VideoMAP and VideoSpaceIcon: Tools for Anatomizing Video Content

Yoshinobu TONOMURA, Akihito AKUTSU, Kiyotaka OTSUJI, Toru SADAKATA

NTT Human Interface Laboratories

1-2356 Take, Yokosuka, Kanagawa, 235 Japan

phone:+81-468-59-3112 email: tonomura@nttvdt.ntt.jp

ABSTRACT

A new approach to interacting with stored video is proposed. The approach utilizes VideoMAP and VideoSpaceIcon. VideoMAP is the interface that shows the essential video features in an easy to perceive manner. VideoSpaceIcon represents the temporal and spatial characteristics of a video shot as an intuitive icon. A video indexing method supports both tools. These tools allow the user's creativity to directly interact with the essential features of each video by offering spatial and temporal clues. This paper introduces the basic concept and describes prototype versions of the tools as implemented in a video handling system. VideoMAP and VideoSpaceIcon are effective for video handling functions such as video content analysis, video editing, and various video applications which need an intuitive visual interface.

KEYWORDS: Video Handling, Visual Interface, Icon, Index, Image Processing, Visualization

INTRODUCTION

Since the introduction of video display boards for computers, many video applications have been attempted. In an early work, Hodges, et al. developed multimedia learning environments using one of the earliest video workstations [1]. Multimedia synchronized editing using the time line was introduced. In the last three years, thanks to advances in the standardization of video compression algorithms and digital signal processing, several multimedia platforms capable of storing, accessing, and displaying video have been developed. The first commercial versions of multimedia operating systems are now becoming available. Various application systems such as desktop video editors and electronic video libraries have been developed by many vendors.

Video Handling Issues and Previous Researches

Although computers strongly support traditional application systems, they cannot yet handle video as efficiently as text. This is because computers do not "speak" the video language. Text handling functions such as automatic full text searches, keyword generation, and structured editing, are powerful and well match the characteristics of texts. On the other hand, video content is extremely vague and difficult to specify. Therefore, most video application systems rely on humans to input the necessary data. To eliminate this dependency on humans, many research issues must be defined and resolved. The key issues are creating intuitive visual clues to help users perform their tasks efficiently, developing a structure of video data management, and utilizing image processing to automatically extract the representation of a video.

A visual interface is essential to activate the user's visual sense and stimulate the user's intuition especially when manipulating video. Brondmo and Davenport introduced Micon(moving icon) to represent the video content of hypermedia journals [2]. One of us (Tonomura) has already proposed a content oriented visual interface using video icons and other intuitive visual interfaces for video handling [3]. The video icon is based on a structured icon model that has a shadow corresponding to the video segment footage. Mills, et al. proposed a magnifier tool for video data that offers a range of views, from wide to close, of video data [4]. The tool is simple but effective in supporting video manipulation. The research to date has tried to develop effective visual interfaces but the clues they offer the user are rather limited.

Video data management is important for establishing flexible systems. The management should resolve the two main questions: what kind of information should be used to represent the video content and how the information should be handled. Davenport, et al. proposed a framework that used layered information management [5]. The important idea of "granularity of meaning" was noted:the degree of information coarseness needed for efficient multimedia handling. MacNeil developed a visual programming environment which uses a case-based reasoning approach to support multimedia designers[6]. A vertical slice of each video frame image is simply shown over time for users to see the visual rhythm of the shots in video.

The development of a system supported by image processing is important because the clues must be extracted automatically if we are dealing with a large video database. One of us (Tonomura) proposed a video handling architecture that used, as the basic clues, the video indexes created automatically by image analysis [7]. The automatic detection of video cuts, one of the video indexes, was realized by analyzing the intensity histogram data. Ueda, et al. reported an editing support system that used image analysis to achieve

cut detection and image flow analysis, in which image processing was performed only by software [8]. In these more recent works, however, the number and styles of clues to video content are still limited compared to the rich information present within videos.

Our Focus

Our research focus is to establish universal clues, that would be useful enough to handle a video in many different ways. We were also interested in creating and testing promising new video tools that would lead to enhanced video applications. Our research results were realized in our prototype video handling system.

