VIDEO INTERFACE FOR SPATIOTEMPORAL INTERACTIONS BASED ON MULTI-DIMENSIONAL VIDEO COMPUTING

Akihito AKUTSU, Yoshinobu TONOMURA and Hiroshi HAMADA

NTT Human Interface Laboratories
1-1 Hikari-no-oka, Yokosuka-Shi, Kanagawa, 239 JAPAN
{acts, tonomura and hamada}@aether.hil.ntt.co.jp

ABSTRUCT

Because digital video is becoming increasingly important for the networked multimedia society, the audio-visual access environment should allow us to do more than just passively watch. We propose a new video user interface concept made possible by multi-dimensional video computing. Multi-dimensional video computing offers a framework for analyzing a video, creating new structures, and restyling and visualizing the video according to the user's demands. The video interface visualizes video content and context structure comprehensibly to allow us to access the spatiotemporal information in videos intuitively.

In this paper, we introduce our research activities toward a video interface based on the information extracted from the video. New video interfaces called VideoBrowser, PanoramaVideo, and VideoJigsaw are described.

1. INTRODUCTION

When we try to extract information from a video, we are forced to watch it in real time. We can fast forward or reverse it, but this provides only a very rough overview of the video contents. It is irritating to replay a sequential tape because of the time taken and the many unsuccessful attempts often needed to locate a specific segment. Therefore, a method for fast video browsing that enables the viewer to grasp the idea of a lengthy video without watching it in its entirety is one of the most desired functions. In addition to the above-mentioned problems, we are compelled to watch the video as set by the director and/or producer. To scan a specific video quickly, we need a more direct and comprehensive interface than fast forwarding and normal viewing. Consumers are now moving beyond the conventional video player. They are starting to use commercial digital video software to access attractive multimedia applications. This means that they are becoming familiar with the flexibility and power of computer interfaces.

This paper introduces a new user interface for video, a more direct and comprehensive visualization of video context, and access to video contents. It provides not only effective interactive video interfaces, but also a variety of user-selectable video viewing styles. A new user interface requires video computing which can accurately and rapidly process final videos. In video computing studies, the most important issues are 1)how to extract a video's inherent attributes as unconsciously perceived by people, such as scene changes and camera operations, and 2)how to transform/combine the attributes to realize new visual representations. The goal is to provide intuitive access to spatiotemporal video information.

The issues of video analysis, video structuring, visualization, and new video interfaces with intuitive access, are very important in stimulating new video applications, especially for accessing digital video libraries.

First we describe the framework and essential functions of multi-dimensional video computing in section 2. In Section 3, we describe the concept of the user interface with videos and clearly present the video features and attributes extracted for spatial-temporal visualization and video access. In Section 4, we propose and discuss the space-based video interface with visualization and video access.

2. MULTI-DIMENSIONAL VIDEO COMPUTING

Figure 1 shows the framework of multi-dimensional video computing. Physical features which represent video content and context are extracted for indexing and structuring by signal video processing techniques. Automatically extracted features allow a sequential video data to be manipulated as processable segmented pieces. Video content structure is constructed by analyzing and indexing the segmented video contents. Video context structure (link structure) is created by establishing new relationships between the segmented videos. A semantic structure is created by combining the created structures and knowledge held in databases. The video structures are visualized and shown to the user in response to his/her demands. Visualization provides us not only with several effective interactive video interfaces, but also a variety of user-selectable video viewing styles.

Recently, the concepts of video structuring and video interfaces have been proposed[1][2][3]. Ways of turning the concepts into video interfaces, such as video indexing, have been discussed[4]. These studies included a video content and context analysis tool for video indexing. Several researchers have investigated the visualization of the video attributes that represent content and context for the purpose of fast video browsing[5][6]. Recently, the focus of research has moved to the visualization of the spatiotemporal information held in a video. Spatiotemporal based video representations offer rich visual cues to the viewer[7][8].

The content and context structures appropriate for video interface construction depend on the user’s purpose and the kind of video. Video structuring is classified into two types, 1)structuring *sequential video data*, for example movies, dramas, news, and so on, and 2)structuring *lumps of raw video data* for example multi-

Figure 1. Multi-dimensional video computing

camera shots, unedited raw video data etc.

Sequential video data implicitly has a hierarchical structure that is created by the director during production. The structuring of this type of video data uses edit information that include cut points, music part, narrations area and so on. The video browser mentioned in 3.1 is appropriate for this type of video data. The video browser is aimed to provide not only effective video browsing, but also a variety of user-selectable video viewing styles according to meet the user's demands.

