

WATCH · GRAB · ARRANGE · SEE

Thinking with Motion Images
via
Streams and Collages

by

Edward Lee Elliott

B. A. Computer Science
University of California
Berkeley, California
1984

Submitted to the Media Arts and Sciences Section,
School of Architecture and Planning,
in Partial Fulfillment of the Requirements for the degree of

MASTER OF SCIENCE IN VISUAL STUDIES

at the

Massachusetts Institute of Technology

February 1993

© Massachusetts Institute of Technology, 1993
All Rights Reserved

Signature of Author _____

Edward Lee Elliott
Media Arts and Science Section
January 15, 1993

Certified by _____

Glorianna Davenport
Assistant Professor of Media Technology
Thesis Supervisor

Accepted by _____

Stephen A. Benton
Chairman
Departmental Committee on Graduate Students

MASSACHUSETTS INSTITUTE
OF TECHNOLOGY

Watch

MAR 11 1993

LIBRARIES

Watch · Grab · Arrange · See Thinking with Motion Images via Streams and Collages

by

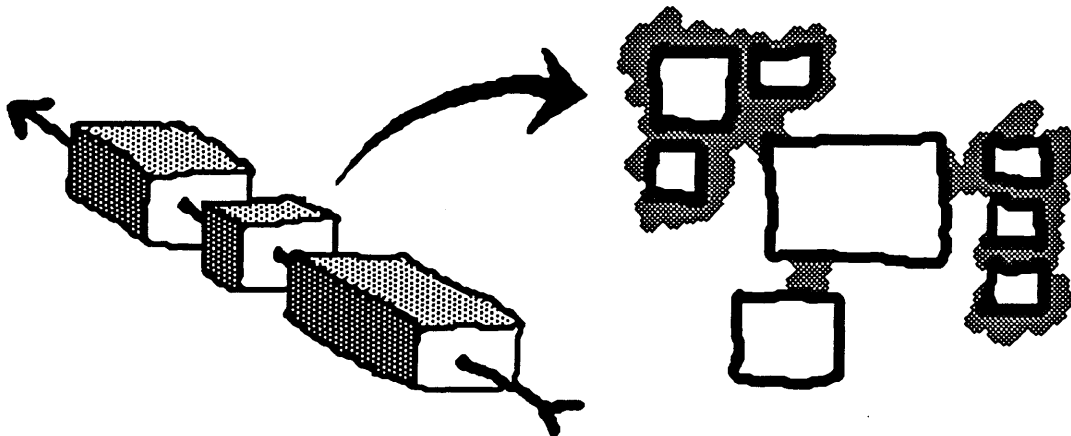
Edward Lee Elliott

Submitted to the Media Arts and Sciences Section,
School of Architecture and Planning,
on January 15, 1993,
in partial fulfillment of the requirements for the degree of
Master of Science in Visual Studies
at the Massachusetts Institute of Technology

ABSTRACT

Filmmakers experience a creative reverie seldom enjoyed by novices. That reverie comes as one pieces together thoughts embodied in motion images. This thesis borrows the manipulation of motion images from editing for the purposes of viewing. It suggests a collection of tools for grabbing elements from video streams and for manipulating them as a way of critical viewing.

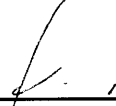
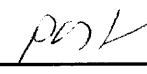
Motion images are usually conveyed sequentially. The tools suggested here allow a viewer to transfer from sequential image streams to collages of parallel images. The *video streamer* presents motion picture time as a three dimensional block of images flowing away from us in distance and in time. The streamer's rendering reveals a number of temporal aspects of a video stream. The accompanying *shot parser* automatically segments any given video stream into separate shots, as the streamer flows. The *collage* provides an environment for arranging clips plucked from a sequential stream as associations of parallel elements. This process of arranging motion images is presented as an engaging viewing activity. The focus is on viewing utensils, but these tools provide an alternative perspective to video elements that also has bearing on editing.



Thesis Supervisor:
Glorianna Davenport
Assistant Professor of Media Technology

This work was supported in part by Bellcore and Nintendo.

Thesis Advisor:

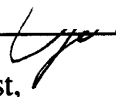
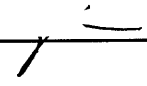
 

Glorianna Davenport
Assistant Professor of Media Technology,
Asahi Broadcasting Corporation Career Development Professor,
Director of the Interactive Cinema Group

Reader:

Edith K. E. Ackermann
Associate Professor,
Epistemology and Learning Group

Reader:

John Watlington
Research Specialist,
Entertainment and Information Systems Group

ACKNOWLEDGMENTS

The following pages reflect the support, criticism, and friendship of many people. I am grateful to them all.

I want to thank my mother, Dolores Ledesma, and my father, Richard Elliott, for all their encouragement, emotional, edible, and otherwise.

Thomas Aguiere Smith and Stephan Fitch have been my IC buddies from my first day at MIT. I hope we never loose touch.

An elite group of test pilots patiently tried out my system and gave me valuable feedback. Thanks to Kevin Brooks, Glorianna Davenport, Nira Granott Farber, Mark Halliday, Gilberte Houbart, Michael Kirschenbaum, and Renya Onasick.

I want to thank Hideki Mochizuki of the Sony Music Entertainment Group and Yoshiji Nishimoto of ARIO for the opportunity to visit Japan and exhibit the video streamer in Sony's Art Artist Audition '92, and Chris Gant for handing me the application in the first place.

The Interactive Cinema Group has been a great place for me to get my head out of video post production and dabble with motion images in new ways. The best part about IC though is the diversity of people there. I'd like to thank Thomas Aguiere Smith, Carlos Alston, Kevin Brooks, Betsy Brown, Amy Bruckman, Stuart Cody, Ryan Evans, Stephan Fitch, Mark Halliday, Scott Higgins, Gilberte Houbart, Hiroshi Ikeda, Hiroaki Komatsu, David Kung, Lee Morgenroth, Natalio Pincever, David Tamés, and Koichi Yamagata for their comradeship, especially late at night.

My thanks also go to others beyond IC for many thoughtful comments along the way: Marc Davis, Debby Hindus, Marilia Levacov, Alan Rutenberg, Kris Thorisson, and John Wang.

My thesis readers, Edith Ackermann and John Watlington, scribbled gallons of red ink for my benefit. Thank you. David Tamés also helped me to polish drafts.

Finally, endless thanks go to Glorianna Davenport for her patience and confidence, and especially for the chance to be here.

CONTENTS

1 - INTRODUCTION 13

Thinking with Video	13
Kernels	16
Reconfigurable Multi-Streams	16
Associative Viewing	17
Time - Capturing and Holding Thoughts	19
Spatial Composition - Cousin of Editing	21
Overview of this Paper	22

2 - BACKGROUND and AIM 23

Video in the Land of 1's and 0's	23
Visual Thinking with Video	26
Traditional Ways	26
Proxies for Chunks of Time	28
Many Screens	30
Unconventional Presentations	31
WOW and SDMS	31
Elastic Charles	32
Salient Stills	33
Timelines	33
The Aim	34
A Hybrid of Viewing and Editing	34
Interface Goals	34

3 - THE VIDEO STREAMER 37

Form Meets Function in VideoLand 37

What's it Like? 39

Video Feedback, DVE Trails, Slit-Scans 39

Flip Books, Mutoscopes, Tape Delay 40

SDMS and Card Catalogs 41

XYT, Videographs, ... 42

How it Works 44

Beginning Notion: Buffer the Stream 44

Video Volumes - Time Will Tell 44

A Fish-Eye View 47

Sensible Sound 49

4 - THE SHOT PARSER 51

Why Segment into shots? 51

Related Work 52

The Algorithm 53

To Discern More 57

Framework 57

Other Information to Detect 60

5 - THE COLLAGE 63

Working in the Collage 65

What's Going On 68

Keeps Shots Fresh in Mind 68

Larger Context 68

Perpetual Motion 68

Collecting Thoughts Visually 69

Filtering and Culling 69

Grouping and Partitioning 70

Ordering - Sequencing and Sorting/Ranking 70

Example: Drawing a String of Shots from a Collage 71

Critical Mass 72

Critical Viewing - To See for Yourself 72

Pictures and Words 73

6 - EVALUATION 75

Evaluation	76
Setup	76
Feedback	76
Streaming in Yokohama and Tokyo	77

7 - DESIGN APPROACH 81

Genesis of the Streamer	82
Design Approach	83

8 - NEXT 87

Continued Work	87
Stream On	87
Collage+	91
Potential Applications	92
In an Editing Toolbox	92
Interactive Installations	94
Video Scratch	96
Browser	96
Interactive Cinema	97
Merely Notions	97
What Does Video Feel Like?	97
Desktop Video	99

9 - CONCLUSION 101

BIBLIOGRAPHY 103

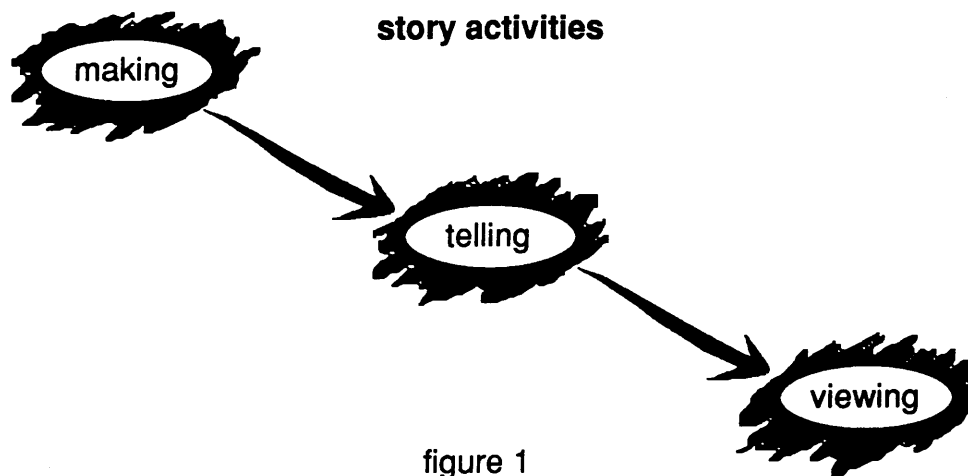
1 INTRODUCTION

Filmmakers seek to maintain reverie in their audience. Reverie is the state of daydreaming, directed by filmmakers, that a viewer succumbs to while watching a film or video. Filmmakers also enjoy their own sort of reverie creating a film or video. This more creative reverie comes as one shapes and influences the story. Novices rarely experience this second reverie, partly because the tools are so cumbersome that it is far easier to get bogged down in the mechanics of motion picture technology than to get caught up in the reverie of composing thoughts via motion pictures. Shooting video has become as simple as taking snapshots (shooting well is another story). Yet it seems the mechanics of editing are so far removed from the medium, motion imagery, that of course it should require the passion of a filmmaker in order for one to withstand the years of apprenticeship necessary for the mechanics to become second nature. Only then can one really think in terms of the creative medium. This thesis seeks to nudge the boundaries of what we currently call editing in a direction that supports more intuitively working and thinking with moving pictures. It explores ways of meaningfully visualizing time and ways of cultivating a collage of motion picture clips to maintain multiple thoughts and perspectives of a body of footage. The focus here is on the viewing experience, but this thesis has implications for traditional editing as well, since editors are often viewers themselves as they become familiar with a body of footage.

1.1 Thinking with Video

One way of explaining the research interests of the Interactive Cinema Group (IC) at the Media Lab is that IC is researching computer assisted story making, telling, and viewing, with special attention to the medium of motion pictures. **Story making**

(production) is typically highly collaborative. It goes through well defined stages, often progressing from concept, to scripting and planning, to shooting, to editing. **Story telling**, or story “presentation”, for film and video has typically been an uninterrupted “performance” of the results of the editing stage from beginning to end, often before a large audience. **Story viewing** is the activity of the audience and it doesn’t normally directly affect story making or story telling. The most influence, or interaction, an audience has is usually just box office receipts or TV rating points. Cinema has historically treated these three activities as *stages* in the life of a story. Influence typically flows in one direction, from making to viewing.



As interactive video evolves, new bonds between these activities are forming that allow influence to flow more freely when appropriate. A few simple examples might help to illustrate:

- A) A lot of work in interactive video to date has concerned the bond between viewing and presenting. Interactive pieces are often produced in a manner very much like the production of non-interactive pieces, but are designed to allow the viewer to influence how the material is presented. (laserdiscs for example)
- B) When the viewer can also contribute new material to the presentation, influence flows from viewing to making as well.
- C) More iterative production results when the presentation influences production. A simple example of this is the introduction of video taps and off-line edit systems to film sets.

**interactive
story activities**

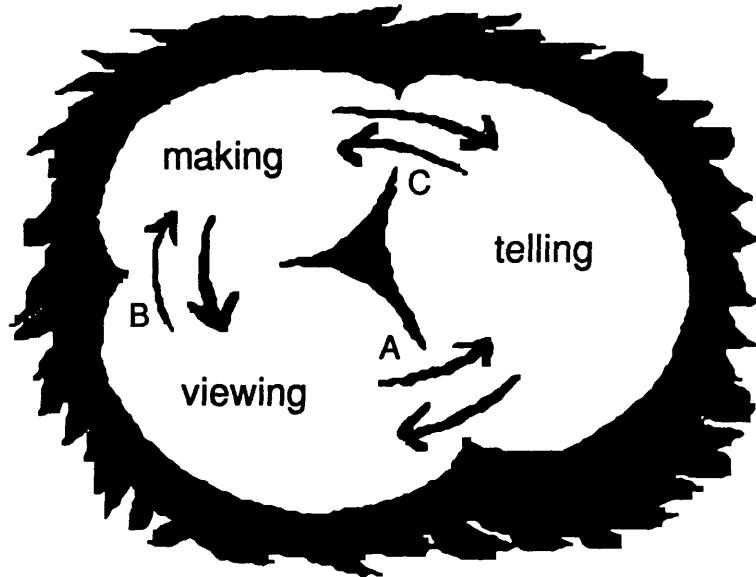


figure 2

Part of the work of the Interactive Cinema Group is to suppose, implement, and evaluate new ways of integrating these story activities. Digital video¹, a highly malleable motion picture technology, is often used in IC's pursuits. This thesis project focuses mainly on the third story activity, viewing, by putting some of the motion picture manipulation of story making in the hands of viewers.

More specifically, this thesis and corresponding research project concern ways to support thinking with video. By *thinking with video* I mean the processes of comprehension (mental reconstruction), that occur in a viewer's mind while watching a video presentation. This thesis portrays video as a stream for a viewer to better perceive characteristics of time. It explores non-linear viewing and thinking of sequential footage by providing basic tools for establishing parallel associations among elements visually in a collage. This thesis concentrates on bonds between viewing and making as they come to share tools for manipulating footage as a way to better to get at what's going on within a body of footage. Although this project targets story viewing, given that part of editing involves reviewing and analyzing footage, it may ultimately have implications for story making as well.

¹ I frequently use the term *video* in this paper to mean moving pictures, regardless of whether they originate on film or electronically, or whether they are conveyed via analog or digital signals. The thinking medium this thesis focuses on is motion pictures, while the particular motion picture technology it employs happens to be interactive digital video.

What's the difference between tools for thinking *about* video and tools for thinking *with* video? To me, the first includes tools specifically designed for analyzing motion pictures, or thinking about the abstract structure of a motion picture stream and its content more than about its expression. The second refers to the combination of video and appropriate tools for manipulating video as a means of thinking about some subject. It's analogous to the difference between thinking about sketch pads and thinking with sketch pads. Thinking *about* X makes X the object of thought, while thinking *with* X makes X the vehicle for thought about some subject, Z. This gets back to the aim of maintaining creative reverie and viewing reverie simultaneously. If the emphasis is on manipulating motion images, more weight goes to creative reverie. If the emphasis is on manipulating motion images as a way of thinking about their content, there is better balance between the two reveries.

Considering the broad view of interactive cinema given above and the specific goal of this thesis, this "thinking with video" project may offer one bridge from figure 1 to figure 2, more likely one piece of that bridge. The tools developed here work with any standard video source, from raw personal footage to highly polished commercial footage. Feeding off this wealth of unstructured video², though it likely has not been produced for the purpose of viewer interaction, it might be useful to think how to take advantage of these types of tools as IC does integrate story activities.

1.2 Kernels

Several notions and terms concerning thinking with video are essential to the discussion ahead. Among them are reconfigurable multi-streamed viewing environments, how we might exploit them for building associations, and a model for time.

1.2.1 Reconfigurable Multi-Streams

Environments containing multiple video sources are common. TV news control rooms, video post-production suites, surveillance systems, and video walls are just a few examples. Also, the dynamic presentation and control of a single video source is

² Unstructured video is just a continuous series of still image frames. Structured video contains additional information regarding how this series of frames is segmented and how various segments relate, even if they are not contiguous.

becoming more common on personal computers. Reconfigurable multi-streamed displays are less common. These are mostly found somewhere in production chains. Reconfigurable multi-streamed environments for viewing are perhaps nonexistent. The possibilities for dynamic display of multiple video clips are endless. An environment combining **many video streams** and **tools for managing** them promises to be greater than the sum of these two features. New ways of appreciating content may surface, especially through the process of tinkering with one's footage. By developing a dynamic viewing environment with a few basic tools for quickly saving clips from a source stream and for rapid recombination of those clips in a collage, this thesis suggests some possibilities for visual thinking tools.

1.2.2 Associative Viewing

Associative viewing is a term I will use to refer to taking simultaneous perspectives of a subject presented in motion images. Anticipating the collage, let's first look at associative viewing with regards to a subject presented in film or video.

In perceptual and cognitive psychology, Constructivist theory holds perceiving and thinking to be active meaning-making processes. Constructivist theory maintains that as a story unfolds, viewers unconsciously undergo hypothesis forming, testing, and confirmation or reforming. Shots and sequences trigger thoughts in the viewing mind, and viewers literally "reconstruct" their version of the story in their minds.³ Watching video, we "digest" its shots and sequences to recast plot and theme in our minds. Whether innate or learned, this capacity for digesting and recasting a story often occurs unconsciously, while we consciously experience the impact or symptoms of this process. Watching a string of shots, we build a family of associations and cause and effect relations. This all happens over time through a series of images that lodge themselves in our viewing minds. Eventually this collection grows to a sufficient body of sequences and associations to establish a program's message or impact. Sometimes this reconstruction doesn't take long, as in the case of a fifteen second commercial. Even a single shot can have complete impact. For example, a shot of baby's first steps in a home video speaks so concisely to its limited audience

³ David Bordwell speaks of this as "the viewer's activity" in *Narration and the Fiction Film*. Bordwell states that "The narrative film is so made as to encourage the spectator to execute story-constructing activities. The film presents cues, patterns, and gaps that shape the viewer's application of schemata and the testing of hypothesis."

of family and friends because that audience already holds the thoughts and emotions that the shot can trigger.