This paper:

- •Discusses a video information reference model to clarify the origin of the information contained with a video.
- •Describes the basic concept of our video handling system to explain how the video indexes work.
- •Proposes VideoMAP which displays useful video features as a tool for video indexing.
- •Describes VideoSpaceIcon which represents temporal and spatial video features as icons.

VIDEO INFORMATION REFERENCE MODEL

As shown in Fig. 1, a lot of information about a video is related to its creation and use. The information includes filming parameters, how it is stored, and how it was edited. The left hand side of Fig. 1 lists the physical information and the right hand side lists how people interact with the video through creation and editing. Our intention is to extract as much information as possible from the video itself. Furthermore, a long term goal is to estimate the intention of the people associated with the video production.

The most basic data in our video handling system are cut points and camera operations. A cut point is a seam in the video sequence generated by camera stop and start or subsequent editing. Camera operations include panning, tilting, and zooming. The director's intent is reflected either implicitly or explicitly in the cut points and camera operations. Editing effects are such as fades, wipes. Telecine conversion is to convert photographic films into videos. By analyzing the video, such information can be extracted and described in a structured manner.

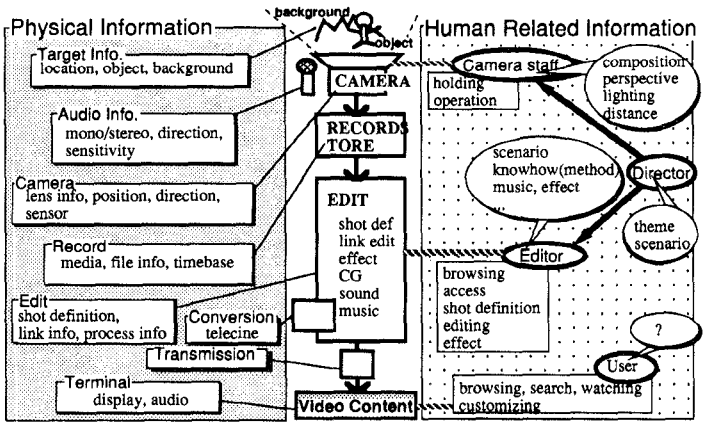


Fig. 1 Video Information Reference Model

FROM ANALYSIS TO APPLICATION: VIDEO INDEXED BASED VIDEO HANDLING

The processing structure of our video handling system is shown in Fig. 2. The attributes of the video data are located lower while applications fall in the top layer. Video data are expressed in HVC (Hue, Value, and Chroma) because this mirrors human image perception. Value means intensity.