The potential structure of a *lump of raw video data* is formed by the linkage of segmented video clips. Clips are linked according to attributes, for example, several shots which are taken by multi-camera are linked by spatial relation, object information, background sound and so on. The interface that uses a content-relation based structure is aimed at intuitively understanding spatial information (for example object locus, size, absolute motion and so on) from the segmented video clips. In 3.2, we mention an interface with this type of video structure.

3. USER INTERFACE FOR VIDEO

3.1VideoBrowser

Videobrowser provides an interface to visualize content and context structure, and to allow us to grasp video contents intuitively[9]. Automatically extracted cut points are used to visualize the context structure. It is assumed that the cut points are features added by the director and are essential video units. The visualization is realized by showing key frames and/or 3D icons(2D image + 1D time) selected and/or created using cut points from each shot. To display the key frames and/or icons along a time line visualizes not only the content structure, but also the context structure. We have developed several video interfaces with browsing facilities. They are described below. *PaperVideo* is a paper-based new video interface in which video information is fixed on paper. The *PaperVideo* system prints on paper representative images of the video. These representative images are extracted by using the cut point information contained in the video. Because paper is convenient and easy to use, The *PaperVideo* allows us to grasp video contents and context easily[9].

We have also realized an effective access interface for a large volumes of video data held in digital libraries[10]. The interface can provide different viewing styles to match the user's purpose. The proposed video interface realizes hierarchical interaction where the viewing styles can range from course to fine. For creating the hierarchical structure, we use attribute information (on air time, production place etc.) and extracted information (cut points, music, voice points etc. in video).

Figure 2 visualizes our proposed video browser which has granularity for the interaction. We can see a large volumes of video data in overview in the coarsest layer. In the fine layer, we can watch the video contents directly on the monitor.

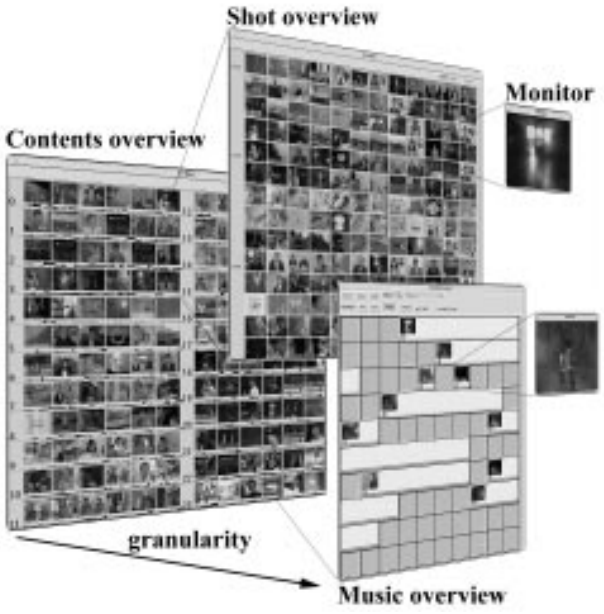

Figure 2. VideoBrowser

3.2Video Interface with video shot spatial synthesis

The space-based video interface supports our understanding of the spatial information among segmented video clips intuitively. For this interface, multiple shots are synthesized spatially. The interface allows us to grasp an object's locus, size, absolute motion, trajectory etc. as contained in the multiple shots.

The following requirements should be realized in the space-based video interface to access spatiotemporal information intuitively.

1) reconstruction of wide video scene space from a video.
2) visualization of object motion and trajectory in scene space.
3) random access to spatiotemporal information by using spatial index.

In 1), we reconstruct a wide video space over a frame, like a panoramic space. For creating an even wider video space, we synthesize some wide video spaces spatially. In 2), for visualizing object motion and trajectory, the extracted object is fixed in a wide reconstructed space. The above visualizations are used as an index for realizing direct random access to

spatiotemporal information. The video interface with 1), 2) and 3) allows us to grasp spatial-temporal video information spatially, and to get the information directly.
In this paper, the video scene space is represented as a panoramic image by using automatic extracted camera operations[11][12] which include the information of the spatial relationship between frames in a shot. We create a stroboscopic panoramic image in which an automatic extracted moving object is fixed. We spatially synthesize panoramic images by using the spatial relationship among each segmented shots.
The space-based video interface with these panoramic representations permits random access to spatiotemporal information.

4. SPACE-BASED VISUALIZATION AND ACCESS OF VIDEO CONTENTS

We collected several video sequences which were captured by a video camcorder fixed on a tripod. The video sequences held information of a common space, but the temporal information differed because the videos were captured at different times. We realized a space-based video interface by video spatial synthesis.

