MPEG-7 Motion Activity and Audio Descriptors," *Video Mining*, Rosenfeld, A.; Doermann, D.; DeMenthon, D., October 2003.
5. Freeman, W., and Zhang, H. "Shape-Time Photography," *Proceedings of CVPR 2003*.
6. Harrison, B.L., Ishii, H., Vicente, K. and Buxton, W. "Transparent Layered User Interfaces: An Evaluation of a Display Design Space to Enhance Focused and Divided Attention," *Proceedings of CHI'95*, pp. 317-324. .
7. Hudson, S., Rodenstein, R. and Smith, I. "Debugging Lenses: A New Class of Transparent Tools for User Interface Debugging," *Proceedings of UIST'97*, pp.179-187.
8. Long, A. C., Myers, B. A., Casares, J, Stevens, S. M. and Corbett, A. "Video Editing Using Lenses and Semantic Zooming," <http://www-2.cs.cmu.edu/~silver/silver2.pdf>, 2004.
9. Mackay, W.E., and Beaudouin-Lafon, M. "DIVA: Exploratory Data Analysis with Multimedia Streams," *Proceedings CHI'98*, pp. 416-423, 1998.
10. Mills, M., Cohen, J. and Wong, Y. "A Magnifier Tool for Video Data," *SIGCHI '92: Proceedings of Human Factors in Computing Systems*, Monterey, CA, pp. 93-98, 1992.
11. Radhakrishnan, R., Xiong, Z., Divakaran, A. and Memon, N. "Time Series Analysis and Segmentation Using Eigenvectors for Mining Semantic Audio Label Sequences", *IEEE International Conference on Multimedia and Expo (ICME)*, June 2004.
12. Robertson, G.G. and Mackinlay, J.D. "The document lens," *UIST'93*, pp. 101-108, 1993.
13. Wittengburg, K., Forlines, C., Lanning, T., Esenther, A., Harada, S., Miyachi, T. "Rapid Serial Visual Presentation Techniques for Consumer Digital Video Devices", *Proceedings of UIST 2003*, pp. 115-124.
14. Zhai, S., Buxton, W. and Milgram, P. "The partial-occlusion effect: Utilizing semitransparency in 3D human-computer interaction," *ACM Transactions on Computer-Human Interaction*, 3(3), 254-284, 1996.
15. Zhang, H., Low, C.Y., Smoliar S.W. and Wu, J. H. "Video Parsing, Retrieval and Browsing: An Integrated and Content-Based Solution," *ACM Multimedia 95*, pp. 15-24, 1995.