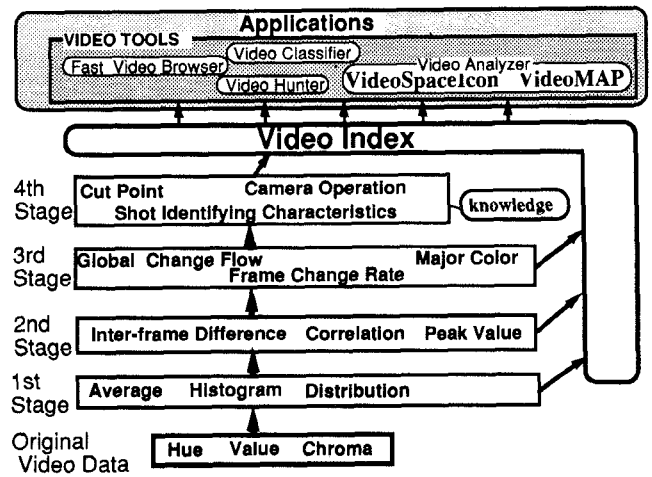


Fig. 2 Processing Structure of Video Handling

The first stage process is image processing for general feature extraction on each frame in the video sequence. For example, the average intensity, intensity distribution, and intensity histogram are calculated from individual pixel values. This stage outputs a new time-based data stream. The second stage filters the output of the first stage to extract more specific features; for example, inter-frame difference, correlation, representative peak value. As a result of the filtering, new time-based data streams are generated. The third stage processes the data streams to generate the basic information needed for characterizing the video. For example, the frame change rate is defined by the inter-frame difference of intensity, which means how rapidly the images change. The global change flows are obtained by analyzing the images of time-based data stream. The details are explained in later sections. The video indexes are viewed from the aspect of movies in the fourth stage. Cut points, camera operations, and other shot-identifying characteristics such as shot color are obtained. The typical usage of cut points is to define the unit of video access and storage. We have established an effective cut detection method using intensity data [9]. In the method, cut detection is performed by thresholding the frame change rate. Camera operation detection is realized by analyzing the global change flows. The processes in this stage can be supported by knowledge of the mechanics of video creation. For example, telecine conversion is detected observing a periodic frame change rate because we know the characteristics of the conversion. We treat the features generated in the first to fourth stages as video indexes.

The application stage offers several very effective video tools such as a fast video browser which shows a long video within a short time, a video hunter which retrieves a specific video, a video classifier, and a video analyzer which analyzes video at various levels of granularity. Real applications are being built using the tools. VideoMAP and VideoSpaceIcon,

described later, are prototype video tools of the video analyzer.

Generated in Off Line, Used in Real Time

One of the advantages of using video indexes is that we don't need very fast image processing computers to use the application. Once a video is indexed, the indexes are attached to the video data and stored in a data base. The application can be run on a relatively slow computer because time consuming image processing is not needed. It is possible that an industrial index maker could generate the video indexes with a very fast task-specific computer.

Granularity of Handling

By using the video indexes generated in different stages, granular video handling is possible. The more abstract the video index is, the coarser the expression of the video content becomes. A multi-layered video index management scheme that will permit access to the video indexes of any stage is needed in order to realize the smooth management of granularity.

VIDEOMAP

External Representation of Video Index Features

VideoMAP is the interface that expresses essential video features as video indexes. It also provides direct access to specific parts of the video. In other words, VideoMAP is the visual representation of the video indexes held in the computer. The indexes are displayed on a time line. The method of feature visualization depends on the feature's characteristics. By pointing at some part of the feature pattern, the corresponding video frame appears in the video window. One of the advantages of this interface is that users quickly develop a concept of the behavior of each feature on the time line and also the correlation between features.

Example: 20 Seconds at a Glance: Six Features

A typical VideoMAP layout is shown in Fig. 3. The time line runs from left to right. In this case, there are 600 frames so the total time line represents 20 seconds. The intensity histogram, intensity average, inter-frame difference of intensity histogram, video X-ray image which is explained later, and hue histogram follow from top to bottom. Features shown in VideoMAP can be selected from the index list to match the user's requirements. The order of the feature rows can also be changed by the user. (See also Tonomura, Color Plate 1.)

The cut points are clearly shown in each feature as pattern discontinuities. In the plot of histogram intensity, the intensity values of all pixels in each frame are quantized into 16 levels. Each quantized frequency value is represented with a gray scale where white indicates high frequency. Intensity average in the second row is calculated from the intensity histogram. The plot of inter-frame difference of intensity histogram (third row) shows the big peaks which clearly indicate the cut points. The image flow patterns (fourth and fifth rows) is obtained by filtering the edge images of spatio-temporal video frames. The number of edge image pixels is summed on vertical and horizontal axes. The result of filtering is a kind of projection from the top and the side. We call this the **video X-ray.** White pixels indicate the existence of a large number of edge image pixels on that axis. The hue histogram (bottom row) has a resolution of 256 hues: from the top, purplish red -> blue -> green -> yellow green -> green -> orange -> red. Cut points are clearly discernable. One or more apparent bars are seen in several shots. The bar in the center indicates green while the upper bar indicates blue. Since the target video is a scene from a baseball game, the green corresponds to the turf and the blue the fence. The players' uniforms are white but white is not directly represented as a hue. White does, however, directly influence intensity. The images on the top are the first frame