As shots and sequences play out, a collection of images grows in our minds. Some are memories. Some are associations. When we have formed enough images we can begin to glimpse constellations of meaning comprised of these visual memories and their interconnections. Part of these meaningful interconnections owes to montage, the meaning produced by combining shots in time.⁴ J. Dudley Andrew's summary of Eisenstein and his theory of montage notes psychologist E. B. Titchener's suggestion that it takes at least two sensations to make a meaning.⁵ As a train of shots triggers a series of thoughts in our minds, memories, the residue of those thoughts, can exist side by side and by the dozens as we continue watching. We create and keep much more than sequential relationships for these images in our minds. Viewing seems to transform the string of montage into a web of collage.⁶ Some of the associations binding the collage are directly motivated by montage. Others, while suggested within the video, are completed by the viewer.

The viewing tools implemented for this thesis include a video collage where viewers deposit clips they've selected from the video they are watching and then shape collections of these clips as a way of thinking about them. These basic tools for arranging a collection of clips are intended to foster this sort of associative thinking by providing an environment for associative viewing. As key moments in the source stream are saved and rearranged, the viewer can embody and explore different

⁴ Montage has many meanings that apply here. Montage literally means *assembly* in French. Regarding cinema, in the U.S. the term *editing* has been used to mean the work of putting together the shots of a film, especially connoting trimming and eliminating unwanted material. This is one meaning of montage. In Europe, montage has also meant editing, but with more emphasis on synthesis, building up from raw material rather than taking away. In the 1920's, Soviet filmmakers laid the foundation for theories of Montage. Pudovkin believed shots to be elements to be joined to build a sequence. Alternatively, Eisenstein employed montage as a means of colliding shots that didn't fit, creating meaning through juxtaposition.

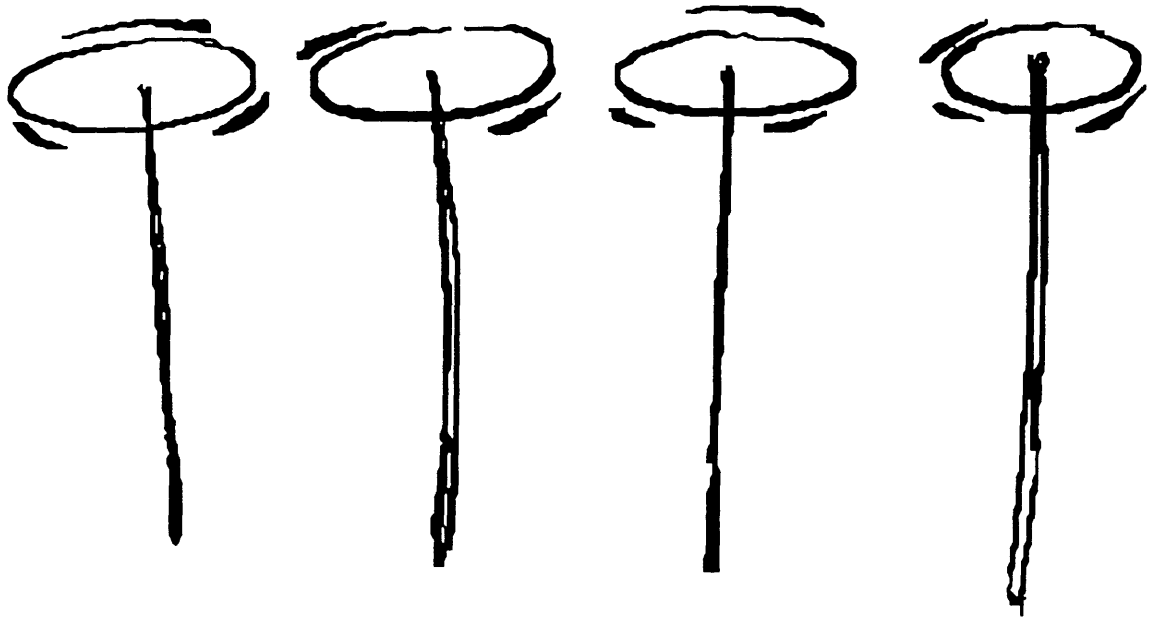
Montage also has a more specific literal translation from French. It originally meant "wiring up in electricity." One would speak of serial and parallel montage. It seems we've come full circle with montage as this thesis employs both serial and parallel electronic motion images. (See also *How to Read a Film*, pp. 183-190, for a general discussion of montage, and *Film Art*, pp. 385-388, for a summary of Soviet montage in particular.)

⁵ J. Dudley Andrew, *The Major Film Theories*, page 55.

⁶ "Collage" also comes from French, meaning literally "to cut out and glue".

perspectives. A viewer's collage may or may not reflect the ordering of the shots in a video stream, but it can allow simultaneous viewing of one element from various times in the source stream, a sort of cubist view featuring simultaneous perspectives across time, and simultaneous views of a variety of things, more of a catalog.

One might think of the collage as a collection of visual landmarks that provides an overview of a video stream by allowing us to see many moments at once. Though our eyes may focus on only one clip at a time, we are free to dart about to maintain all the clips in our mind. This reminds me of the old plate spinning stunt. The performer can only throw a spin to **one plate at a time**, but by rapidly jumping from one to another he can spin **many plates at once**.

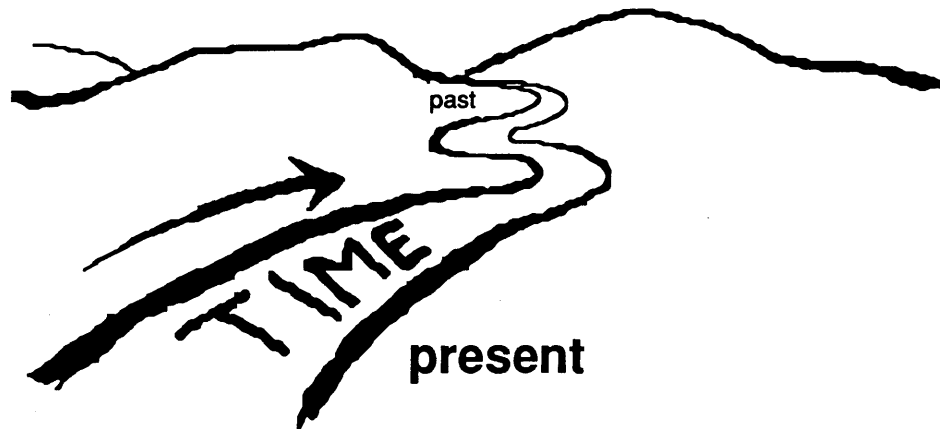


Looking at time, different details are apparent at different time scales. We can think about the frame, the shot, the sequence, and an entire program. Each of these provides context for the shorter ones. As montage creates meaning through context, it is important to consider a shot not only by itself, but also within the context of its scene and within that scene's context of the program. This is another example of **associative viewing**. By seeing many scales of context **at once**, we gain a holistic view of an element by associating it with its larger wholes.

1.2.3 Time - Capturing and Holding Thoughts

Let's look at time as an ever flowing river of events. Standing on the banks of this river we capture events from the stream of time and hold them in our mind. If we call

the events nearest to us our short-term memory, and the events further away our longer-term memories, it seems we can dwell on a given event to increase its prominence as it inevitably drifts from short-term to long-term. If we don't magnify events as they increase their distance, they eventually disappear across the horizon.



This is a very simple model for time and memory, but serves well for this discussion. Now, given time as a stream, there is a variety of time streams and views of them in our perceptual landscape. First, there is the roaring, unstoppable river of TIME, the unbroken flow we experience collectively. Then there is instantaneous time, the close-up view of this river, or of any other, at a single moment, a snapshot. There is recorded time, the capturing from a stream of events to form a repeatable stream at a later point in TIME. Given a recorded time-stream, we can divert it into TIME and re-view it, view it again. And from certain vantage points we can see multiple time-streams at once. Home videos might be lazy meandering streams of time. MTV is a torrent of images, a flash flood. With a music video's less-than-a-second shots our memories are pressed to the limit.

Note taking is one way of aiding memory. Saving thoughts on paper helps to store them somewhere less volatile than our minds. Additionally, the act of writing notes helps reinforce the memories in our minds. As we perceive and think about visual information we need ways to take visual notes, video memories. A large difference between video and print is that it's pretty cumbersome to review a piece of video, especially while in the midst of it. The viewing tools implemented for this thesis offer both a way of freezing a video stream for review and a way for keeping visual notes as its set of tools for thinking with video.

1.3 Spatial Composition, Cousin of Editing

Conventional video editing suites typically have many source monitors, but usually only one is playing at a time (except while constructing a mix effect). Also, traditional video editing is a process of ever forward progression, concentrating on one or two shot relations at a time. Film editing and, just recently, non-linear video editing provide alternatives to this, but they are still twiddle, review in motion, pause, and twiddle some more processes. Due mostly to motion picture technologies, we've only ever worked with one picture in motion at a time. One question this thesis poses is "Can we, and do we want to, pay attention to multiple motion images at once?"

There has been very little research into motivations for amateur videography, especially with regards to editing.⁷ One reason many people don't edit their camcorder footage might be that they don't have a narrative in mind when they shoot it. Therefore, if the composition of home video footage is not part of a planned production with a specific vision in mind, what can motivate one to rearrange their material? In other words, how can we support the reverie of filmmaking, especially editing, for novices? Part of the intent of this thesis is to provide tools for non production oriented "editing". These tools might be regarded as utensils for constructive viewing. They have strong ties to editing's video manipulation tools, but their purpose is very different. The focus here is on assimilating and reorganizing a stream, while editing is concerned with creating a stream. The tools developed for this project include a utility for quickly capturing clips from a running stream, and a collage environment for viewers to hold clips/thoughts and mull them over while watching a source stream. Editing is the *process* of combining shots, working towards a sequence for presentation. The environment suggested here focuses on the *experience* of collecting and combining shots as a way of engaged viewing, working towards better understanding.

⁷ Richard Chalfen studies how home moviemaking actually occurs in "Cinéma Naïveté: A Study of Home Moviemaking as Visual Communication".

1.4 Overview of the Rest of the Paper

I've structured this paper in nine parts. The rest of this thesis will progress through a discussion of precursors and related work (section 2), a description of each of the components to the system (the streamer, section 3; the shot parser, section 4; the collage, section 5), an evaluation of how they function together (section 6), a short discussion of the design process (section 7), directions for further work (section 8), and a conclusion (section 9).

Section 2, Background and Related Work - Outlines “visual thinking” in the realms of video editing and presentation. Summarizes some less conventional presentations. Presents the problems and challenges this project addresses.

Section 3, The Video Streamer - Introduces the streamer by relating it to other systems, then describes how it works.

Section 4, The Shot Parser - Presents motivations for automatic shot parsing. Describes the parsing algorithm implemented for this project. Suggests a framework for a more sophisticated parser.

Section 5, The Collage - Describes how one works with the collage. Describes the functionality built into the collage and ways of organizing clips. Weighs user's involvement against machine assistance.

Section 6, Evaluation - Describes evaluation sessions, both informal user testing at the Media Lab and public poking in Japan, and what was learned.

Section 7, Design Approach - Summarizes design criteria and guidelines by outlining the genesis of the streamer. Discusses how design challenges were met.

Section 8, Further Work - Outlines some open issues for the streamer, parser, and collage, and suggests some further development. Lists some possible applications for the video streamer. Poses a couple more distant possibilities for video objects.

Section 9, Conclusion

2 BACKGROUND AND AIM

The previous section introduced our cast of characters: story activities; thinking with video; time; and spatial composition. This section ultimately presents the aim of the project. First it sets the stage by outlining traditional ways of working with motion images and of representing time. A few examples from the growing collection of innovative time renderings and multi-streamed presentations are also described here. We begin by summarizing the benefits digital video brings to this project.

2.1 Video in the Land of 1's and 0's

As a motion picture technology, digital video offers, and in many cases still only promises, many advantages over analog tape and film. The capabilities of digital video most salient to this project are reconfigurability, multiple streams, random access, and amenability to image processing.

Reconfigurability - Of these four, the attribute most essential to the remaining discussion is reconfigurability. Russell Sasnett defines reconfigurable video in his master's thesis, titled "Reconfigurable Video", to be a:

*"delivery method for video products, whereby subject matter and presentation format may be varied for different audiences using the same video source materials without recourse to programming. The purpose is to increase the versatility of video resources by allowing new content-specific accessing modes to be created for them by viewers."*⁸

⁸ Russell Sasnett, "Reconfigurable Video," p. 4.

This definition is a good starting point for how I use the term reconfigurable video in this thesis. For the purposes of this project I take reconfigurable video to mean video that is especially susceptible to rearrangement both temporally and spatially, by both viewer and maker. In other words, reconfigurable video is motion pictures presentable and viewable in other forms than single image, full screen, continuous, linear time. Reconfigurable video can be resized, relocated, and resequenced. It can even be re-rendered in ways that portray time in an instant. Ultimately, reconfigurable video is video whose presentation can be reformed at viewing time to shift emphasis or meaning of its content. Webster's first definition for *represent* is "to bring clearly before the mind." Reconfigurable video is motion imagery that can be re-presented by the viewer.

Multiple Streams - Digital images and digital image streams are often more elastic than their analog counterparts.⁹ Image elasticity affects bandwidth usage. The bandwidth requirement for conveying an image stream is a function of image resolution, frame rate, and encoder efficiency¹⁰. Image content can also affect compression (encoding), which in turn affects image quality and bandwidth usage. For analog pictures these attributes are fixed for a set bandwidth. Digital images have the potential to vary these attributes dynamically as necessary. Add to this that digital images are much more easily buffered and accessed than analog pictures. This gives digital images an additional benefit: if their need can be anticipated and if extra bandwidth is available ahead of time, they can be transmitted early.

⁹ *Elasticity* refers here to how flexible an image stream is at fitting various presentation requirements. A *brittle* image can only display at one size and frame rate. An elastic image is more variable and adaptable.

¹⁰ Image resolution and frame rate are related. **Resolution** is a measure, in pixels, of the spatial dimensions of a picture. For example, a picture might be 400 pixels wide and 300 pixels tall. A third dimension of image resolution is pixel depth, the color resolution of a pixel. Pixel depths commonly vary from 1-bit to 32-bits. Combined with display size and viewing distance, resolution affects the perceived pixel density. **Frame rate** is a measure of temporal density, or sample rate (e.g., 30 frames/second). Image quality is a subjective measure that has to do with how well a given resolution is exploited to reproduce a picture. It depends very much on the method of encoding, both spatial and temporal, and on the image content. Lossless encoding attempts to minimize bandwidth while preserving measurable quality. Lossy encoding sacrifices image attributes to minimize bandwidth while attempting to maintain perceived quality and is generally more efficient than lossless encoding in terms of how much it can compress a picture. Of course, resolution, frame rate, and encoder efficiency are all interdependent and optimizing bandwidth is not simply a matter of optimizing each element individually.

$$\text{bandwidth} = \frac{\text{resolution} \times \text{frame rate}}{\text{encoder efficiency}}$$

This digital elasticity makes for more efficient use of transmission channel bandwidth, and, given certain encoding schemes, for more graceful degradation of digital image streams as their requirements exceed bandwidth available.¹¹ When channel bandwidth and processing speeds are saturated, the elasticity of digital image streams can allow them to deteriorate incrementally. It's not an all or nothing affair. First to go might be image quality, then resolution, then frame rate, and finally sound. This way if you need a fifth stream across a channel large enough for only four full resolution streams, perhaps you can have them all at eighty percent resolution.

Random Access - Many digital video media offer instant access to any single frame. Video tape, analog or digital, can take minutes to shuttle from one end to the other. Analog laserdiscs have reasonably fast seek times, but are not commonly rewritable. Given the relatively quick access of many current digital video technologies (magnetic drives, optical drives, RAM), it's possible to maintain a continuous flow in playback time regardless of how that time is actually distributed on the physical media. This is essential for making time reconfigurable. Instant access also makes looping quite simple, which the collage exploits and will be discussed later.

Image Processing - Digital pictures are quite amenable to image processing across the span from specialized hardware to general purpose programming languages. The shot parser developed for this project (described in section 4) processes video image streams, detecting probable shot boundaries. Algorithms evolve faster in software than in hardware, so the shot parser was developed in C very quickly.

¹¹ Sub-band encoding and Apple Computer's "road pizza" codec allow for graceful degradation. MPEG does not.

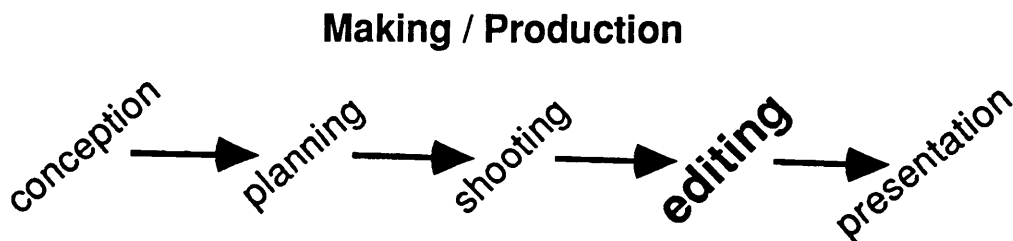
2.2 Visual Thinking with Video

A discussion of some existing ways of working and thinking with motion images here will help the discussion of alternatives explored by this project later.

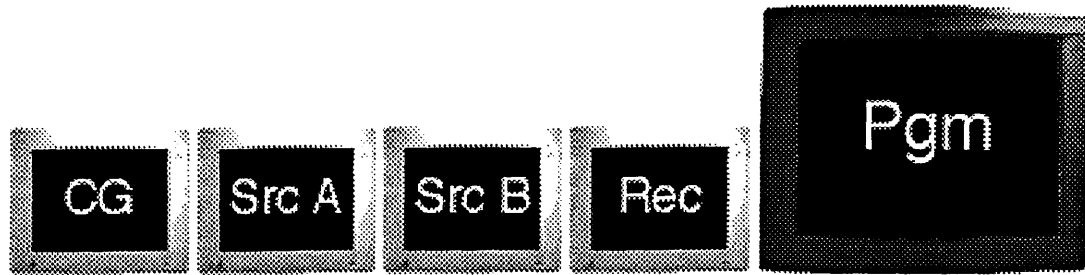
2.2.1 Traditional Ways

When we think about manipulating motion pictures and about drawing associations between motion pictures, we most commonly think about editing and montage. Editing usually connotes tight description of a sequence. Montage has more to do with associations across time. Montage creates meaning by juxtaposing shots in time. Multi-streams do the same by combining elements spatially in a collage. The methodologies of editing and the psychology of both montage and editing have implications for multi-stream viewing and composition.

Editing - Traditionally, motion picture production is a staged process that includes conception, planning, shooting, editing, and presentation. The inelasticity of analog media, film or video, keeps this a staged process. Production costs usually prohibit iterative production. Editing is just one step in this production pipeline and is the final creative stage. More specifically, editing may be thought of as the coordination of one shot with the next. The fundamental task of editing is to sequence shots.



Presentation of motion pictures is usually full-screen on a single screen. We normally consume our movies and television one stream at a time (although, fast channel flipping might be akin to spinning many plates at once). On the other hand, video editing suites usually have an array of monitors providing multiple views of the various sources for composition. Typically, each monitor is dedicated to a single tape deck or effects device, for example: source A, source B, the character generator, and the record deck. Their layout is fixed. Though these multiple screens support thinking about multiple shots a little better than a single screen does, it's rather inconvenient that monitors have to be tied to individual tapes rather than to individual creative elements, namely shots.