4.1Space-based visualization

4.1.1Space reconstruction

The panoramic scene space is created by changing the image position and size of each frame according to the camera operations[12]. We call this spatial synthesized scene space PanoramaVideo. A normal panoramic photography has only 2 dimensions, but the PanoramaVideo offers 3 dimensional data.
Figure 3 shows the PanoramaVideo produced from shots captured with pan and tilt camera operations. This visualization allows us to understand spatial information intuitively without replaying the video. Next, we created a spatial synthesized scene space which was wider than PanoramaVideo by using camera operations and the spatial relationship among shots.
The spatial synthesizing concept is called VideoJigsaw. The wide PanoramaVideo is the one of embodiments of the VideoJigsaw.
If we watch each shot individually, we can not understand the spatial relationship of the objects captured in each shot. VideoJigsaw allows us to grasp the spatial object structure in terms of position and size etc.

Figure 3. PanoramaVideo

4.1.2Object trajectory representation

We fix the extracted objects in a reconstructed scene space to represent motion object. We call this Stroboscopic PanoramaVideo. Stroboscopic PanoramaVideo has two types of representation as shown in Figure 4, 1)fixed at spatially regular intervals, 2)fixed at temporally regular intervals. Stroboscopic PanoramaVideo allows us to understand not only the object's absolute motion but also changes in the object's form.
By using object information and the relationship between shots, we can overlap the Stroboscopic PanoramaVideo. The overlapped PanoramaVideo is also a form of VideoJigsaw. Figure 5 shows the overlapped Stroboscopic PanoramaVideo. We used two shots of different ski jumpers. This representation allows us to understand the spatiotemporal difference between the two jumper's forms intuitively. Watching the two shots on different monitors is not as effective.

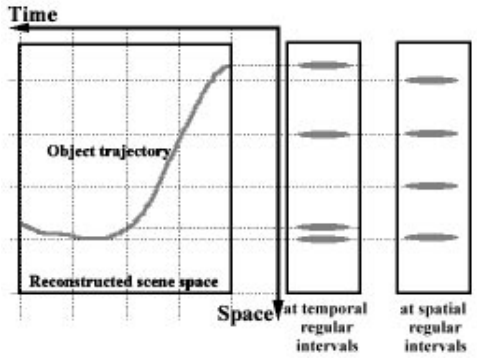

Figure 4. Types of stroboscopic representation

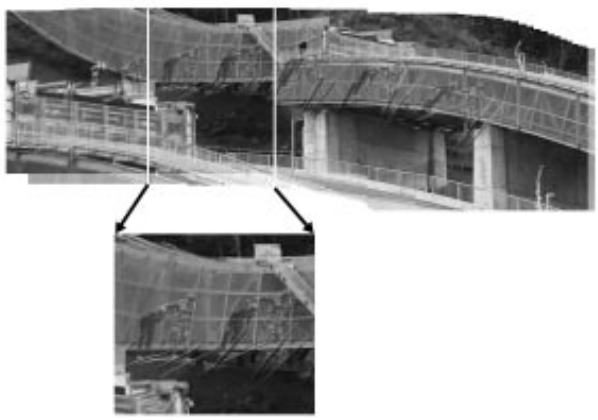

Figure 5. Overlapped stroboscopic PanoramaVideo

4.2Space-based interaction

In this paper, we propose a space-based interface with direct access to temporal video contents. The proposed interface uses a panoramic representation to provide access cues without using time code. The video access mainly consists of two types of operations: 1)selecting the access point and 2)manipulating(play, stop and so on) the video. The space-based visualization that is mentioned in 5.1 is effective for selecting the manipulation points on the video directly. If you click on a specific part of the panoramic representation, the video will be replayed from the point selected. Spatially accessing the panoramic view equals temporally accessing the video sequence.
In the real world, we try to understand some events by trying

various viewing styles. We assume that there are three stages in the viewing styles: overview, place-based point-view, and motion-based point-view. In the overview stage, we can overview the whole space. In the point-view stage, we watch and concentrate on the events happening. We use the place-based point-view, when we want to pay attention to a specific place and observe what is happening. When we focus on the moving object, we utilize the motion-based point-view. We get information from the real world while going back and forth between these stages.

Figure 6 shows our implementations of the space-based play back modes corresponding to the above viewing styles: 1)panoramic playback, 2)place-based playback and 3)motion-based playback. The panoramic playback mode can replay a shot using the panoramic scene space. We can watch the moving objects as if we were there. If we want to watch some part of the scene space in which some events are happening, the place-based playback mode supports us in accessing that spatial part directly. Playback is performed only for the place of interest. The motion-based playback mode replays the momentary object motion selected by the user by clicking on the fixed object in the scene space. When we want to see the temporal events in the shot, this mode is effective to understand occurrences.

The proposed video interface offers both spatial and temporal manipulation. If you touch some part of the panoramic spot and move it, the video will be replayed at the speed matching your movement. This manipulation provides us seamless and/or intuitive manipulation (playback, stop, forward, rewind etc.).