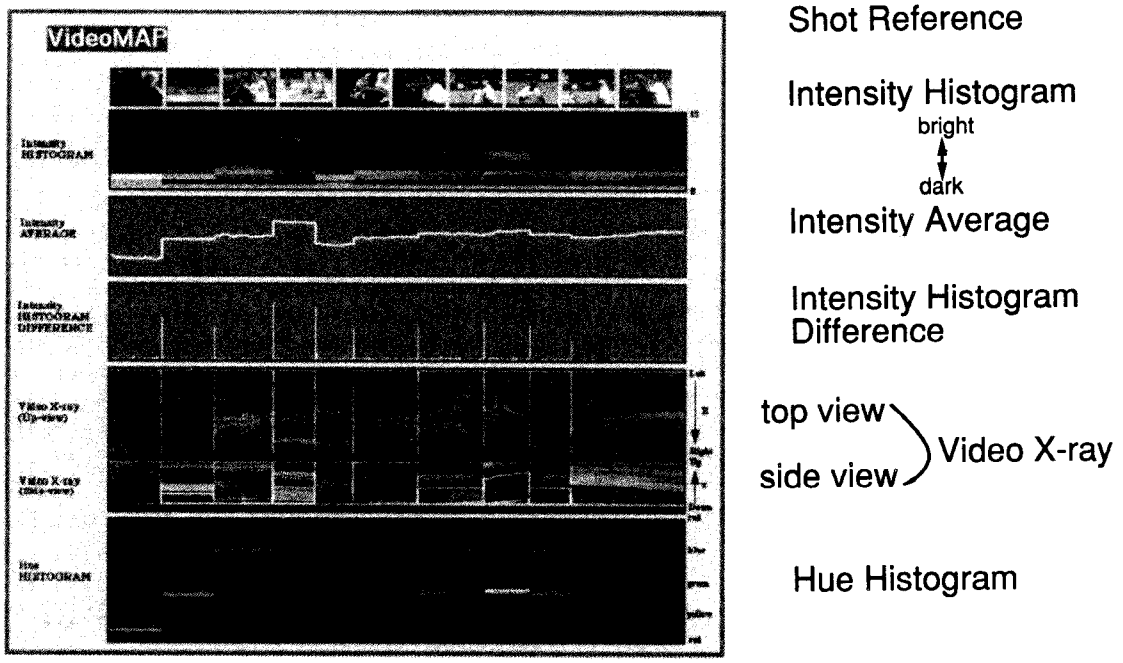


Fig. 3 VideoMAP (600 frames, 20 seconds)

of each shot. By comparing the image with each of the cut's patterns we can readily discern pattern origin.

Video Sign: Identification Information of Shot

The intensity histogram and image flow contain interesting patterns that include important information. We call this kind of pattern **Video Sign** and use it for identification. Video Sign is equivalent to the voice print of the audio field. The exact form of Video Sign depends on the content, but it is possible to perceive general patterns. For example, patterns in the intensity histogram can indicate the presence of editing effects commonly applied at the beginning or the end of a shot. Note that the example shown in Fig. 3 does not contain such patterns. Camera operations and object motion within a scene can be discerned from the video X-ray. For example, the first shot in the top view of the video X-ray (fourth row in Fig. 3) contains slanting lines which show that camera was panned. The last shot contains lines that diverge. This is a typical zooming pattern. Detailed analysis of these patterns led us to develop the VideoSpaceIcon. With further research we could employ pattern matching to identify scenes or shots. We are still testing VideoMAP to determine the most effective features that will fully engage the power of human intuition.

VIDEOSPACEICON

VideoSpaceIcon is a tool that displays spatial and temporal video features as icons. It utilizes the camera operation parameters that can be automatically detected from the video X-ray of VideoMAP.

Video Space: Space Oriented Video Content View

An example of a VideoSpaceIcon is shown in Fig. 4. It is an extended (3-D) rectangle not the conventional square 2-D image icon. The original video shot showed a girl appearing on the left through a door, walking to the right and sitting on a chair. The camera was panned from left to right following the girl. Conventional icons or video screen can only show actual frames, but a VideoSpaceIcon can show the entire physical space. We call this the **video space**. The conventional video display process employs only a time oriented approach, but VideoSpaceIcon is space oriented; it allows us to grasp the physical space of the shot at a glance. The shape of the VideoSpaceIcon is drawn in Fig. 6. Our system is able to display views in different angles other than front. The three dimensional figure of the icon allows us to grasp how the camera was operated at a glance. The top and side view can be replaced by video X-ray images in X-ray mode. Fig. 5 shows the VideoSpaceIcon in that mode: modified video X-ray images, top and side view, are displayed with a front view. The straight lines parallel to the time line in Fig. 5 are created from the background and objects that remain stationary over a long period of time. The top and side views are useful to grasp object movement because movement is indicated by non-parallel lines which are clearly visible.