Segmentation - Just as editing is only one stage in the production pipeline, it also is usually accomplished in stages. Projects are frequently broken into an off-line stage where raw footage is filtered and recombined to create a first draft of the final piece, and an on-line stage where the off-line edit is tightened up and visual and audio effects are added. Preliminary off-line activities include reviewing and logging material, discarding unusable footage, marking locations on sources, and describing the footage. Current technologies often limit non-linear editing to the off-line stage where full “broadcast quality” is not absolutely mandatory. Since more and more non-linear edit systems employ digital video, while almost all video still originates on film or analog videotape, the processes of digitizing the video source (introducing it into the digital domain) and segmenting it into separate shots occur simultaneously as a tedious chore for an editing assistant. This involves shuttling the source tape to find the beginning of a take, or some more meaningful beginning of a shot, marking that point on the tape, and then searching for a meaningful end of the shot and marking it. Then with the desired portion of the source tape defined, some digitizing utility is launched to automatically shuttle the tape to the head of the shot and digitize it. Since this is a preliminary stage, some sort of description at this first viewing helps the editor organize the material later while sequencing it. Unfortunately, this process of segmenting can be extremely tedious and far too dislocated from the more creative parts of editing later on. For production, especially of a documentary, this may be where the “story” begins to unfold for the editor or director. This review and selection process is part of a larger process of assimilation of the material and deriving some meaning from it, and ultimately deciding how to present it.

Filtering and Selecting - There is another task that goes with segmenting at these early stages of editing. As you view the raw material and define the boundaries of elements, you decide which pieces you are going to use in your edit session. This task of filtering out the bad, or selecting the good, is a key stage in the life of a shot.

In film editing this is when a clip finds itself in a bin, or “on the floor”. In most video editing this is when a shot’s time-code is logged, or forgotten. Digital non-linear edit systems sometimes re-employ the concept of bins and display at least a representative picture from each shot residing in the bin.¹²

Montage - Turning from production to presentation, motion pictures are almost always conveyed in a single stream. Montage is the effect of creating meaning, or inducing sensations, by sequencing and juxtaposing shots within that stream. A shot of a character fleeing the scene of a murder followed immediately by a shot of the same character behind bars suggests to the viewer that the character was caught, tried, and convicted. A shot of a crowded intersection followed by a shot of a busy ant colony makes an analogy by juxtaposing the two shots. These are examples of associations made through sequence. Founded on sequential associations, montage is inherently single streamed and temporal.

Multi-streamed presentations, such as video walls, also convey meaning by combining streams spatially, just as montage does temporally. They create meaning by associating images in a parallel as well as a serial manner.

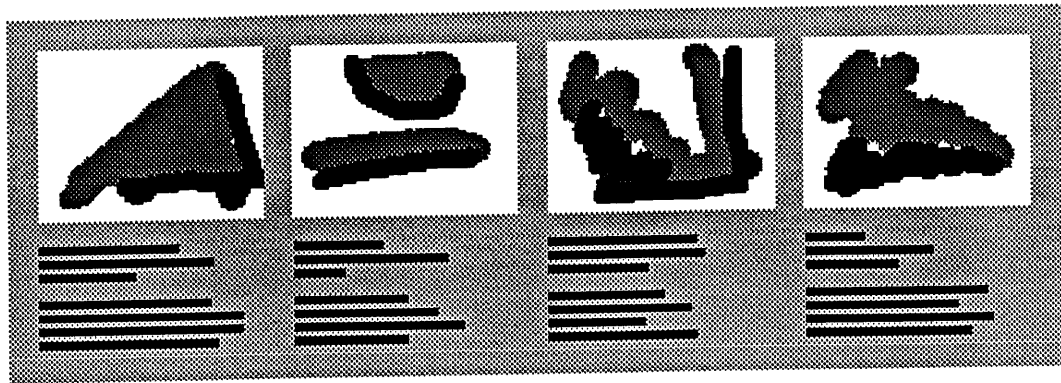
2.2.2 Proxies for Chunks of Time

Working with motion pictures poses the challenge of appreciating how to see the instant, the moment, and the whole sequence at once. Tools for visualizing these multiple time scales frequently employ an iconic still or clip to represent a larger chunk of video and a grid arrangement of these thumbnail views to depict a sequence.

Single Frame “tags”: Storyboards and Comic Books - Comic books tell stories through an array of still pictures. Similarly, motion picture editing often uses storyboard views to portray a sequence. Sketches are used in pre-production to represent each shot in a sequence. In post-production, many non-linear edit systems use a single frame to stand in for each shot. The stand-in frame is captured from the shot it represents. A comic book frame or a sketch in a storyboard can be rendered to

¹² The Montage Group’s non-linear video editing system, the *Montage Picture Processor*, first introduced at NAB in 1985, uses the notion of bins for holding shots. Each pair of monitors corresponds to a conceptual bin, regardless of where the elements of that bin are actually located on tape.

portray more than an instant in time, but there is always a question about which single frame in a series of video frames best represents the content of its shot and all of that shot's temporal characteristics. Some systems simply use the first frame from the shot. Others allow the person defining the shot to select a more appropriate representative frame (QuickTime's poster frames for example¹³).



Head and Tail Frames - For many editors, the most critical element is the transition between shots. Focusing on the transition, the most pertinent parts of a shot are the end of the previous shot and the beginning of the new shot. With this in mind, another approach to representing a chunk of time in an iconic view is to display the head and tail frames of a shot.¹⁴ In some shots, these two frames may appear identical. For example, talking heads usually look the same at the beginning and tail of the shots. In other shots, such as a panning camera move, the head and tail pair really do display some of the temporal aspect of the shot. For example, if the shot contains a pan and the end frames overlap, that camera motion might be apparent in the pair.

Micons - While single frames, or even pairs of frames, in a storyboard help to portray a sequence, they are not very good at showing the temporal characteristics of the single shot they represent. Micons are more dynamic stand-ins. Micons are short

¹³ QuickTime is an operating system extension for the Macintosh that manages time based media. QuickTime offers standard controllers for time media and standard codecs for compressing and decompressing images and image streams. In QuickTime terminology, a *poster frame* is a user defined frame that is used to refer to an entire video clip.

¹⁴ The Montage Picture Processor editing system uses head and tail frames to represent shots. The Montage uses a 7x2 array of monitors to display the head and tail frames from the current clip of each of its seven bins. The top row of monitors displays the head frames from each shot and the bottom row displays the tail frames.

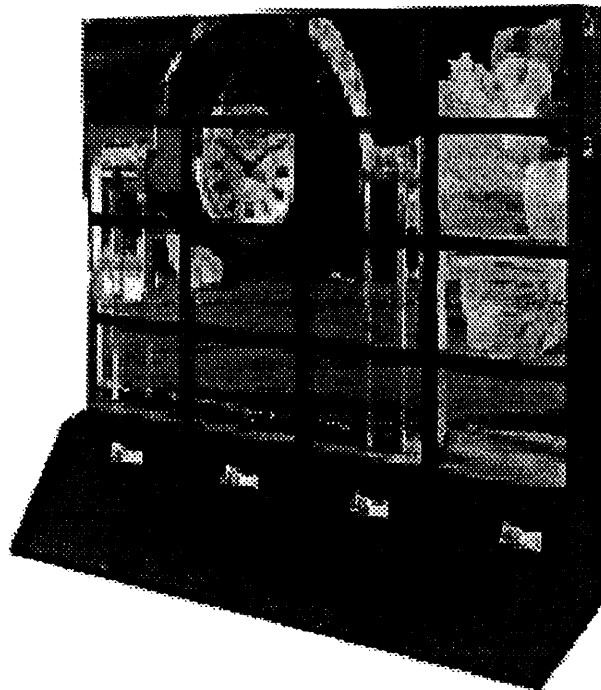
postage stamp sized clips of digital video that represent longer full-screen clips of video. Because micons are in motion, they represent the shots they refer to better than do still frames. Micons are also useful for representing short sequences of shots. Some of the most effective micons span a cut, perhaps owing again to the idea that meaning comes from a combination of elements.^{15,16}

2.2.3 Many Screens

There are many examples of multi-streamed viewing environments. On the production side, a network news control room or any video post-production suite offer examples of multi-streamed systems. On the receiving end, video walls and picture-in-picture (PIP) televisions are examples of ways viewers experience multiple image streams. These all have a fixed layout though and are not reconfigurable.



Picture-In-Picture television



video wall

¹⁵ Micons were first developed for "Elastic Charles", an electronic magazine prototype produced by the Interactive Cinema group in 1990. See Brondmo's and Davenport's "Creating and Viewing the Elastic Charles - a Hypermedia Journal" for a complete description.

¹⁶ "A thing in itself never expresses anything. It is the relation between things that gives meaning to them and that formulates a thought." -- Hans Hofmann, *Search for the Real* (1967).

2.3 Unconventional Presentations

2.3.1 WOW & SDMS

In a 1981 MIT Architecture Machine Group project called “World of Windows”, Richard Bolt and Mike Naimark implemented a simulation of a display providing upwards of 20 moving pictures at once. It also explored passive viewer control of audio. The content and its layout were prerecorded onto laserdisc, therefore it was not reconfigurable, but it did provide a glimpse of what viewing multiple video streams in a dynamic display with some viewer control might be like.¹⁷



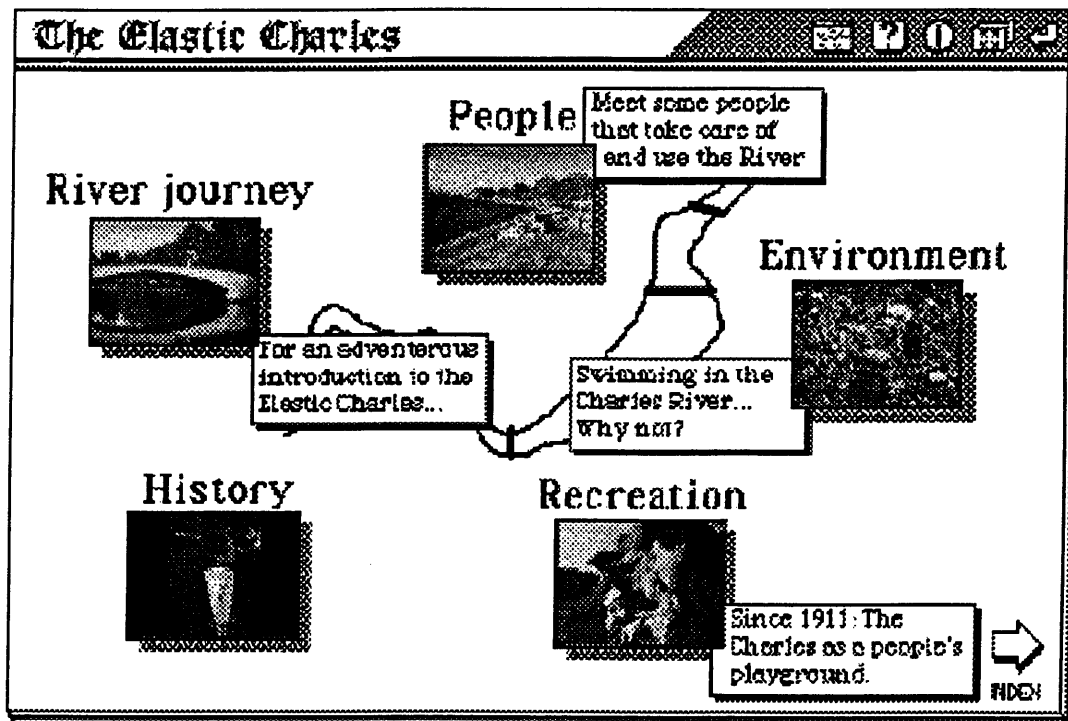
Richard Bolt was also a co-conspirator with Nicholas Negroponte in a slightly earlier Architecture Machine Group project called the Spatial Data Management System (SDMS). The fundamental notion of the project was that “we find items on the basis of a more or less definite sense of their location in a familiar space, which space may be actually present or remembered.”¹⁸ The SDMS was a very early example of using a virtual desktop on a computer display as an organizational and memory aid for multiple media types.

¹⁷ See Richard Bolt’s “Gaze-Orchestrated Dynamic Windows” for a description of the World of Windows project.

¹⁸ Richard Bolt, *Spatial Data-Management*, p. 7.

2.3.2 Elastic Charles

The Interactive Cinema Group's interactive documentary about the Charles River, "Elastic Charles", and its accompanying "Mimato" tool set for managing micons and video segments provided for scripted control by an author as well as for reconfigurability by the viewer. Within "Elastic Charles" one could view multiple micons on a screen at once, and there was HyperTalk script level support for managing their display. The Mimato tool set's limitations on the viewing end were that clip digitization was not real-time, and its digitized clips were often too short, though they served well as micons. These concerns have since been addressed by desktop digital video.¹⁹ While "Elastic Charles" did embody some of the notion of viewer as maker, allowing the subscriber/viewer to alter and add new content links, it was designed to fit the paradigm of interaction within a prescribed presentation. Since "Elastic Charles", the Mimato tools have also been used to supplement other projects with micons.



¹⁹ The "Elastic Charles" was produced on a Macintosh in 1989/1990. With the appearance of QuickTime in 1991, video digitizing has become real-time and clips on the order of minutes and hours can be captured and replayed.

2.3.3 Salient Stills

Laura Teodosio, working with Walter Bender in the Media Lab's Electronic Publishing group, has developed an innovative way of transforming moving images into still images she calls salient stills.²⁰ Salient stills render time in a single still frame by algorithmically correlating a series of motion picture frames in terms of size and position and then averaging them together into a single frame. For example, all the frames in a pan are correlated to find the overlap between each pair of frames. Then a single panorama frame wider than the video frames it was derived from is rendered by averaging the frames together while shifting their positions so their overlaps match up. A sequence of frames containing a zoom would yield a salient still with more resolution in the part of the image that the zoom was directed at. Salient stills don't reflect the order of events in time (for example, the panoramic salient still doesn't indicate whether the camera panned left or right). Instead, they attempt to show in a still frame everything that happened within the camera's view within a certain span of time.

2.3.4 Timelines

David Small, at the Media Lab's Visual Language Workshop, has developed another new way of rendering some of video's temporal nature in a still picture. Small's work takes the shot length information derived by a shot parser to place a representative frame from each shot into a timeline, spacing between shot frames according to the shot lengths. It doesn't give any indication of camera motions within shots or transitions between shots. However, using the vast screen real estate of the VLW's 6000 x 2000 pixel display, it gives a good visual sense of pacing at a glance.

Working with motion image sequences, we often need to see them at a variety of time scales. Michael Mills, Jonathan Cohen, and Yin Yin Wong, at Apple Computer, have created an interface prototype called the "Hierarchical Video Magnifier" that allows one to nest timelines in a hierarchy of finer and finer temporal detail.²¹ This technique provides multiple temporal contexts for video elements within a limited screen space.

²⁰ Laura Teodosio, "Salient Stills", MIT Media Lab master's thesis, June 1992.

²¹ Michael Mills, Jonathan Cohen, and Yin Yin Wong, "A Magnifier Tool for Video Data", *Proceedings of CHI '92*.

2.4 The Aim

One primary goal of this project has been to explore ways of thinking with video, especially ways that engage the viewer and ways that might not be so sequence oriented, not as linear. Another way of stating this goal is to describe some of the challenges the project has faced.

2.4.1 A Hybrid of Viewing and Editing

Editing is always the most evident example of how and why one would manipulate video clips. Designing and presenting this project, it has been too easy to think of active manipulation of video clips solely in terms of editing as it refers to constructing a sequence for presentation. The tool set developed for this project is better described as an environment for exploration than for production. This tool set is intended more to help users discover associations within footage as they watch it than to help a designer implement an interactive video scenario for others. However, this project does have implications for editing, especially when we consider editors to be viewers as well, such as in the early stages of editing when they need to review footage to become familiar with the story it will tell.

2.4.2 Interface Goals

The tool set developed for this project makes up an interface for viewers to digest motion image streams. Designing such an interface poses many challenges. Some of the larger ones have been:²²

- How to balance the viewer's/user's attention to the source stream against their manipulation of collected clips.
- How to keep the motion pictures visually prominent. In other words, how to keep the motion pictures the medium for thinking rather than dwelling on buttons and sliders as the interface to motion pictures.
- How to invite exploration. How to make the system easy to learn and use.
- How to combat creeping featurism. How to avoid adding more and more tools past the point of being understandable and usable. How to design a few flexible tools rather than a complex set of specialized tools.

²² Some of these interface goals are borrowed from a broader list in Donald Norman's *The Design of Everyday Things*.

Motion Images carry time. Portraying time effectively is a complex problem. Among the family of issues concerned with rendering time are:

- How to portray motion in a still image.
- How to display video frames so they portray their own temporal characteristics clearly (shot durations, transitions, camera motions, editing rhythm, etc.).
- How to associate a view of motion within the frame with a view of motion beyond the frame.
- Different ways to effectively render different time scales and how to relate them.

An overriding challenge to the whole interface is how to streamline the tools so that thinking with them can keep pace with just straight viewing. Motion pictures want to stay dynamic. Viewing needs to stay fluid. If the tools trip up either, they may hinder the wonderfully complex and subtle thinking we do in linear viewing to the point where their cost exceeds their benefit. Though the tools cannot approach the complexity of viewing, they should approach the speed.

Sometimes there are so many more things we can do with a computer than without. Sometimes there are so many fewer. Ultimately, the tools need to be simple and loose enough to support the diversity of ways people think and the diversity of ways motion images offer themselves to be thought of.