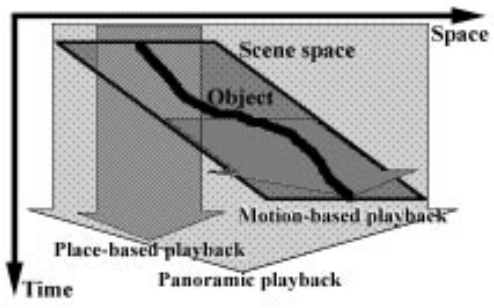

Figure 6. Space-based interaction modes

We usually interact with a video to get spatiotemporal information for various purposes, for example, editing, retrieving, entertainment and so on. The best video interface should have various levels of granularity to support our interaction style and intentions. When we watch TV, video movies etc., we are satisfied with rough granularity(for example title, action scene, music scene etc.) in terms of interaction. The proposed VideoBrowser provides many more levels of interaction granularity. Video creators and/or editors, sports coaches and/or commentators etc. requires fine grain interactions. To edit videos, individual frames should be handled, for example “In point” and “Out point” selection. To analyze captured motion, larger frame groups are needed. PanoramaVideo and VideoJigsaw allow us to interact at the shot and/or frame level which is fine grain interaction. This allows them to catch and handle the important action points from the shots effectively and intuitively.

5. CONCLUSION

We proposed a new video user interface concept made possible by multi-dimensional video computing. Multi-dimensional video computing offers a framework for analyzing a video, creating new structures, and restyling and visualizing the video according to the user's demands. The video interface visualizes the video’s contents and context structure comprehensibly to allow us to access spatiotemporal information in the video or group of videos intuitively. In this paper, we introduced our research activities on a video interface based on the information extracted from the video. New video interfaces called VideoBrowser, PanoramaVideo, and VideoJigsaw were described.

Acknowledgments

We are grateful to Dr. Yukio Tokunaga, Executive Manager of the Advanced Video Processing Laboratory, NTT Human Interface Laboratories, for his encouragement during this research. We would like to thank our colleagues for their helpful advice and discussions.

References

[1]Y. Tonomura, “Video Handling Based on Structured Information For Hypermedia Systems”, Proceedings of ACM Int’1 Conference on Multimedia Information Systems, pp.333-344, 1991.

[2]H. Ueda, T. Miyatake and S. Yoshizawa, “IMPACT: An Interactive Natural-Motion-Picture Dedicated Multimedia Authoring System”, Proceedings of CHI'91, pp.343-350, 1991.

[3]G. Davenport, S. Aguierre and N. Pincever, “Cinematic Primitives for Multimedia”, IEEE CG&A, vol.11, no.4, pp. 67-74(July 1991).

[4]Y. Tonomura, A. Akutsu, K. Otsuji and T. Sadakata, “VideoMAP and VideoSpaceIcon: Tools for Anatomizing Video Content”, INTERCHI 93 Conference Proceedings, pp.131-138, 1993.

[5]F. Arman, R. Depommier, A. Hsu and M-Y. Chiu, “Content-based Browsing of Video Sequences”, Proceedings of ACM Multimedia 94, pp.97-103, 1994.

[6]P. Aigrain, P. Joly and P. Lepain, “Representation-based user interfaces for the audiovisual library of year 2000”, SPIE Vol.2417, pp.35-45, 1995.

[7]R. Irani and S. Peleg, ”Motion Analysis for image enhancement: Resolution, Occlusion and Transparency”, Journal of Visual Communication and Image Representation, vol.4, no.4, pp.324-335, 1993.

[8]L. Teodosio and W. Bender, “Salient Video Stills: Content and Context Preserved”, ACM Multimedia'93 Conference Proceedings, pp.39-46, 1993.

[9]Y. Tonomura, A. Akutsu, Y. taniguchi and G. Suzuki, ”Structured Video Computing”, IEEE Multimedia, Vol.1, No.3, pp.34-43(1994).

[10]Y. Taniguchi, A. Akutsu, Y. Tonomura and H. Hamada, ”An intuitive and efficient access interface to real-time incoming video based on automatic indexing”, Proceedings of ACM Multimedia 95, pp.25-33, 1995.

[11]A. Akutsu, Y. Tonomura, “Video Tomography: An efficient method for Camerawork Extraction and Motion Analysis”, Proceedings of ACM Multimedia 94, pp.349-3566, 1994.

[12]A. Akutsu, Y. Tonomura and H. Hamada, ”VideoStyler: Multi-dimensional video computing for eloquent media interface”, Proceedings of ICIP 95, Vol.1, No.1, pp.330-333, 1995.