How VideoSpaceIcon is Created

The VideoSpaceIcon is created by changing the image position and size for each frame according to the camera operation parameters. Camera operation information is

Fig. 4 VideoSpaceIcon: Example 1 (front view: panning shot)

Fig. 5 VideoSpaceIcon (temporal mode of Fig. 4)

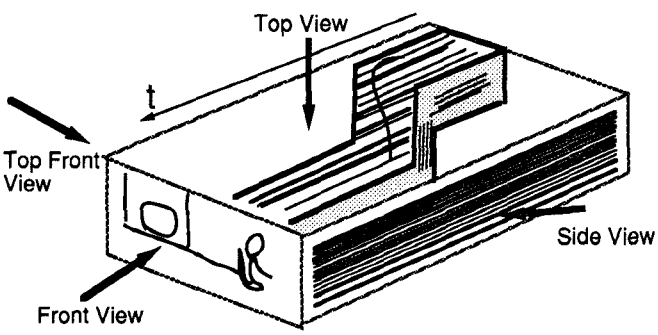


Fig. 6 Views of VideoSpaceIcon

available beforehand only in specific applications and demands the use of a special camera. However, what we are interested in is creating video spaces for video shots for which the camera operation is unknown.

What the camera captures in a frame has more or less global changes. For example, when the camera is panned to the right, background and static objects move to the left. This movement can be tracked with motion vectors. By analyzing the flow generated by the motion vectors, camera operation parameters can be determined. Our former research [10] and Ueda's IMPACT [8] are based on motion vector analysis. Unfortunately, it takes too long to compute the results and robustness is suspect in many situations. Our newly proposed method uses the video X-ray images, is rapid, and is robust enough to analyze most videos. When the camera is fixed, the video X-ray images contain many lines parallel to the time line that were generated from the background. When the camera is panned, the lines are slanted; the degree of slanting depends on the panning speed. When the camera is zoomed, the lines diverge. A moving object generates non-parallel lines. In our experiments, as long as the background contains some distinctive features, global flows are clear enough to calculate camera operation parameters.

The degree of slanting can be converted into the three major camera operation parameters: panning, tilting, and zooming. Tilting is vertical panning. Panning and tilting vary the camera's optical axis. Zooming changes the camera's focal

length. Other operations include lateral camera movements such as tracking, and dollying. These basic camera operations are shown in Fig. 7. In our current system, panning, tilting, zooming, and any combination of these operations can be represented by VideoSpaceIcon. The other operations are not currently supported because reconstructing the video space would be much more complex.

Constructing a VideoSpaceIcon from camera operation parameters is shown in Fig. 8. The icon image position shifts to indicate panning and tilting, while the image size is changed to express zooming. Zooming has little effect on the video space itself, but the zoomed part has higher resolution than usual. Another example of VideoSpaceIcon is Fig. 9. Its form mirrors the camera operations: tilt down, pan to right, fix, and pan to left. (See also Tonomura, Color Plate 2.)

Structured Icon Model

An icon is a representation of one or more features of something and permits interaction with the something. By clicking on the video space, the correspoinding frame image is accessed and displayed. While the actual creation of a VideoSpaceIcon is performed as is explained above, logical icon handling for visualization and interaction in software is achieved by using the structured icon model. We have proposed the structured video icon model and the simple video icon that has shadow [4]. VideoSpaceIcon is an extension of this technique. The structure model of the VideoSpaceIcon is shown in Fig. 10.