3 THE VIDEO STREAMER

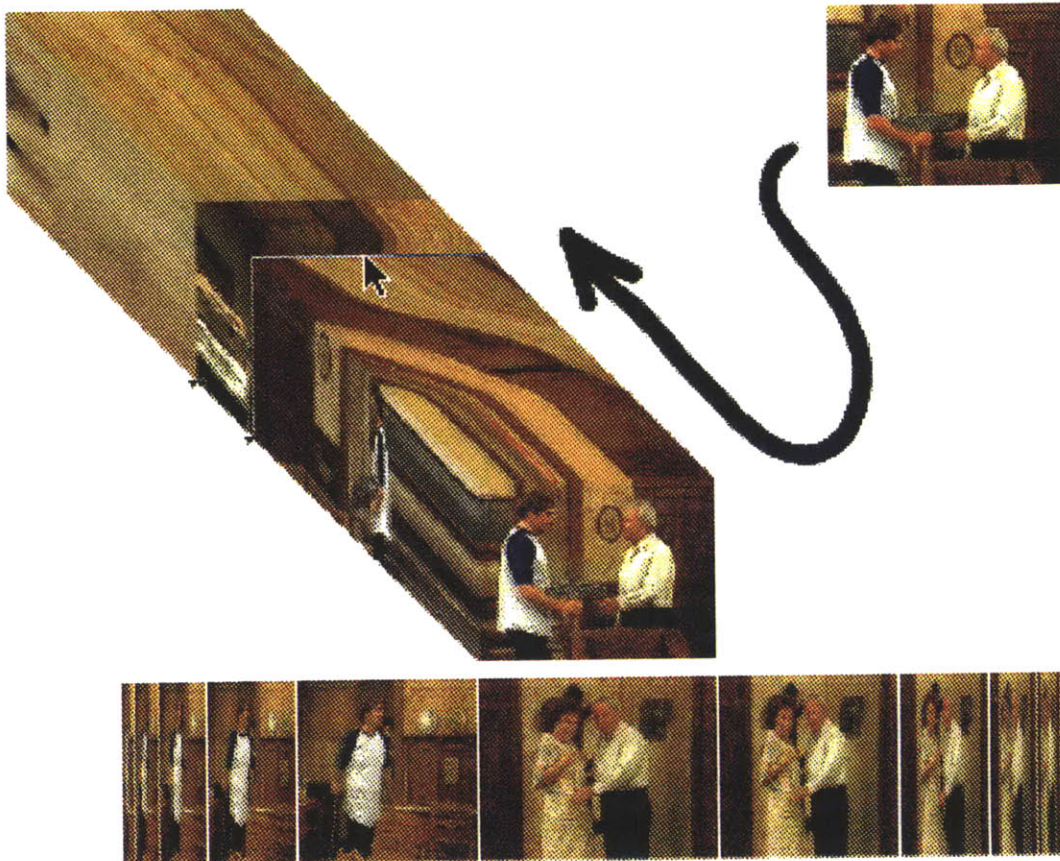
Many of us work with video often. Most of us see and watch video daily. Video, the electronic moving pictures and sounds that carry movies, sports, music, and news to us every day, is usually viewed full screen and in full motion. It is a continual stream of images and their information that incessantly flows into and out of our minds. New images are quickly displaced by even newer ones, often inducing a sensation of deluge without lending themselves to comprehension.

The video streamer serves many purposes, ranging from a utility to review and select segments of video for recombination later, to a dynamic painting of a moving picture stream that provides alternative ways for one to appreciate the rhythm of colors and motions embedded within video.

3.1 Form Meets Function in VideoLand

The video streamer works in two simple modes: flowing for capture, and paused for review. When the streamer is flowing, the video source, which can come from broadcast, from a camera, or from videotape, is arranged as a three dimensional solid, stacking the picture frames in a block, much like the pages in an animation flip book. As new frames enter, they push the older frames towards the upper left of the display, further away in distance and in time. The sides of this block, formed by the edges of all the frames, reveal a number of temporal attributes of the stream. Shot boundaries and the editing rhythm are clearly visible along the sides of the block, as are many camera motions.

This is simply a way of rendering a video buffer, a series of frame buffers, that paints and repaints each frame at its position relative to the current time rather than at its absolute physical position in the buffer, honoring the dynamic nature of video streams.



By pausing the streamer, the user can review the contents of the block of frames. Using a mouse or stylus, individual frames are selected for display by placing the pointer over the edges of those frames in the extrusion. Moving the cursor quickly across many frames displays those frames in motion, similar to thumbing through a flip book.

The video streamer displays approximately twenty seconds of video, portraying second to second dynamics of the video stream. The *frame-viewer* shows us more precise frame to frame relationships, based on which portion of the video stream currently pointed to. Presenting and synchronizing a view of a shot and a view of a frame at the same time allow us to see the flow of the moment and the detail of the instant at once. We can see the forest and a tree at the same time.

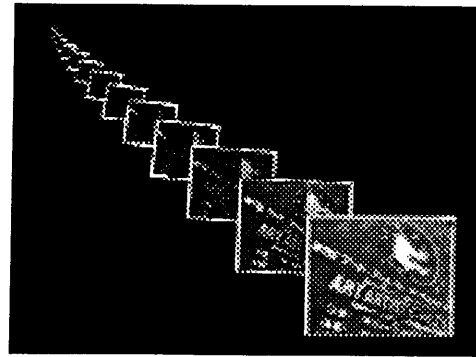
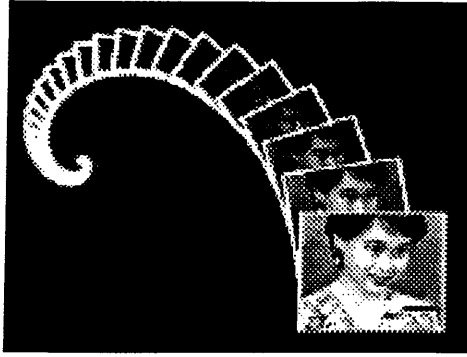
The shot parser uses a frame differencing algorithm to segment the stream into manageable chunks. The video streamer and its accompanying frame-viewer and shot parser provide a small tool set for analyzing and appreciating a stream of moving pictures in various ways. This tool set can combine with other tools and applications to present multiple perspectives of video streams. In this thesis project it supplements a display system for collages of video clips. Because it presents itself both in motion and paused, and because one can view the stream as is or dive in and manipulate it, the video streamer offers a series of conceptual bridges, from the concrete experience of watching moving pictures, to a slightly removed perspective (stepping back from a view of the frame to a view of time in the stream), to an entirely abstract analysis of the signal as it might relate to features of the stream's content (shot lengths and cutting rhythm). These separate views are synchronized and displayed simultaneously, providing a coherency that helps bridge concrete and abstract appreciation of the stream.

3.2 What's it Like?

I am almost convinced that nothing is unique and that almost nothing is original. Sometimes it is helpful to describe something by describing what it is like. This section illustrates the streamer's view of motion images by describing some of its parents and siblings.

3.2.1 Video Feedback, DVE Trails, Slit-Scans

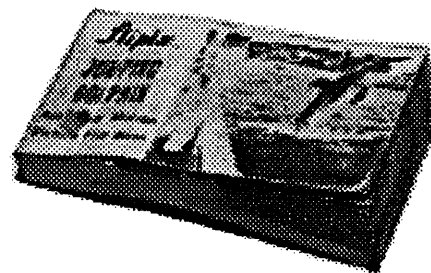
I've been asked whether the streamer has been inspired by drugs or by video feedback. No comment on the first. The streamer does owe something to video feedback, and perhaps DVE trails, photo finish slit-scans, flip books, and mutoscopes. The streamer looks similar to feedback in the way older images flow away from the foreground into the distance. It also looks like DVE trail effects where the background is not refreshed as the frame of video moves across the screen, except in the streamer view the current frame is stationary and the background trail flows away from it instead.



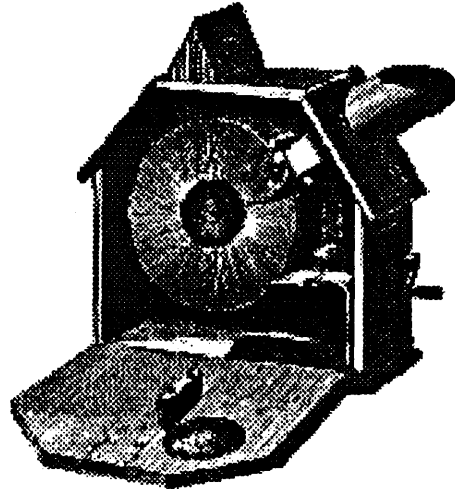
The sides of the streamer also owe something to slit-scan photography, especially the photo-finish technique used for horse races. Slit-scan photo-finishes expose a narrow strip of film through a vertical slot and move the film past the slot without shuttering. This yields a horizontal strip of film that shows a line in space, a finish line, across a short span of time. If the video camera is stationary and somebody walks into the frame, the effect on the side of the streamer is identical to a photo-finish. A similar visual effect happens on the sides of the streamer when the video camera pans or tilts across the scene. (Notice “Meathead” on the left side of the stream on page 38.)

3.2.2 Flip Books, Mutoscopes, Tape Delay

Operationally, the streamer is similar to flip books, mutoscopes, and talk radio’s seven second delay. The streamer and flip books both stack a series of frames to form a 3D volume of moving pictures. Stroking the edges of the streamer for playback is analogous to thumbing through the pages of a flip book. With certain flip books, if you bend them to reveal the edges of the frames, you can get a reproduction of the streamer’s slit-scan time view.



The mutoscope is a viewing apparatus developed in 1894 where the picture cards mounted on a revolving axle are “flicked” by turning a handle. The mutoscope is essentially a rotary flip book. The streamer’s wrap-around video buffer is conceptually the same as the mutoscope’s rolodex-like stack of pictures.²³



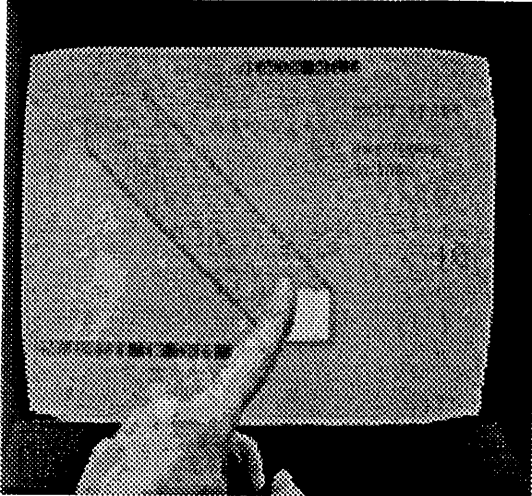
In a way, the streamer also functions like the seven second delay used in talk radio. The streamer buffers the video for selection purposes and the seven second delay is a looping audio buffer used for filtering. A viewer can select a segment of *video to keep* from the streamer’s buffer of recently viewed footage. A talk show host can select *audio to censure* from the delay’s buffer of recent conversation. Both use buffering the same way, except, without user intervention, the streamer discards, while the delay broadcasts.

3.2.3 SDMS and Card Catalogs

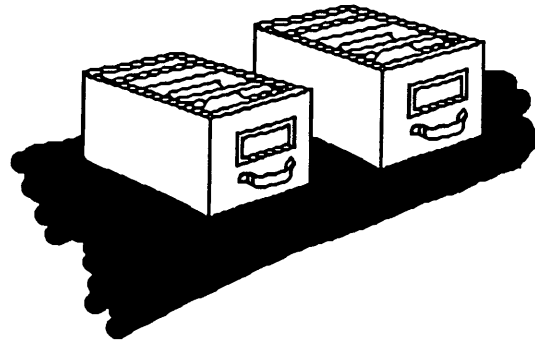
The Architecture Machine Group’s Spatial Data Management System contained a slide collection. This collection was presented on a small screen as a stack of blank frames representing the slides. To access individual slides the user would slide a pointer along the stack, called a “key map”. This would index the frame pointed to and display it on a large screen. The video streamer and the SDMS key map are analogous to a library card catalog as a volume for indexing material.²⁴

²³ Franz Paul Liesegang, *Dates and Sources*, p. 58.

²⁴ Richard Bolt, *Spatial Data-Management*, pp. 42-45.



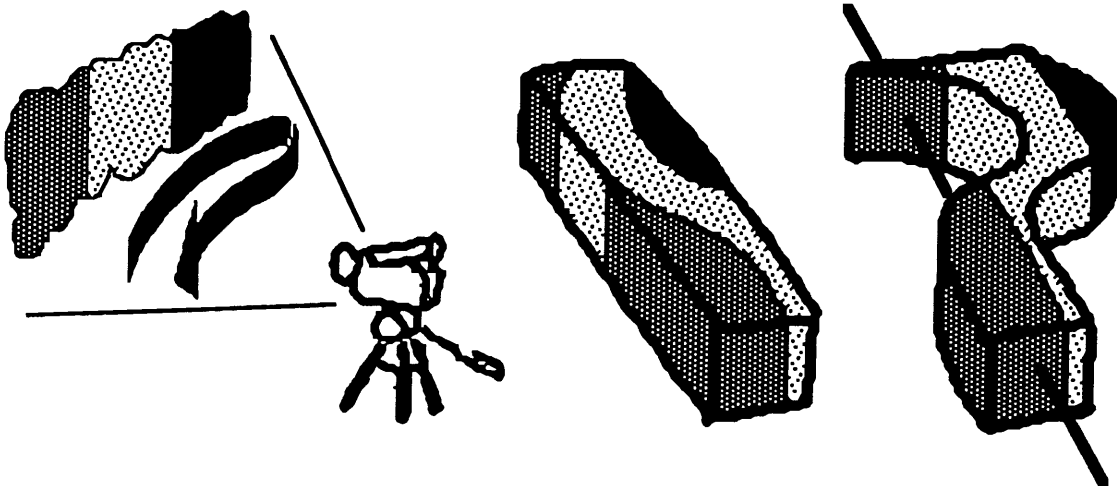
SDMS "key map" for slides



3.2.4 XYT, Videographs, ...

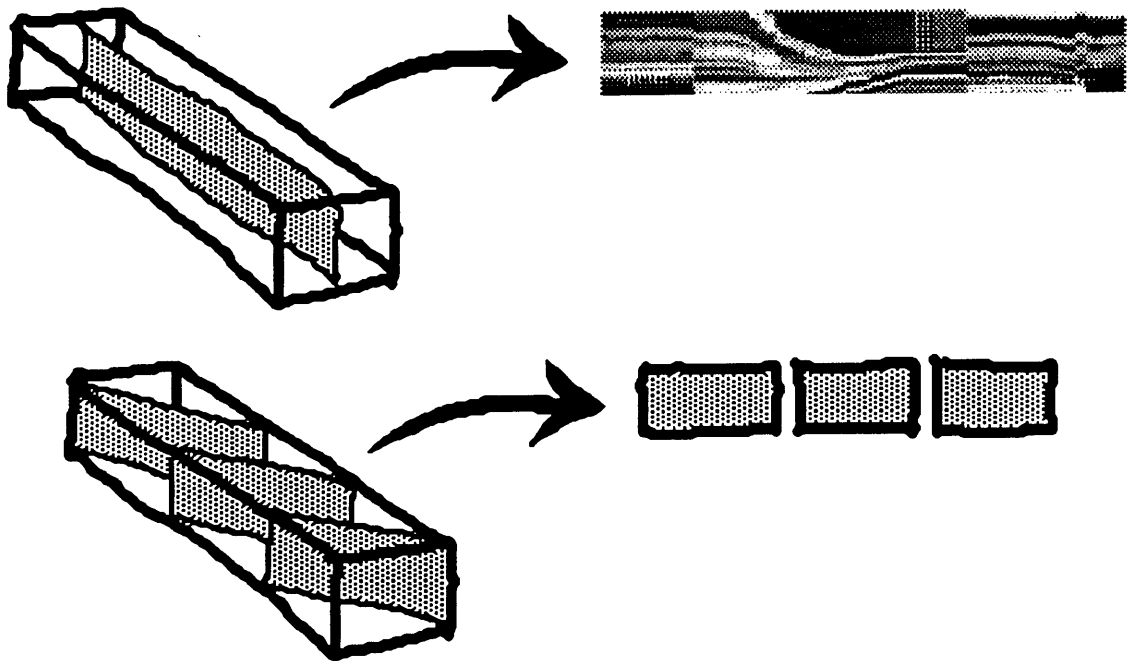
The streamer uses a volume of video for its video capture interface and for review and selection of captured footage. Video volumes are also currently being used for image processing. And slit-scan cross-sections of video volumes are being used as static representations of video clips across time.

Ted Adelson and John Wang, of the Media Lab's Vision and Modeling group, are exploring the use of video volumes for image processing. They have coined the term XYT for three dimensional arrangements of video frames stacked along the time axis. Successive frames are correlated horizontally and vertically to line up the frames so that a ray through the time volume would pierce a single line of sight. Then the redundancy of the frames is used to filter out image noise for each ray of pixels.



Ron MacNeil, at the Media Lab's Visual Language Workshop, Michael Mills, Charles Kerns, and Mitch Yawitz, at Apple Computer, and Akio Ohba, at Sony, are using slit-scan views of video clips to present the dynamics of those clips across time within a still image. Ron MacNeil's *videograph* takes a vertical slice, four pixels wide, from the center of each frame and lays these slices out horizontally on a timeline. Videographs are tied to an annotation system and are used to represent the rhythm of a moving picture in a still image. Apple and Sony, in similar manners, approach the same problem of representing the dynamics of motion in a still by also taking a slice from each frame, but offsetting the location of that slice horizontally across time. This yields a diagonal slit-scan from the left side of the volume to its right.

Ohba's "Video Browser" also presents motion picture data in a timeline for quick browsing across a large span of time.²⁵



²⁵ Akio Ohba, "Interactive Video Image Coding & Decoding, (Two new tools for video editing and computer animation)", 1990.

3.3 How it Works

There are many aspects of the video streamer to describe here, from the basic notion of buffering a video stream, to how the streamer portrays time, to how the user works with the streamer for reviewing images and sound.

3.3.1 Beginning Notion: Buffer the Stream

A typical scenario for digitizing video from a live source:

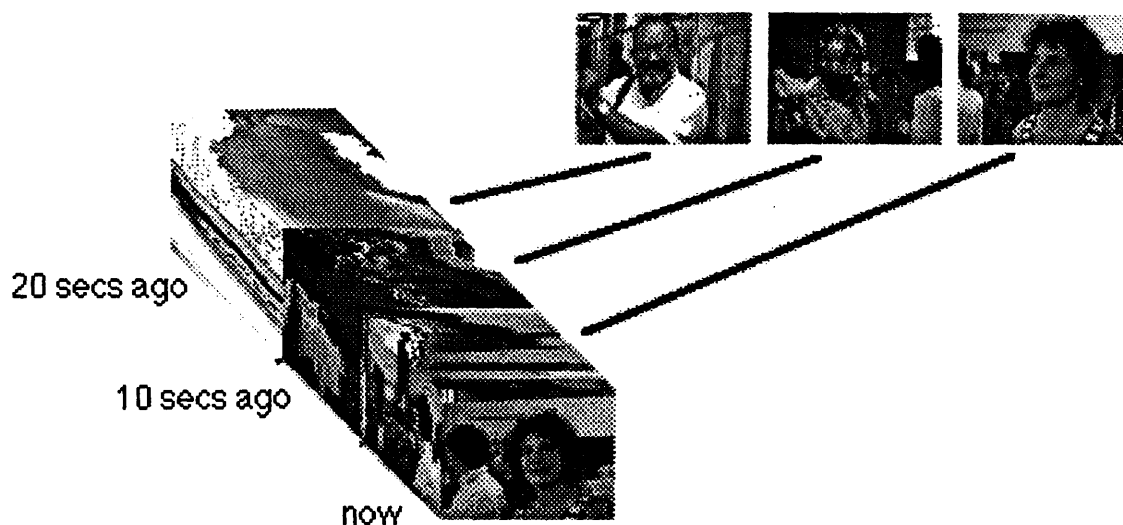
You cannot digitize perpetually, so you videotape perpetually instead. Then, when you see something you're interested in you wait until it runs on into something you don't want so you can stop recording and rewind the tape. You most likely don't have frame accurate computer control of the tape deck, so you have to start the digitizer and then roll the tape. Once your selected piece runs out, you can stop digitizing. Then you have to go back and trim the head and tail of the digitized shot. And so it goes.