Moving VideoSpaceIcon: Reproducing Object Motion in Video Space

One interesting aspect of the VideoSpaceIcon is that object movement can be reproduced in a very simple manner. By overlapping sequential image frames in the video space, a moving object actually moves. Fig. 11 shows this idea using Fig. 4. In this example, the girl's image really walks. This is a new effect that is not possible with traditional video. We can see the entire space in which the shot was actually taken. When creating a video from a cinemascope movie, the original movie is often trimmed and some of the picture is lost. Under some constraints, it is possible to reconstruct the

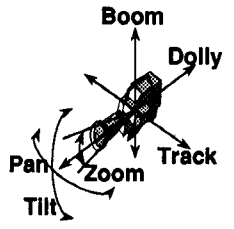


Fig. 7 Basic Camera Operations

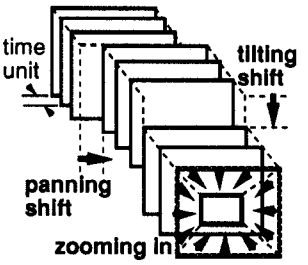


Fig. 8 Construction of VideoSpaceIcon

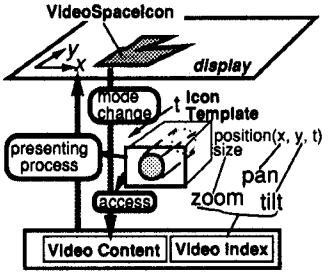


Fig. 10 Icon Structure Model

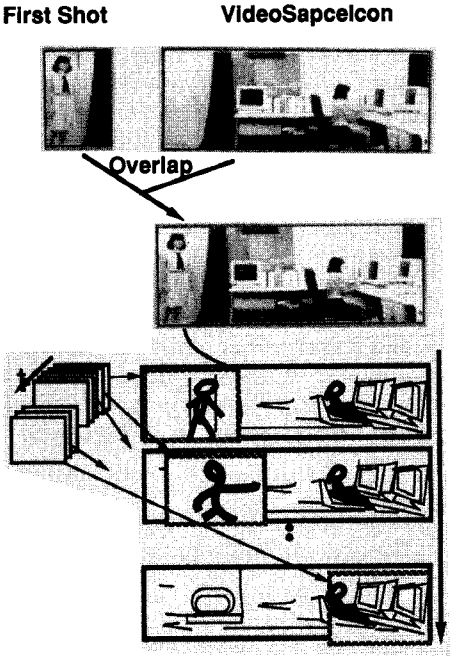
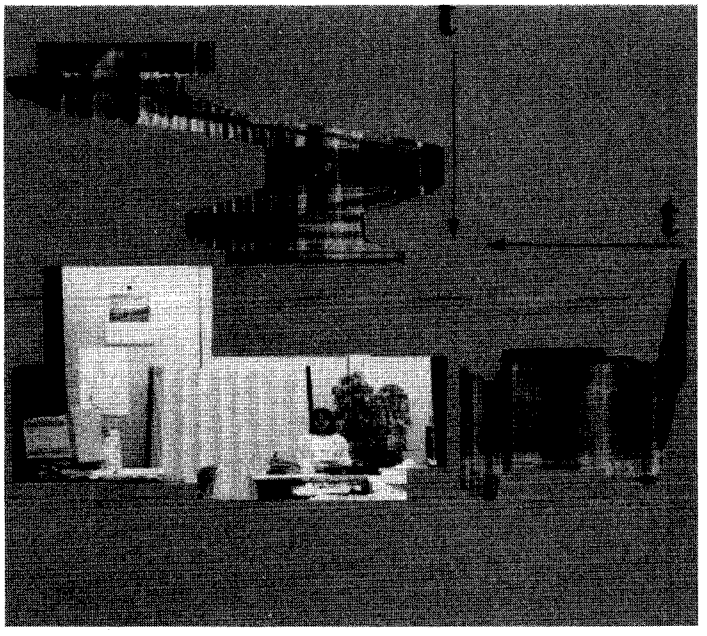
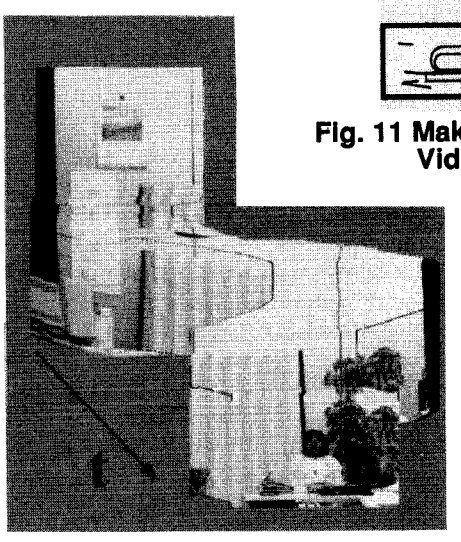


Fig. 11 Making of Moving VideoSpaceIcon



(a) X-ray mode



(b) normal mode: top front view)

Fig. 9 VideoSpaceIcon: Example2 (panning, tilting)

original image with our method. Also, by windowing the video space created, a video with a different view from the original one can be obtained.