The notion I have been pursuing with the streamer is that *by the time you see something you're interested in, it should already be available for manipulation*. The first step towards that is to buffer the video stream digitally. By buffering the stream digitally we avoid having to buffer on videotape and can eliminate some of the steps above.

3.3.2 Video Volumes: Time Will Tell

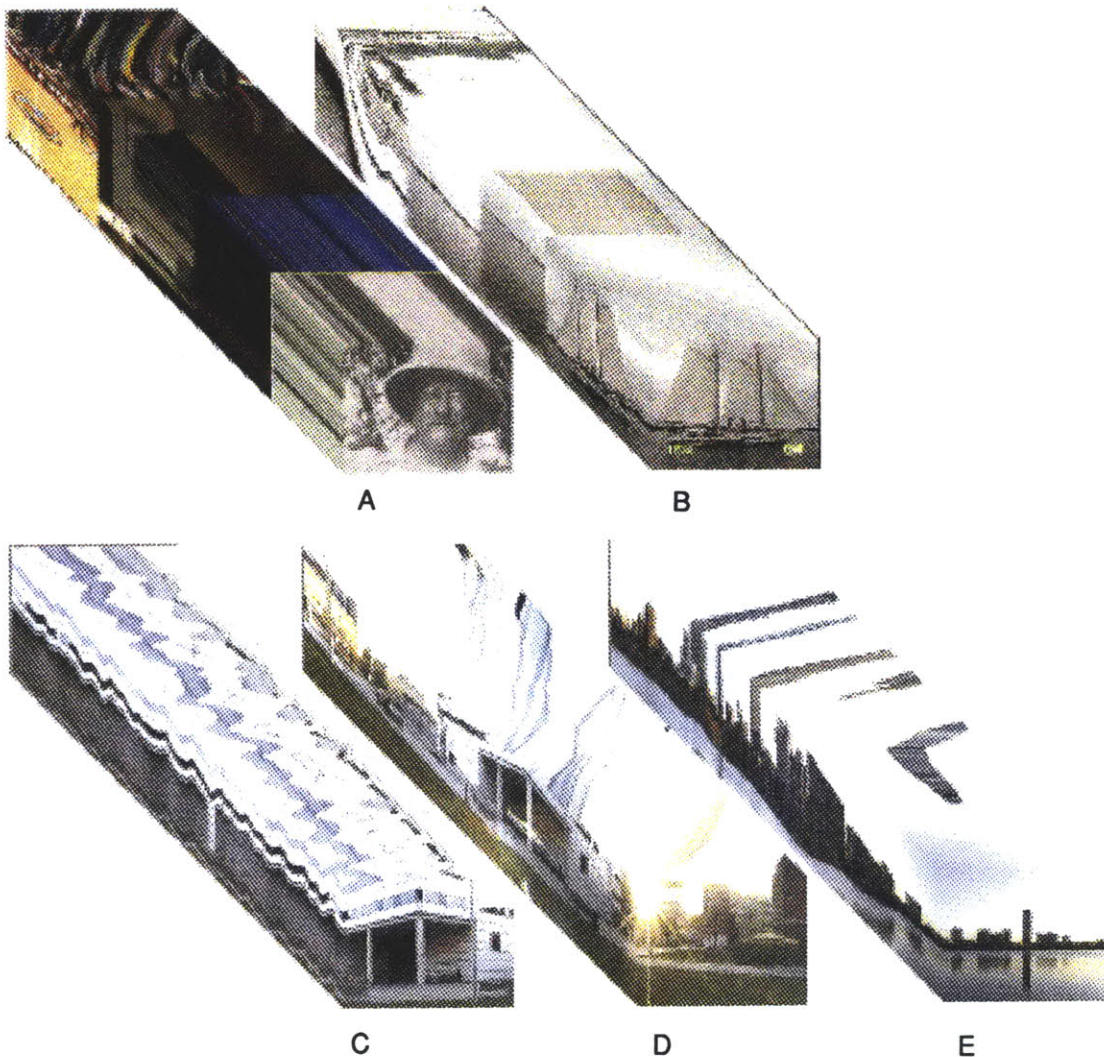
Digital video permits a variety of ways of presenting a video stream. Moving pictures are normally presented in a rapid succession of still frames that produce the illusion of motion. We might view this string of pictures as a stream flowing past us in time. When we stop the stream and step back from it to gain a perspective beyond a single frame we usually imagine strings of frames side by side, as though they were still on celluloid strips. The video streamer offers two alternatives to these types of views. First, it stacks the frames flip book style so that we can view more of the stream in the same space. Lining up the frames so that we see only their edges, we can see temporal characteristics that are perceived in the normal full-screen full-motion view but are not visible in film-strip views. Characteristics such as camera motions, transition types and locations, shot length, and shot rhythm stand out. Second, since the video stream is inherently dynamic, the extruded view also presents itself as a flowing stream. The edge patterns in the extruded view give clues to the

characteristics of the stream's contents, but they rarely indicate exactly what is found within the frame. As we view the stream flowing, the images are implanted in our minds. Then, when the stream is paused, its edges provide reminders to what we have seen, as well as providing a view of temporal attributes. Perhaps this reminder characteristic of the stream caters to our spatial memory. It's sometimes easier to remember where something is than when it is.²⁶



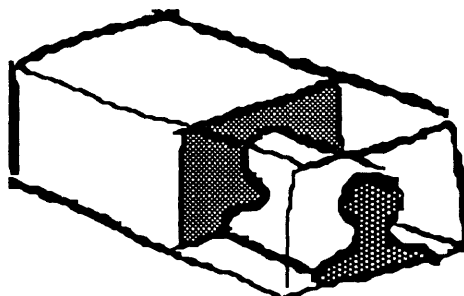
There are numerous temporal characteristics visible on the sides of the video stream. Transitions pop out and paint a picture of editing rhythm. Even basic types of transitions are visible: a cut is marked by a sharp ending of one set of streaks and the beginning of another, while dissolves show up as blending transitions. The sides of the streamer also indicate camera motions: flaring streaks indicate zooms; diagonal streaks and swishes reflect pans and tilts; shaky shots leave squiggly streaks; and straight continuous streaks usually indicate a locked down camera, often for a talking head shot. One learns to “read” these patterns after working with the streamer for a short while. Watching the original video leave its telltale streaks is the fastest way to learn what the various patterns can indicate. After a while, one learns to read even more subtle streaks. For example, I’ve noticed that three streaks, two white and one colored streak in the middle, appear occasionally on the bottoms of streams. The two white streaks are a white shirt and the colored streak is a tie, likely indicating a male talking head shot.

²⁶ Just an aside: Thinking about how memories are displaced evokes some questions about cinematic perception: How long do viewers remember the exact sequence of shots? How rapidly and in what fashion does the ordering deteriorate in the viewer’s mind? How much longer can viewers catalog content than they can list its sequence? How do these memories relate to content and its presentation?

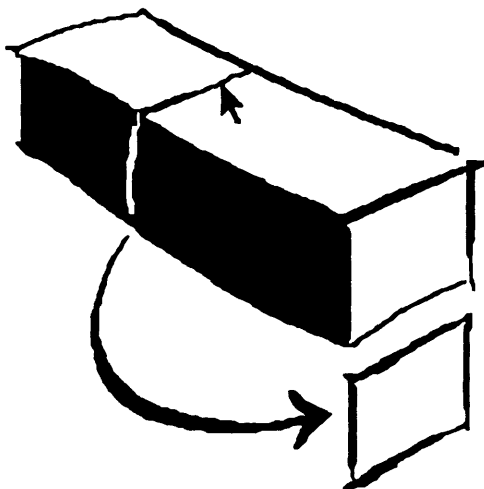


- A - Series of cuts.
- B - Series of four shots joined by dissolves. The camera tilts up in the first shot in the upper left. The last shot pans across a ship, leaving a slit-scan image of it on the side.
- C - Shaky shot of the Media Lab.
- D - Pan across the Media Lab.
- E - Pan across Boston, followed by a zoom out.

The current implementation of the streamer portrays the video stream as a solid opaque block. If the computer can separate foreground from background, or static elements from moving elements, we might also employ transparency in the rendering of the extrusion to selectively view certain elements. For example, we might render the background portion of each frame transparent so that we only extrude foreground elements. This might be a way of seeing the contents within a block.



As the video streamer stacks incoming frames, it flows from front to back with the farthest frames comprising the portion of the stream that is downstream in time. When the stream is paused, the extrusion invites poking at it to review its contents. As you move the pointer across the extrusion, the frame it touches is highlighted and presented in full in a viewer below. By “stroking” the extrusion from back to front, the user essentially replays a portion of the stream. Stroking front to back replays it in reverse. This provides an intuitive way of indexing the contents of the stream and to quickly jump from a fast-forward-like overview to a meticulous frame-by-frame review. It also relates the relatively longer term characteristics visible on the sides of the extrusion to the contents of the individual frames that comprise the extrusion.

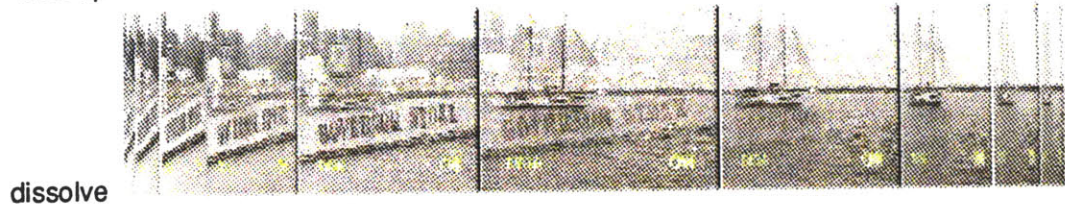
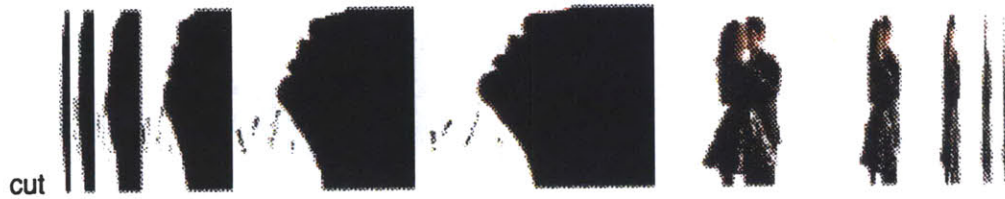
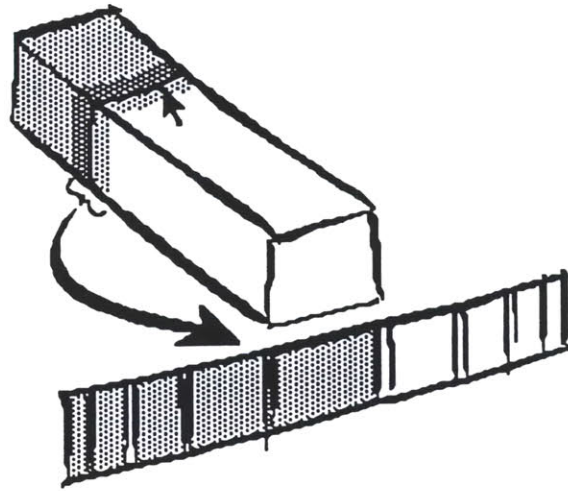


3.3.3 A Fish-Eye View

The video streamer shows characteristics across hundreds of frames in the stream. Sometimes it is useful to look at things on a more microscopic scale. The frame-viewer can also present a series of frames side by side, with the center frame representing a user selected moment in the video stream, the frame lying under the mouse pointer. The frames preceding and following this selected frame in the stream are presented to the left and right, but are more and more foreshortened in a fish-eye fashion the further they are in time from the central frame. This allows us to show a

longer string of frames within a given space. This expanded view extends about a quarter second on each side. The foreshortened frames in this anamorphic view have just enough resolution to allow us to spot a cut coming along. Transitions between shots show up especially well in this span.

Moving the cursor across the stream extrusion scrolls the frames in the frame-viewer. This dynamic relation helps connect the temporal characteristics visible in the extrusion with the instantaneous frames of the frame-viewer.



3.3.4 Sensible Sound

Stroking across the streamer plays back the audio stream along with the frames. The streamer captures audio into a sound buffer synchronized with the video buffer. Digital audio allows for more meaningful ways to play back sound than just at the speed and direction the video is played. The streamer always plays sound **forward** at **normal speed**, the most natural way for us to make sense of sound. Each video frame has an index into the sound buffer. As the user strokes across the paused stream, each time a new frame is displayed in the frame viewer, that frame also halts any audio currently playing and starts playing sound forward at normal speed from its index into the sound buffer. Stroking slowly forward in time produces a “Max Headroom” like stuttering of the sound playing back. If each frame were to only play a sound chunk equal to the frame’s duration, stroking slowly would produce gaps between the chunks. At twenty frames per second, these sound chunks are too short to be intelligible when there are gaps separating them. The stuttering has proven a worthwhile tradeoff for eliminating the gaps of silence. Also, given this method of playing sound indexed by video frames, stroking slowly in reverse has the effect of playing words forwards as the sentence comes out backwards. This method of playing back sound is especially useful when we select a chunk of video based on elements in the sound track, such as the beginning of a phrase or a beat in the music.

Who so desires the ocean makes light of streams.

-- Ahmad Ibn-Al-Husayn Al-Mutanabbi, Syrian (915-965)

4 THE SHOT PARSER

The streamer makes it easy to recognize shot boundaries at a glance, but shots still have to be manually designated before they can be saved for work elsewhere. To select a series of frames in the streamer to be viewed as a clip in the collage, the user drags the pointer across the streamer while holding the mouse button down, like selecting a run of text in a word processor. This sets begin and end points for the selection. The streamer also has a shot parser that partitions the video stream into a string of shots, automatically defining the begin and end points. Given machine tagged shot boundaries, the user simply taps on a shot to select it. It's a small savings of effort for the user, but it helps streamline the process of introducing source footage to the collage. The shot parser only recognizes cut transitions and it isn't perfectly accurate, but it works in real time as the streamer captures video. The current implementation of the shot parser also suggests some possibilities for more sophisticated parsers.

4.1 Why Segment into Shots?

Shots are fundamental cinematic elements. Shots are among the building blocks for editing, so they seem a natural element for deconstruction and recombination as well. There are other meaningful ways of segmenting a stream, by sequence or scene for example, and some streams cannot be segmented merely according to splices between takes (e.g., Hitchcock's film without edits, "Rope"). But shots have a good granularity for attaching a concise description to, and so it is useful to regard them as atomic elements of an image stream. In other words, shots are bite size chunks for

consuming video streams.²⁷ Shots are not the only useful units of meaning, just the ones the machine can efficiently detect right now. Section 4.4.2 will describe possibilities for other types of information the machine might eventually detect.

Compared to our amazing vision system, our eyes and our mind, computer “vision” is extremely rudimentary. Still, even basic signal processing offers one way for us to exploit the machine to help us understand a stream.

4.2 Related Work

Partitioning incoming video into separate shots is a useful first step towards organizing the stream. In 1990 I built a non-real-time shot parser for recognizing and marking gross scene changes that runs on a Macintosh. There is also a body of research concerning methods for parsing video. Russell Sasnett describes a shot parser in his master’s thesis, “Reconfigurable Video”.²⁸ Ueda, Miyutake, and Yoshizawa of Hitachi have pursued the problem of parsing a stream into separate shots as well as the problem of recognizing types of shots such as zooms and pans.²⁹ Tonomura and others at NTT have also made strides towards machine recognition of scene changes and camera motions.³⁰ Marc Davis, Ron MacNeil, and David Small, at the Media Lab, have implemented interfaces for helping a user quickly locate shot

²⁷ See “Cinematic Primitives for Multimedia”, by Davenport, Aguiere Smith, and Pincever, for more about granularity of meaning.

²⁸ Russell Sasnett’s scene detector recognized nearly 90% of valid scene changes in one piece of test footage (a documentary film), but false detections were also relatively high. “Reconfigurable Video”, pp. 64-72.

²⁹ Ueda, Hirotsada, Takafumi Miyutake, and Satoshi Yoshizawa, “Impact: An Interactive Natural-Motion-Picture Dedicated Multimedia Authoring System”, *CHI '91 Conference Proceedings*, pp. 343-350.

³⁰ Tonomura, Yoshinobu, “Video Handling Based on Structured Information for Hypermedia Systems”, *International Conference on Multimedia Information Systems '91*.

boundaries empirically and algorithmically.³¹ There has also been related research into parsing temporal information based on audio.³²

In some systems an alternative to implementing an algorithm for detecting scene changes is to monitor a video compressor's temporal compression factor. The QuickTime video compressors currently available on the Macintosh usually emit keyframes for changes more subtle than cuts between shots, so they have proven unusable as a gauge for shot boundaries in the video streamer.

4.3 The Algorithm

The shot parser attempts to slice the video stream at shot boundaries, saving the viewer the tedium of finding precise beginnings and endings of shots. It compares successive frames, or series of frames, watching for gross changes in the signal, tagging suspected shot boundaries. The shot parser currently works best at recognizing cuts between shots. A more elaborate video stream parser could conceivably watch for signal characteristics that signify other types of transitions, such as dissolves, fades, wipes, and keys. It might be tuned to recognize camera motions as well, say pans and zooms.³³

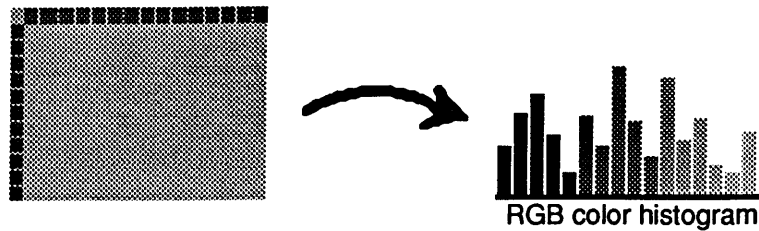
Each shot in a video stream has its own signal signature. Just like a written signature, a shot's signature has various characteristics that we can analyze to distinguish it from other shots. The shot parser works in tandem with the streamer as a video capture utility, so there are a number of aspects of the algorithm that have been intentionally kept simple to minimize the CPU load for parsing, maximizing the capture frame rate. In other words, we are looking at an especially blatant characteristic of the video signal's signature that costs little CPU time to analyze.

³¹ Marc Davis's "media timeline" and Ron MacNeil's "videographs" present slit-scan views of the video stream, similar to the sides of the video streamer, so the user can quickly spot transitions. David Small has also recently developed a shot parser that analyzes color composition to automatically detect scene changes.

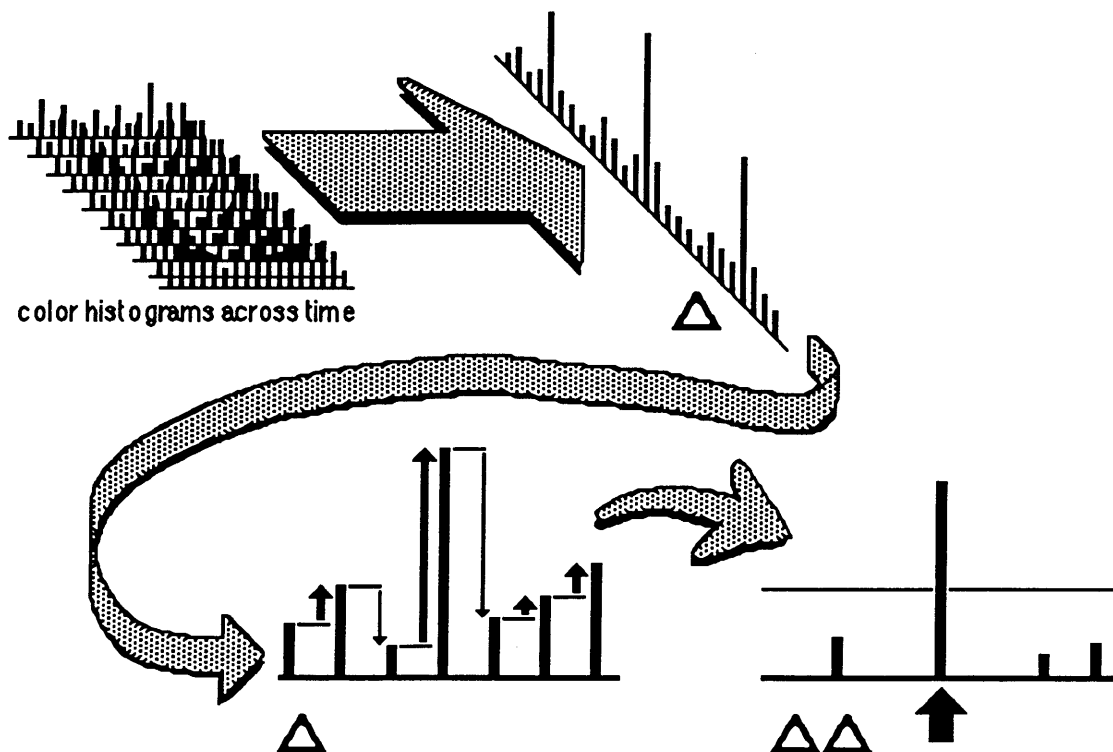
³² Natalio Pincever's master's thesis, "If You Could See What I Hear", proposed a method for analyzing the audio portion of video material to detect a variety of characteristics, including scene changes.

³³ Laura Teodosio's and John Wang's methods for correlating frames spatially can also be used to detect pans and zooms. "Salient Stills".

The shot parser samples each frame along its edges for RGB values and builds a color frequency histogram for that frame. It samples only along the top and left edges of the picture. It needs at least one vertical and one horizontal sample region in order to eliminate the possibility of recognizing pans and tilts as scene changes. It samples at the edge since in most shots the background is relatively stable and appears at the edges, while much of the action appears within the frame. A more sophisticated sampler might convert to another color space, say HSV or YIQ, but the parser currently works solely with RGB frequencies in order to minimize processing.



The system compares the color histograms of a pair of frames and sums the absolute differences between the histograms at each color value. This summed difference may continually ride at a relatively high value for some footage, especially if the signal is noisy or there is a lot of camera motion. So, it takes the difference between subsequent difference values to arrive at a plot with fairly well defined spikes. When the spikes in this second differential rise above a "magic number" threshold the given frame is marked as a transition point.



The following is pseudo code for the parsing algorithm:

```
for each frame      /* compare it with its predecessor to see if it starts a new shot */
{
    /* Clear the color histogram. */
    for each bin, b, in the histogram
        colorHistogram[b] = 0

    /* Sample the pixels to find the frame's color set. For each pixel sampled, */
    /* increment the histogram bin that corresponds to that pixel's RGB value. */
    for each pixel, p, in top row
        colorHistogram[RGBval(p)] = colorHistogram[RGBval(p)] + 1
    for each pixel, p, in left column
        colorHistogram[RGBval(p)] = colorHistogram[RGBval(p)] + 1

    /* Now compare this histogram with the preceding frame's histogram */
    /* to get the difference value, delta. */
    delta = 0
    for each bin, b, in the histogram
        delta = delta + abs(colorHistogram[b] - prevColorHistogram[b])

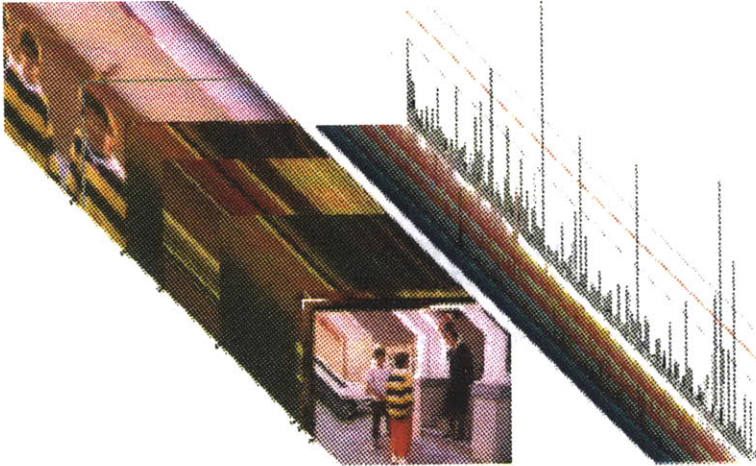
    /* Obtain the second difference to filter out some noise by setting deltaDelta to */
    /* the difference between the current delta and the delta derived earlier for */
    /* the previous frame. Note: only positive jumps between deltas will cause a */
    /* positive deltaDelta above our threshold. */
    deltaDelta = delta - prevDelta

    /* If deltaDelta is above the threshold, tag that frame as likely beginning */
    /* a new shot. */
    if deltaDelta > threshold
        tag this frame as a probable shot boundary!

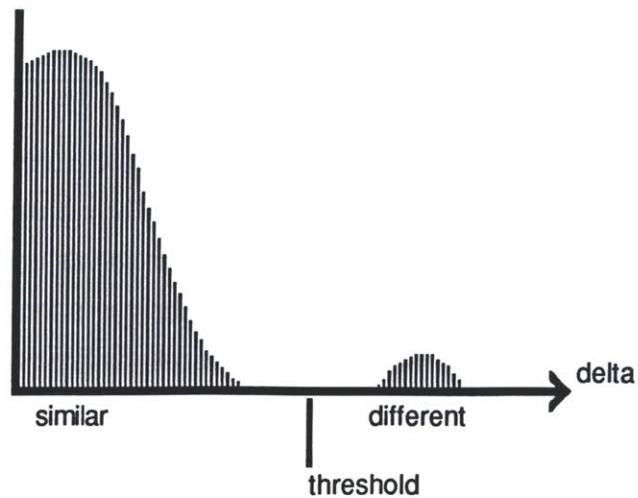
    /* Assign the current values to the "previous" values to be used next time */
    /* through this loop. */
    prevColorHistogram = colorHistogram
    prevDelta = delta
}
/* QED! */
```

The small sample and simple analysis allow the shot parser to maintain real-time capture simultaneously, but it does slip occasionally. It correctly ignores pans as potential scene changes, but it usually ignores dissolves and fades as well. Soap operas fool it frequently. They often have little change in the background, frequently cutting from one wood toned background to another. Black and white movies also fool the current system since their changes in luminance from shot to shot are much smaller than the combined changes in luminance and chrominance the shot parser recognizes in color pictures.

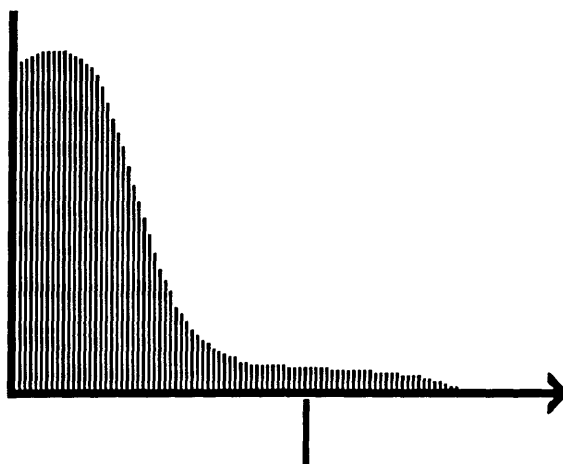
The shot parser can also display the histograms and resulting differentials flowing beside the video stream.



It's useful to imagine the shape of data produced by a perfectly accurate shot parser. A histogram of frame differences detected by a shot parser has bins for the delta values ranging from 0, for identical frames, to the maximum delta possible, for completely different frames. The value of each bin reflects the number of frames in a given stretch of time that differed from their preceding frame by that bin's delta value. A graph of this ideal histogram would have tall bins at the low delta values, reflecting the fact that the majority of frames in an image stream lay within shots rather than at shot boundaries, and are therefore quite similar to their neighboring frames. There would be a smaller mound of bins at higher delta values, representing shot boundaries detected. Ideally there would be a gap between these two mounds, where the threshold is. Delta values near the threshold represent more ambiguous frames. The frames with these values are not definitely similar to or definitely different from their neighboring frames.



This project's shot parser doesn't come close to this ideal graph. It has the mound on the left, but no gap. It merely attenuates towards the higher delta values. The parser occasionally misses scene changes and sometimes throws in false cut markers. It seems to be accurate maybe eighty percent of the time. But it does work in real time and it offers a usable selection most of the time.



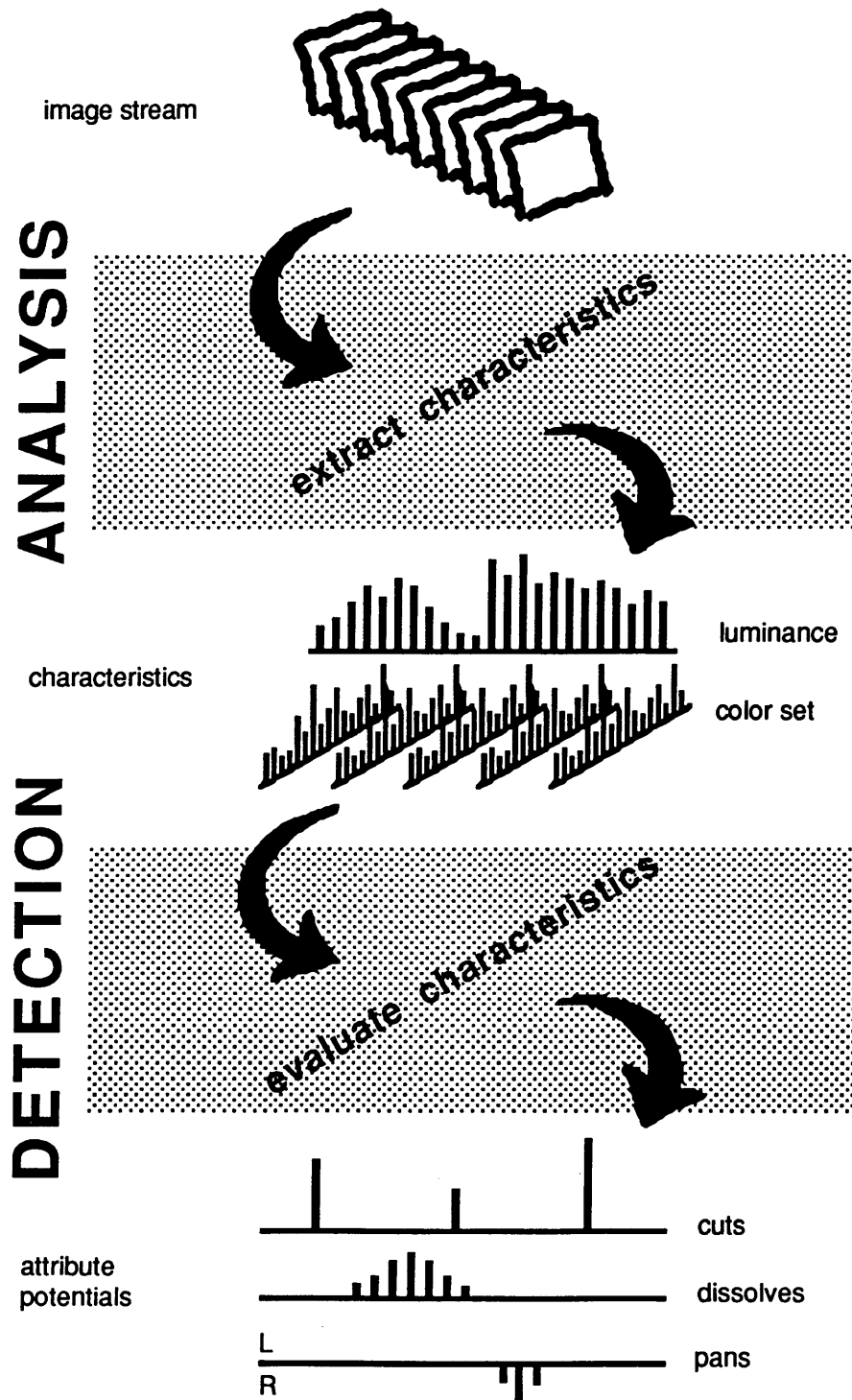
4.4 To Discern More

There are many types of transitions to recognize in a stream beyond just shot boundaries. Shot boundaries do give us a useful first slicing. The basic method of the current shot parser suggests a scheme for combining more elaborate analysis and interpretation algorithms in order to eventually tease more information from the stream with computer assistance.

4.4.1 Framework

The implementation of the shot parser, which currently only detects cuts and very fast dissolves, suggests a framework for more extensive parsing. This framework separates parsing into two phases, analysis and detection. Analysis extracts characteristics from the image stream's signal. Detection evaluates these characteristics across time to derive probabilities for various potential attributes of the image stream. Some examples of signal characteristics are luminance values, color set histograms, and noise levels. Some valuable propensities might be cut locations, location and rate of dissolves, or direction and rate of zooms and pans. The current parser only extracts the color histogram for a portion of each frame and derives the likelihood that that frame is at a transition between shots. Either phase may go through multiple iterations. The current shot parser first extracts the color histogram

characteristic for a frame, then compares it with the same characteristic from the previous frame to get a color histogram differential characteristic, then compares these differentials from frame to frame to get a second differential characteristic that is finally evaluated against a threshold for detecting probable cut frames. The second differential is a cheap way of filtering out some false cut detections caused by noise in the signal.



A very simple characteristic that comes for free in some systems is a given frame's temporal compression factor. If some means of temporal redundancy removal is employed, the compressor usually returns a compression factor for each frame it processes, reflecting coherency between it and its preceding frame.³⁴ Similar frames compress better, so bulges in the resulting stream of compressed data can indicate potential scene changes.

A more sophisticated parser that incorporated a variety of characteristics would need to employ some weighting algorithm to arrive at a difference value, or perhaps a collection of values that translated the various signal characteristics into propensities for content attributes (fade to black or pan left for example). This weighting may call for a mixture of some standard statistical analysis and some AI heuristics. The accompanying audio stream also may provide clues to the overall comparator.^{35,36}

Obviously, sample size and frequency impact the rest of the algorithm. Dissolves and fades usually slip through the parser because each pair of frames is only slightly different. The same algorithm might detect the neighborhood of a dissolve simply by reducing the sample frequency to maybe twice a second. Then, if no outright cuts were detected in that half second, we could tag that as a probable dissolve. For some characteristics it may even be useful to vary the sample region over time. Perhaps the nature of the sample region could be defined dynamically by the parser. For example, say the cut detector notices a series of ten frames that have definite, but below threshold, differences. It might wake up the dissolve detector at that point to check with its lower sample frequency if that stretch of frames comprises a dissolve.

The difference threshold has been derived empirically and is set relatively low in order to catch false hits rather than miss border-line differences. The threshold might also vary in a dynamic fashion determined by the analysis algorithm, as was suggested above for determining the sample region dynamically. Another method for

³⁴ MPEG and the QuickTime video codec employ temporal compression. JPEG does not.

³⁵ Natalio Pincever, "If You Could See What I Hear".

³⁶ Debby Hindus's master's thesis, "Semi-Structured Capture and Display of Telephone Conversations", concentrated on telephone conversations, segmenting an audio stream to add structure to it. It is easy to imagine applying some of her system's analysis to a video sound track as well.

deriving thresholds is to train the system. By feeding the algorithm frame pairs or sequences of frames that a viewer recognizes as similar or different, the system can obtain its values and adjust its interpretation of those values to correspond to the similar/different label provided by the trainer. Given a parsing algorithm, such a “training” session might yield a maximum delta value below which a pair of frames will be deemed similar and a minimum delta value above which frames will be deemed different. Hopefully the highest similarity value is below the lowest difference value. If so, then there is a range of ambiguity where we need to decide whether to include potentially false differences or discard slight differences. Otherwise, this range of ambiguity will extend into the high and low delta ranges the system has determined, rendering them not so definite.

4.4.2 Other Information to Detect

It is hard to gauge what leaps might be made in machine vision, but hopes are high. There are many vision capabilities that could be employed today for automatic annotation of image streams. The Media Lab’s Vision and Modeling group has successfully implemented algorithms for detecting when there is a person in a shot.³⁷ Taking this a step further, giving the machine a collection of mug shots, it can also sometimes detect who is in a shot. These are examples of more sophisticated machine recognition of content that goes far beyond the mere signal filtering of the shot parser that just happens to fortuitously relate signal to content.

More abstract cinematic information is also sometimes helpful. If the machine can recognize the shape of a person in a shot it seems trivial for it to also tag whether that shot is a close-up or a wide-shot. There is also a growing body of work that detects camera motions. If these algorithms become real-time, they might flow their insights along beside the video streamer.

There is a mountain of other information that could be quite useful to stream along beside pictures. Things such as audio to text transcription or simple descriptive annotations such as whether the shot is indoors or outdoors would have many uses, but are extremely ambitious for machine understanding and vision right now. Much of this information is logged manually in production. As digital video becomes more

³⁷ Matthew Turk, “Interactive-Time Vision: Face Recognition as a Visual Behavior”.

prevalent, hopefully more of this information will stick to the pictures and travel with them, sort of like how closed captions currently piggy back analog video. Until machine vision can derive some of these things for us, it would be helpful to have more of this information that is produced anyway continue on with the pictures and flow in time with them when appropriate.