Video Space Monitor: New Display Possible

If the size of the video space is increased, it is no longer just an icon, but it is also a display screen on which video content can be viewed. Two types of display mode are possible: space oriented and anti-vibration.
In the space oriented mode, the video space is simply enlarged. In this mode camera operations are not discerned. In the anti-vibration mode, the video space shows only those frames for which rapidly changing camera operation is detected. In this mode, normal camera operations are observed. This is what we call the video space monitor; it is different from the conventional monitor display in that its screen shape dynamically changes and does not remain square. However, this idea is not fully implemented because of the problems of lens skew. Given some reasonable constraints, however, it is possible.

CONCLUSION

VideoMAP and VideoSpaceIcon were proposed as new tools for interacting with videos. VideoMAP realizes a direct visual interface with intuitive video features. VideoSpaceIcon fully utilizes the power of the video space oriented icon approach and suggests the possibility of many new video handling techniques. Both tools are visual representations of the video indexes that can be automatically extracted by video analysis. We believe that video-indexing-based video handling is important and useful for future video applications.

We are currently testing VideoMAP to optimize existing video features as video indexes as well as to uncover new features. VideoSpaceIcon is now being used to analyze camera usage in movie making. These tools were developed and implemented as a prototype interface for research purposes.

Problems and Future Study

Finally, we must point out several issues regarding our research.

First, we need more sophisticated video features that can be used as video indexes. We currently have about a dozen features based on intensity and hue data. For future applications, we need to consider how movie and video experts deal with video to discover more features.

Second, the robustness of the algorithms used to detect the features must be increased to allow all videos to be processed. We have successfully performed cut detection and camera operation extraction on several videos, including some that were over 2 hours in length. We now know what algorithm parameters are appropriate for what conditions. More experience is needed to confirm that we have a sufficient number of parameters.

Also needed is an enhanced visualization style for the video indexes. The interface shown in this paper is a prototype system and its design and functions may not satisfy the general user. The visual interfaces of these tools should be refined considering human factors. This is our next step.

VideoMAP does not use color to enhance the features to prevent sensory overload and subsequent confusion. An argument has been made that the limited use of color is warranted and we will try to confirm this.

The VideoSpaceIcon described above is not always effective because of the problem of camera lens skew. In one sense, this is not a problem when the icon is small and used only as an icon. To use it as a video space monitor, however, this problem should be resolved. In such conditions, anti-skew transformation is needed. The parameters needed for the process could be obtained from the video X-rays.

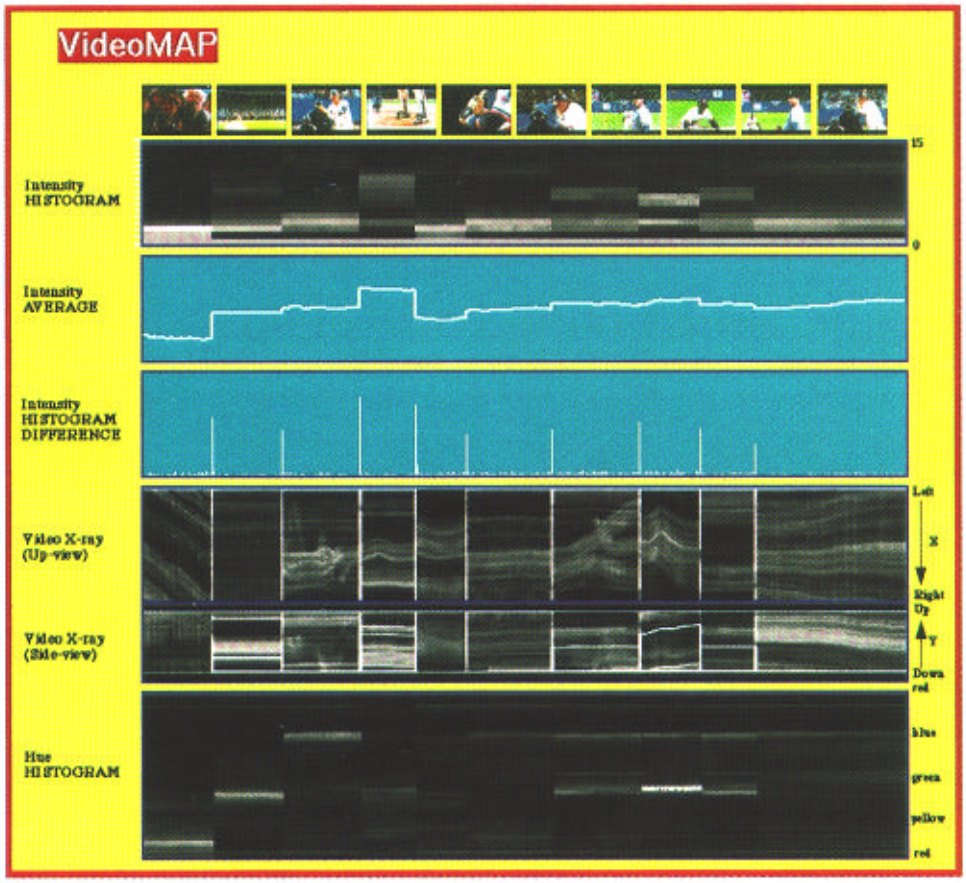
Lastly, when the camera operation is tracking or dollying, in which the camera position changes, the above method of creating VideoSpaceIcon can not be simply applied. A three dimensional space model is needed. This is for further study.

ACKNOWLEDGEMENTS

The authors are grateful to Mr. Tomio Kishimoto, Executive Manager of Visual Media Laboratory, NTT Human Interface Laboratories, for his encouragement of this research. We would like to thank Susumu Ichinose for his helpful advice.

REFERENCES

1. Hodges, M. Sasnett, R. and Ackerman, M. A Construction Set for Multimedia Applications. IEEE Software, 6, 4, (January 1989), 37-43.
2. Brondmo, H.P. Davenport, G. Creating and Viewing the Elastic Charles-a Hypermedia Journal, in Hypertext, State of the Art. R. McAlesse and C. Greene, eds., Intellect Ltd., Oxford, England, 1990.
3. Tonomura, Y. Abe, S. Content Oriented Visual Interface using Video Icons For Visual Database Systems. JVLC, 1, 2, Academic Press, 1990, 183-198.
4. Mills, A Magnifier Tool for Video Data. in Proc. CHI'92, (Monterey May 1992), 93-98.
5. Davenport, G. Aguierre, S. Pincever, N. Cinematic Primitives for Multimedia. IEEE CG&A, 11, 4, (July 1991), 67-74.
6. MacNeil R. Generating Multimedia Presentation Automatically using TYRO, the Constraint, Case-Based Designer's Apprentice. in Proc. IEEE Workshop on Visual Languages, (Kobe Oct. 1991), 74-79.
7 Tonomura, Y. Video Handling Based on Structured Information For Hypermedia Systems. in Proc. ACM Int'l Conf. on Multimedia Information Systems, (Singapore Jan. 1991), 333-344.
8 Ueda, H. Miyatake, T. Yoshizawa, S. IMPACT:An Interactive Natural-Motion-Picture Dedicated Multimedia Authoring System. in Proc. CHI'91, 343-350.
9. Otsuji, K. Tonomura, Y. Ohba, Y. Video browsing using brightness data. in Proc. SPIE VCIP'91, 1606, (Boston Nov. 1991), 980-989.
10. Akutsu, A. Tonomura, Y. Hashimoto, H. Ohba, Y. Video Indexing using motion vectors. in Proc. SPIE VCIP'92, (Boston Nov. 1992).
11. Ripley, G.D. DVI - A Digital Multimedia Technology. CACM, 32, 7, (July 1989), 811-822.



Tonomura, Plate 1 VideoMAP

top-front view front view

Tonomura, Plate 2 VideoSpaceIcon