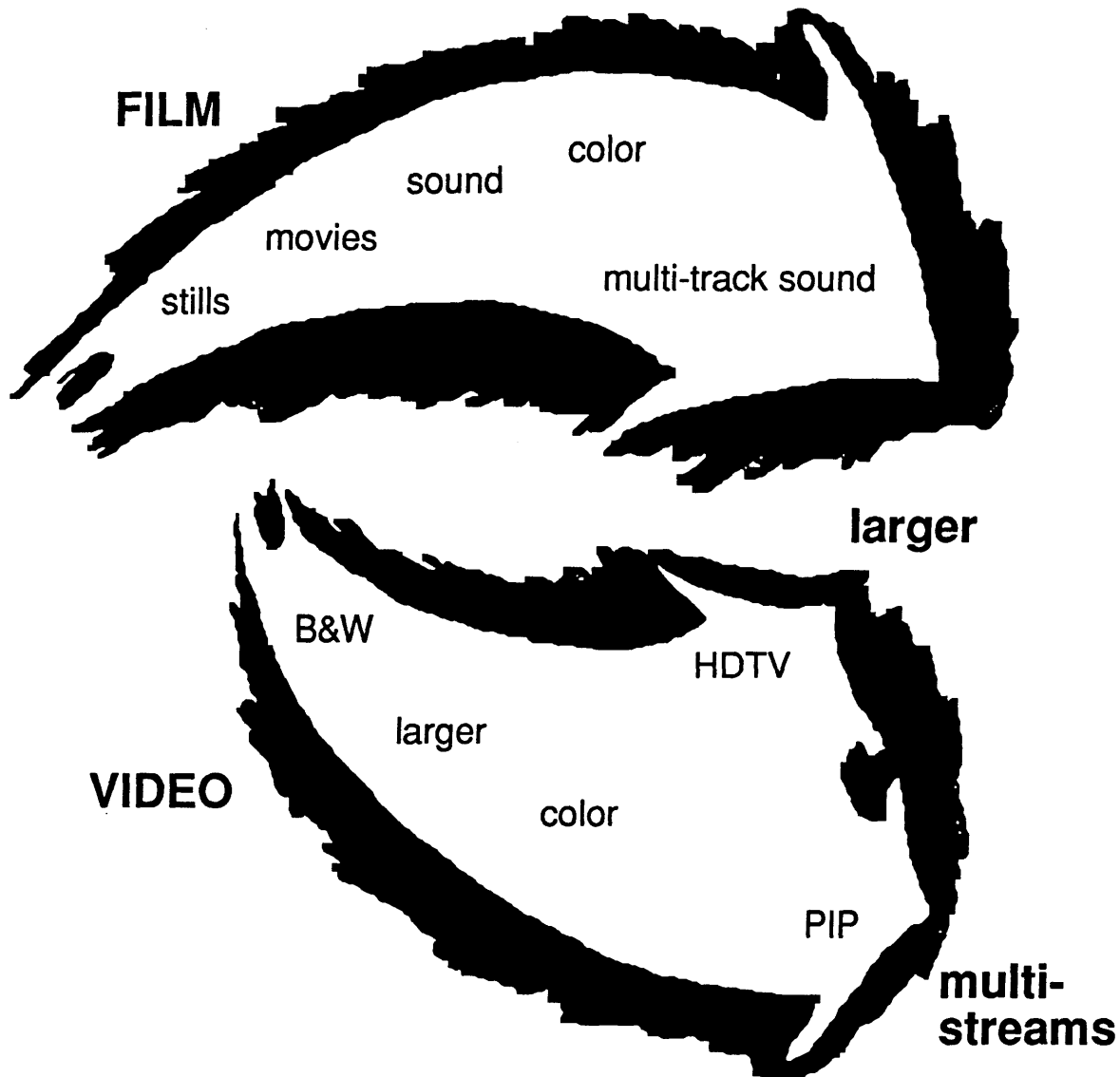
5 THE COLLAGE

The mind, we are told, can cut pieces from the cloth of memory, leaving the cloth itself unchanged. It can also make collages from memory material, by imagining centaurs or griffins.

-- Rudolf Arnheim



We are at a branch in the evolution of motion picture technology. Film and video have both progressed with ever increasing bandwidth. Some milestones are outlined in the figure below. The current push for HDTV marks the latest expansion of the video signal. New technologies providing more bandwidth promise yet larger and clearer pictures.³⁸ Increased bandwidth and a lean towards digital signals can also be exploited for multi-streamed presentation and viewing. The collage part of this project takes advantage of multi-stream possibilities to support thinking about the relations among a collection of images plucked from a stream.



³⁸ I couldn't pass the chance to quote Samuel Goldwyn here: "A wide screen just makes a bad film twice as bad."

This collage has been intended as a place to hold thoughts and mull them over while watching a source stream. I am more interested in how the process of collecting and organizing clips might affect viewing than I am in any particular configuration, linear or non-linear, resulting from the collage. However, regarding video editing, the collage may represent a workspace for a way of thinking not bound to the sequential nature of a final linear presentation. This may pertain more to early stages of editing, especially editing of unscripted material, where one becomes familiar with the material before focusing on the precise sequential organization. In terms of editing's production orientation, the collage might be seen as a mill where a collection of shots is spun from a web into a linear string.

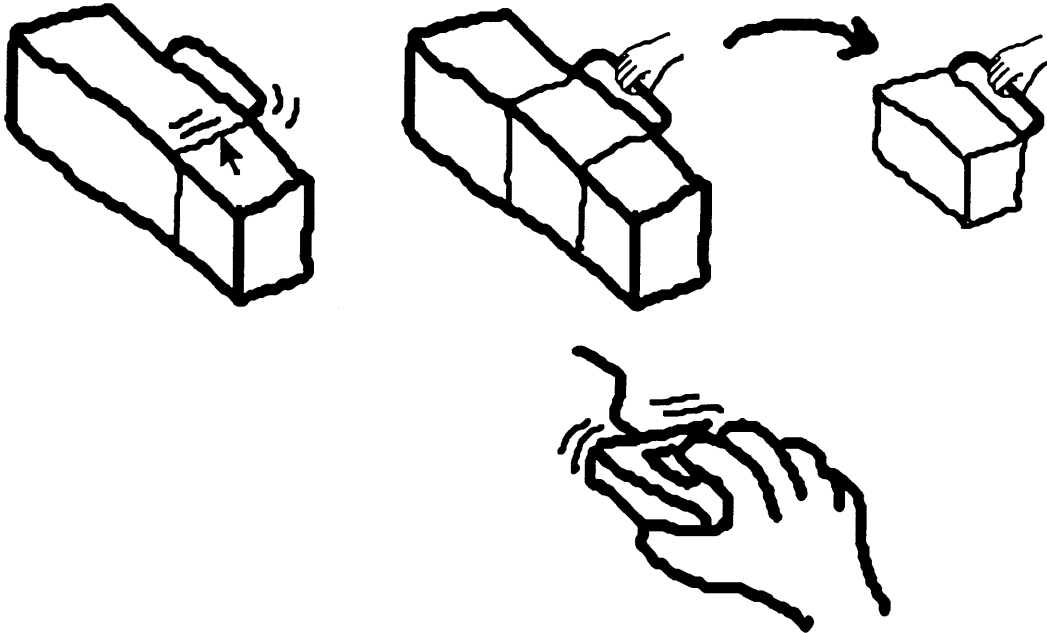
The collage implemented here is really a notepad for thoughts embodied in motion pictures. It offers a beginning tool set for managing collections of shots. I had originally intended a more extensive set of tools supporting the collage, especially for managing temporal relations among clips, but the streamer ended up consuming some of this development time. The collage features some simple windowing capabilities common to most GUI's: resize, reposition, open, and close. On top of that it also records a history of every change in layout so the user can refer back to earlier arrangements, earlier trains of thought.

5.1 Working in the Collage

The collage is implemented in HyperCard and uses the QuickTime XCMDs for accessing and controlling video clips.³⁹ When a chunk of video is selected in the streamer it can be dragged over to the collage. The streamer and HyperCard are separate applications, so when a chunk gets deposited outside the streamer's window, the streamer establishes a dialog with the collage through HyperCard by notifying it that the user would like to work with that chunk in the collage. The collage then responds with a name for the clip-to-be. On receiving the name, the streamer creates a QuickTime format file for the pictures and sound of the chunk and employs a QuickTime codec to compress the video frames. When it's done, it notifies the

³⁹ Working on the Macintosh, HyperCard provided a flexible environment for developing the collage. The QuickTime XCMD's (XCMD's are extensions to HyperCard's programming language, HyperTalk) were at a convenient level of versatility and complexity between the C API and the ultimate viewing environment for the user.

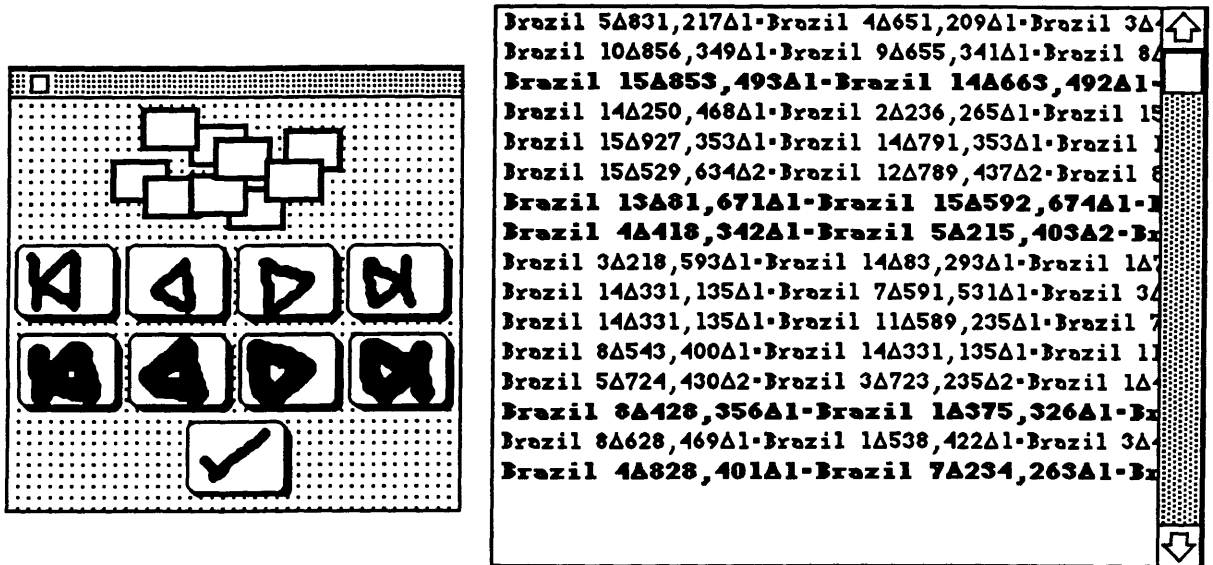
collage that the clip is completed and available. Then the collage opens the clip as another element in the collection. Voilà. The user merely drags the chunk from the streamer to the collage and then waits for the clip to be created. Right now this delay presents a slight interruption to fluidly working with the clip. As compression speeds improve, this interruption should fade away.



As clips appear in the collage, they play themselves out by looping endlessly. There is no synchronization among the various clips, although the user can pause, scan to, and play from any point in any clip. To arrange the clips the user simply resizes by dragging the corners of a clip and repositions by grabbing the clip in its middle and dragging to a new position. There are also miscellaneous other utilities for adjusting individual clip audio levels, retrieving from disk, and disposing.

Each rearrangement, resizing or repositioning, is automatically recorded in a layout history, a list. The user never actually deals directly with the list of text, just with the layouts it represents. The user can review and continue working from earlier layouts by scrolling through the list. There are four buttons provided on a layout palette for jumping to the first, for moving to the previous, for moving to the next, or for jumping to the last layout. In addition, the user can tag any layout as a key configuration by toggling on a highlight of that entry. As highlighted layouts become sprinkled throughout the layout history, the user can scroll among these through an additional row of four similar buttons for moving to the first, previous, next, and last

highlighted layouts. I chose to implement the list this way so that the user can't delete any entries in the layout history, but is still able to quickly jump over intermediate layouts. This keeps a record of how each collage evolves for my pondering later. This palette of buttons serves pretty well, but a visual browser probably would make the layout history far more useful.



HyperCard's graphics tools also provide a useful supplement to this tool set. The graphics tools can be used to create a backdrop for the collage of clips. Combined with the collage, they also hint at how one might use object oriented drawing tools that attach graphics to collage elements.

Recording and displaying many video streams, more than a couple, is especially difficult on a personal computer right now. Implementing a reconfigurable display of multiple streams on one screen requires some quality for quantity compromises. The first trade-off employed in this project was to sacrifice "hi-res" color for low-res black-and-white in order to slip past throughput bottlenecks. 1-bit and 2-bit monochrome moving images take up very little space to begin with, and can compress substantially better than color pictures. That's not too great of a sacrifice anyway. Cinema didn't wait for talkies and color before entertaining audiences. We don't have to wait for full motion, full screen, full color, digital video before exploring multi-stream possibilities.

5.2 What's Going On

I sketched a picture earlier of time as a river that carries events from our present to our past. Like a notebook, the collage is first a place to deposit clips rescued from that river. Saving clips helps to keep them fresh in the viewer's memory, and working with them helps to shed different light on them, and thus modify them. Also, by painting an overview of a body of footage, the collage provides a larger context to fit shots into.

5.2.1 Keeps Shots Fresh in Mind

The video streamer operates under the “out of sight, out of mind” principle. Once an image streams off the back end, it is gone for good. Fortunately, we have much better memories than the streamer's twenty seconds or so. But images in our mind do fade with time, or transform, especially when they are displaced by newer images. Dragging shots from the streamer into the collage keeps shots fresh in our mind by keeping them in front of us. In this way, the collage works as storage, as memory. Also, the very act of selecting and saving probably adds weight to a given shot, which helps keep it fresh in mind. It also affects viewing. Consider how selecting and saving shots might change how you watch, even if you don't ever reorganize saved clips.

5.2.2 Larger Context

Just as the streamer provides context for the snapshot view in the frame viewer, the collage offers a context for shots in the streamer. As the frame viewer shows an instant, the streamer shows a moment containing that instant, and the collage shows a collection of moments. This larger context sometimes exposes relationships that aren't so apparent in the tunnel vision of linear viewing.

5.2.3 Perpetual Motion

Looping clips beg to be better understood through repeated viewings. Non-linear storage media (RAM, magnetic disks, etc.) make it easy to loop video clips seamlessly. There is no need to rewind. Looping clips “stay alive”, or reinforce themselves through repetition. Repeating themselves, they also expose themselves to compounded interpretation. The clips in the collage play in loops in order to invite this sort of iterative viewing where more subtle elements become apparent on the third, the tenth, or the twentieth loop.

5.3 Collecting Thoughts Visually

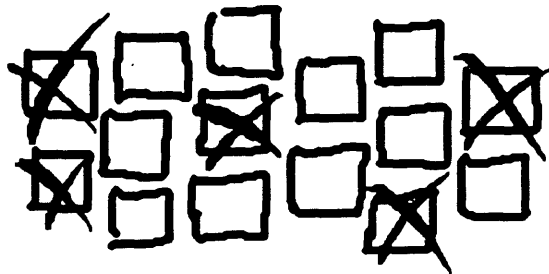
Well-connected meaning-structures let you turn ideas around in your mind, to consider alternatives and envision things from many perspectives until you find one that works. And that's what we mean by thinking!

-- Marvin Minsky

Time is linear. Videotape and film are linear. But do we think in a linear sequential fashion? It seems we've been stuck with the idiosyncrasies of motion picture technologies for so long that we often presume those characteristics to be natural traits of the medium, motion pictures, as well. Even though motion picture storage technology is becoming non-linear, we still commonly work within a linear array emphasizing sequence. The collage is intended to mimic the way our minds can hold multiple thoughts simultaneously without nice neat rows. Applying basic organizing methods (filtering and selecting, grouping, ordering) to moving images, other ways of viewing are enabled, or at least enhanced, in the collage.

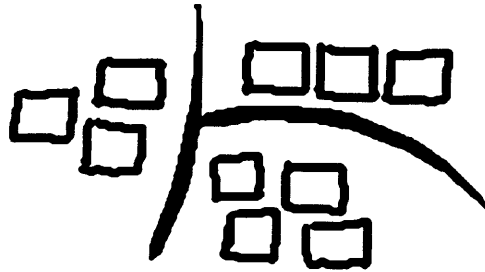
5.3.1 Filtering and Culling

Collecting clips in the collage is not much different from collecting and organizing anything else. The first way of organizing any collection is to decide what belongs, or at least what might belong. One might only admit elements to a collection which belong and then work with the collection from there. In this case the collection only ever holds elements that fit. Or one can gather everything even vaguely relevant and then filter from that collection what doesn't belong. The first method makes the decision up front, while the second only delays it. Either way, the resulting collection establishes a basic relation of association among its elements. For example, the collage might be a collection of hilights from a baseball game, all the examples of montage emphasizing the rhythm of a song in a music video, or perhaps a group of shots of each person appearing on a tape.



5.3.2 Grouping and Partitioning

Once we have a collection of something, how do we organize it further? One way is to partition it into smaller subgroups. This is really just a continuation of the original act of collecting, establishing structure within the collection by way of hierarchical groupings, only it doesn't involve so much filtering out. An example of subgrouping the baseball plays might be to separate the collection into great hits, scoring plays, and spectacular outs. We might think of this as categorizing elements by association, where the significance of a particular shot is indicated partly by the types of shots it is grouped with.



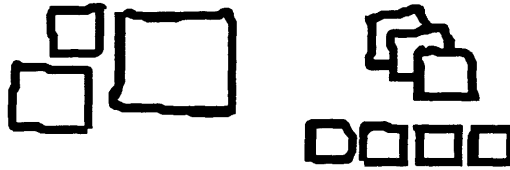
5.3.3 Ordering - Sequencing and Sorting/Ranking

Ordering is another means of building structure within a group. There are at least two senses of ordering, sequencing and scaling. Sequencing has to do with an arrangement where attention progresses from element to element in a certain order. A cause and effect relationship might be set up in a sequence. For example, a shot of the pitch, followed by a shot of the batter swinging, and ending with a shot tracking the fly ball to the outfielder's catch.

Scaling is another type of ordering. Shots collected together into a group may be further arranged by sorting them according to some criteria. An example: the shots of spectacular outs in the baseball game might be ordered according to controversy, with the most questionable call on the left and the most obvious call on the right. Scaling relates within a category by intensity.

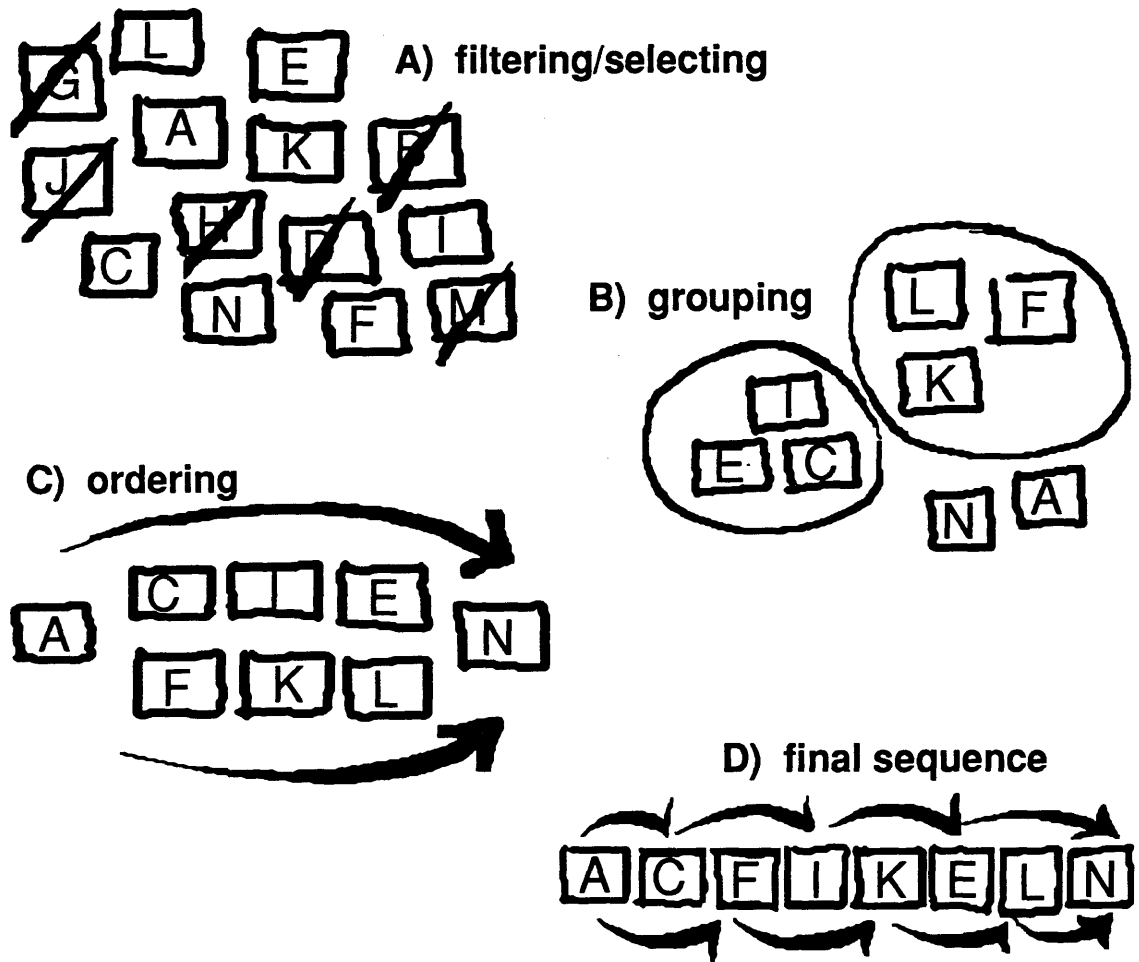
Both these types of ordering can be presented spatially. A storyboard is one example of a sequence of images. Continuing with the baseball example, the spectacular outs might be enlarged according to their degree of controversy, then sorted by size to reflect a continuum of controversy.

Additionally, a spatial display of clips aids unordering when that is useful.



5.3.4 Example: Drawing a String of Shots From a Collage

Capturing video in the streamer and transferring it to the collage is a transformation from a linear temporal arrangement to a non-linear spatial arrangement. An example going in the opposite direction, progressing from a collage to a sequence, might help to illustrate some of these non-linear ways of working with clips in the context of editing. A) Beginning with a random collection of clips on some subject we might review them and discard the unusable clips. B) Then we split them into subgroups that have to do with separate but related issues. C) After that, we order the clips to make a good mini-sequence in each group. We might also choose clip A as a good beginning shot, and clip N as a summary shot. D) Finally, we spin these into one main thread, crosscutting between the two sub-threads.



All of this also happens in conventional editing, only then the collage is in the mind of the editor. Experienced editors can juggle large collections of images in their mind. Perhaps a video collage might help novices to juggle as well.

5.3.5 Critical Mass

Observing a few people using the streamer and collage, the path to organizing thoughts seems to bounce around with no obvious strategy, but always involves culling and grouping, and sometimes continues to ordering. It is interesting though that each person had a different threshold for the number of clips they accumulated before giving thought to associations. Of course, you have to have a collection before you start arranging. But the number of clips at which people began to arrange varied among the small group of people who spent time with the collage from just a few to the collage's maximum of sixteen.

5.4 Critical Viewing - To See for Yourself

Again, Constructivist theory regards perceiving and thinking while viewing as an active meaning making process. But, as Sean Cubitt notes, watching video is not a visibly active process.⁴⁰ The video streamer and collage allow the externalization, or projection of the act of viewing, and thus make it visible. They at least lend some tangibility to visual thoughts, thereby aiding thinkers, and perhaps observers of thinkers, and definitely collaborators.

I don't suggest that the thinking that occurs while working with a collage of clips cannot happen in usual linear viewing. Yet, I believe that having more tangible thoughts/"objects-to-think-with" helps to induce a stronger sense of involvement.⁴¹ Pausing the flow of images, selecting chunks, and recombining them lends to critical thinking in ways different from just absorbing a continual flow does. It keeps the critical mind active.

⁴⁰ Sean Cubitt, *Timeshift: On Video Culture*, p. 4. Cubitt: "It's always easier to see a direct physical activity like domestic editing as an engagement in the processes of production. It is less in our cultural make up to recognise video viewing as a serious practise."

⁴¹ Seymour Papert defines "objects-to-think-with" in *Mindstorms* (p.11) as "objects in which there is an intersection of cultural presence, embedded knowledge, and the possibility for personal identification."

Turning to the specific case of watching television for a moment, I would like to suggest that *appropriation* of an idea is part of critical thinking, and that the streamer and collage can be used as tools for appropriating short elements from television's polished presentation and thinking about them in one's own context at one's own speed. Appropriation means taking something and putting it to one's own use. Critical thinking means challenging ideas for one's self. Critical viewing requires appropriation and tools for making that act tangible to reinforce the process. Sean Cubitt sees the act of appropriation alone as adding value to material: "The domestic video cassette recorder (VCR) is itself a kind of production device, as it can be used for seizing moments from TV's incessant flow, compiling, crash editing. ... However small my input, this kind of activity does seem to me to be adding value to the material which, unselected, remains just another blade of grass on the pampas of television output." ⁴²

5.5 Pictures and Words

Interactive Cinema has many examples of using machine processable representations and machine processable annotations of video clips to aid retrieval and for machine generated presentation.⁴³ How can we employ annotations in the collage for the viewer's benefit? So far, I've avoided that question by dismissing it as being outside the scope of this project. Machine processing of annotations is beyond my expertise and well beyond the time frame for this master's thesis. However, it is worth considering how annotations might eventually enhance the collage.

The streamer and collage are an environment for a viewer to collect and organize visual elements as a way to understand the content of those elements in as close to *real-viewing-time* as possible, *real-viewing-time* meaning at the pace the mind thinks while watching something. Making effective use of annotations in the collage will be challenging. The time and thought required to generate annotations is an investment.

⁴² Sean Cubitt, *Timeshift: On Video Culture*, p. 4.

⁴³ Amy Bruckman's "Electronic Scrapbook" offers assistance to the home video editor by taking textual annotations to clips and fitting them to story templates to give suggestions for sequences. Ryan Evans's and Mark Halliday's "Dynamic Movie Orchestrator" system uses sparse textual descriptions of footage and various cinematic models in the form of filters to weed out inappropriate material and sequence what survives. Benjamin Rubin's master's thesis, "Constraint-Based Cinematic Editing" explored generating narrative sequences in a similar manner.

That investment is worthwhile for editing, a non-real-viewing-time activity, but it may disrupt the dynamics of a once only viewing experience more than it is worth. The two cases where the collage sorely needs textual annotations though are for footage that relies more on its sound than on its visuals and for collections of similar looking footage. And just as the act alone of selecting and saving a piece of footage helps to emphasize it in the user's mind, the act of naming and describing a piece of footage can aid understanding of it, even if that description is never used again.

One plea for more machine processable annotations is so that the machine can assist the user in his or her task by processing footage through its annotations. But the "task" the collage supports is understanding through manipulation on the part of the viewer. The collage is intended as an environment for a viewing activity that encourages critical thinking. Aiming towards that goal, it is more of a system that enables the thinking viewer than one that assists. With annotations and some knowledge representation the machine might eventually assist the collage user by suggesting certain ways to organize the clips or maybe even by gleaning something from the user's arrangements that it can then use for better cooperating with the user. My main ambition has been to develop ways to foster critical thinking in the viewer. Fostering "thinking" in the computer is a much greater challenge.

6 EVALUATION

Such startling advances and cost reductions are occurring in microelectronics that we believe future systems will not be characterized by their memory size or processing speed. Instead, the human interface will become the major measure, calibrated in very subjective units, so sensory and personalized that it will be evaluated by feelings and perceptions. Is it easy to use? Does it feel good? Is it pleasurable?

-- Nicholas Negroponte, 1979

The collage is nowhere near complete (Can a master's thesis really be complete anyway?). At this point it only serves to suggest some possibilities for playing with motion images in a non-linear fashion. When the collage began to reach a workable form I sought feedback regarding how both it and the streamer functioned. Early feedback came in the form of casual conversations and as questions during demos. Later, I invited a handful of friends to "test drive" the streamer and collage. Those sessions yielded insights I could not have imagined working alone. By the time the streamer was done and I was finishing work on the collage I had an opportunity to exhibit the streamer as an interactive installation in a couple of art exhibitions sponsored by Sony.⁴⁴ The rigors of making the streamer bullet proof for public poking and the responses I got from the thousands of people who attended the exhibitions were also an extremely valuable opportunity to evaluate how the streamer works.

⁴⁴ The Art Artist Audition '92 was an open competition sponsored by the Sony Music Entertainment Group/AD Project and was judged by a jury and the public. The Yokohama exhibition ran for seven days in July. Nine finalists gave presentations in Tokyo and exhibited again for three days in September.

6.1 Evaluation

6.1.2 Setup

To get specific feedback about how the tools function and broader evaluations of what it's like to watch something with these sorts of tools I invited friends to try the streamer and collage with footage of their own choosing. I introduced the tools to about ten people, but many got pulled away prematurely by other commitments. Six people made it to the point of working with the system and then discussing it. This small group definitely has specialized interests as they are comfortable working with computers and video, and they are already enthusiastic about potentials for digital video. My aim was not for comprehensive testing and evaluation, but for suggestions and thoughts from a few people working with the system first hand.

For each person, I spent about fifteen minutes introducing the tools, fifteen minutes more for them to "practice", and then they each worked with the system as long as they wanted. A couple of people spent as short as fifteen minutes streaming and collecting. Another person spent about four hours. Most gave it between a half hour and hour. Afterwards, I barraged each with a series of questions.

6.1.2 Feedback

Suggestions - This "quick and dirty" pilot study furnished many specific suggestions and requests. One request I managed to quickly implement was for a way to mark points in the video streamer as it is flowing. The marker is useful for tagging things that are not visible on the sides of the streamer, such as the beginning of a sentence or when somebody turns to face the camera, so you can make sure they don't stream off the far end. Many other suggestions will have to wait. Among them: micons in the collage for a more concise view; longer stretches of time in the streamer; a visual browser for the collage's layout history; a picture of the audio stream; voice and text annotations of clips in the collage.

Frequent Stops - The first thing I learned observing people use the streamer was just how much they want to stop the flow of video and chew on what they have. Perhaps their footage was densely packed. Perhaps the streamer and collage are still too slow to keep up with the flow of interesting content. Almost everybody insisted on pausing their source tape whenever they paused the streamer. I think they wanted to

make sure they got it all. Nobody chose to stream and collage television, a source you cannot pause, but my own experience streaming TV has been that I just grab what I can. Grabbing from programming that is especially rich in streamer sized bites, the television coverage of the Los Angeles riots for example, I soon become content to not get it all. Grabbing some is better than grabbing none.

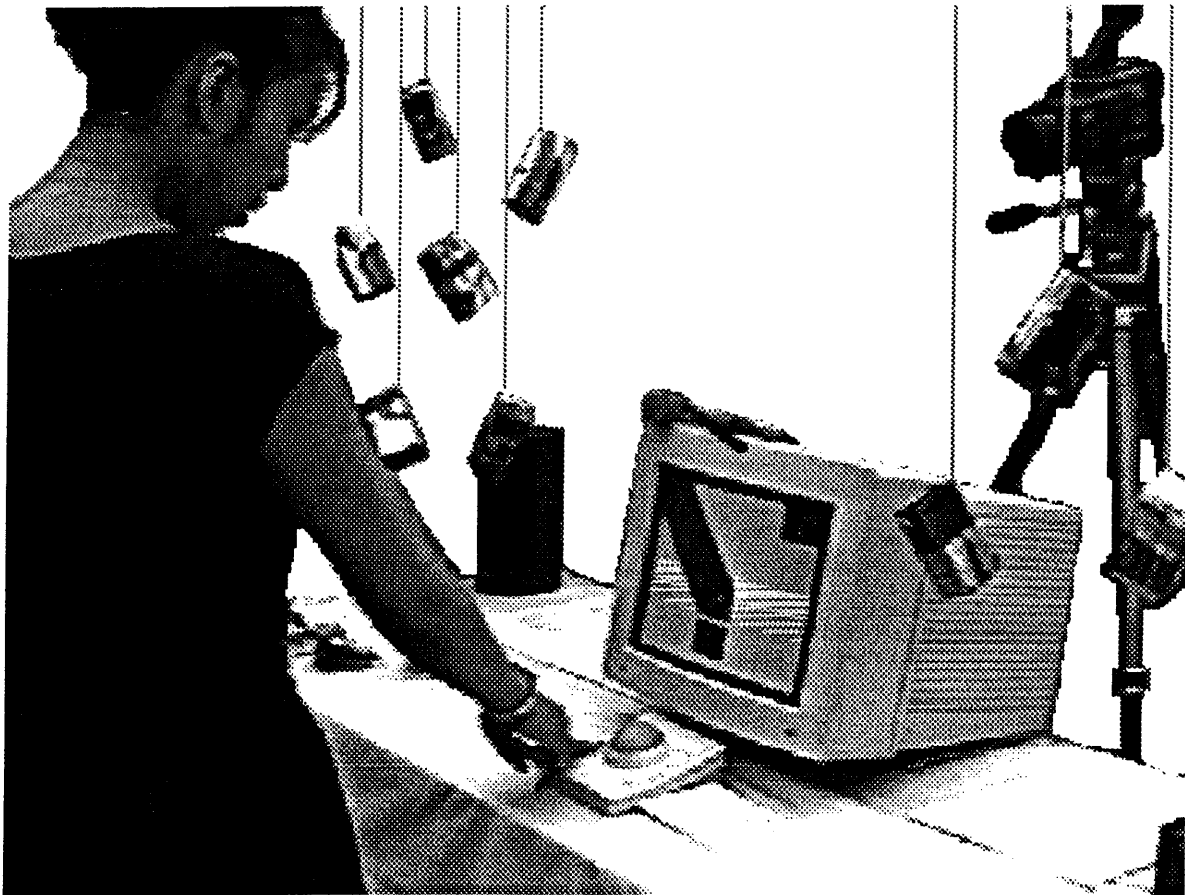
Longer Bites - Since the streamer only holds about twenty seconds, it works fine for a lot of broadcast video because of its relatively fast pace. However, even for broadcast video, twenty seconds is often not long enough to hold a complete chunk. A lot of personal footage begs for longer chunks. Even tightly edited commercial footage often wants more than a single shot to make up a meaningful chunk. The twenty seconds fell far short with speeches and interviews. It is often quite difficult to fit a complete and representative sound bite into that short of a time. It probably reflects an inattention to audio on my part, but the collage was also inadequate for footage of speakers. Having ten clips of the same person on one screen doesn't tell much visually about how ten different sound bites relate.

Need a Reason - Most people want some task besides viewing in order to warrant manipulating video. However, just plain diversion can be sufficient justification. Some people spent a long time with the streamer merely for the sake of playing with its perspective of time. Not many tinkered much with the collage. I'm not sure how the streamer's playfulness can be manifested in the collage, but the collage could stand to be more inviting. Perhaps people need more time working with the collage for possibilities to surface. It will also be interesting to invite people who are less experienced with video and computers to work with the collage. They might relate the collage less to conventional editing tasks and have a more open vision for other possibilities.

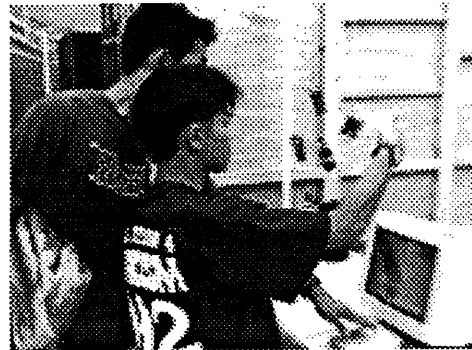
First Hand Experience - Everything, especially the streamer, made more sense to people when they used it on their own footage. Also, just playing with a live camcorder feeding the streamer went a long way towards helping people to read the telltale streaks on the sides of the stream. For example, if you simply tilt the camera yourself and simultaneously see the vertical streaks trailing down the left side of the streamer you will probably sooner recognize similar streaks in recorded footage as indicating a camera motion.

6.2 Streaming in Yokohama and Tokyo

The streamer installation at the Sony art exhibition was a chance for informal testing and evaluation of self-monitored “streaming”. The installation featured a scaled down version of the streamer where the visitors could pause and resume its flow and could flip through the paused video stream with a trackball, but could not select or save a chunk of footage. The installation did not include the collage. The exhibition site had poor television reception, so a camera pointed at the visitors provided a constant source of video. I hung streamer boxes, three dimensional soap bar sized printouts of captured video streams, about the monitor. These were intended to give clear examples of the types of things you can “read” on the sides of the video streamer and to subtly suggest video as something to hold. I also laid a few flip books on the table to reinforce the notions of motion images as a series of still frames and that stacking these frames forms a volume.



I learned a lot about designing installations for public spaces. I also learned a few things about the streamer, and more generally about getting people's attention and inviting them to poke at an interface. Using a live camera as a video source helped rope people in close enough for the streamer to grab their attention. Even if visitors never paused the stream, just swaying and waving in front of the camera to "paint" down the sides of the streamer not only engaged them as creative producers but also attracted the attention of others, allowing these others to watch and make sense of it before they took their turn at making goofy gestures.



There was a microphone next to the monitor for feeding the sound stream. It mostly went unnoticed. This may be partly due to the chaotic acoustics of the site. There were quite a few noisy exhibits. I think having a visual audio stream could have helped suggest playing with sound too. A few people among the week's six thousand visitors did record their voices and play with scratch audio and scratch video together in the streamer.

I hung the streamer boxes about partly for the artsy notion of holding time in your hand. I was surprised by how strong is the desire to learn things through touch. The most familiar part of the exhibit was the flip books. Some people began by

strumming through the flip books, then moving on to the streamer, flip book in hand. Some would explain the streamer to their friends first by grabbing a streamer box and describing it. Children seemed more prone to grab at the flip books and streamer boxes than adults were. One little boy, about four years old, was glued to the trackball. His hand never left it for forty-five minutes. He eventually got tired and slumped to the floor, but kept his hand on the trackball! How can we take advantage of this instinct to know things through touch as we develop systems for working with motion images?



7 DESIGN APPROACH

We have a book to write about the Gulf of California. We could do one of several things about its design. But we have decided to let it form itself: its boundaries a boat and a sea; its duration a six weeks' charter time; its subject everything we could see and think and even imagine; its limits---our own without reservation.

-- John Steinbeck ⁴⁵

Most of the preceding material in this thesis is hindsight. This thesis has been much more of a “do and then learn from it” experience than a “suppose and then prove it” endeavor. Steinbeck’s introduction above fits this thesis project well. Paraphrasing: This project’s domain has been tools for thinking with motion images; its duration about a year; (the rest is the same). Initially, the project was focused solely on the collage. But, as necessity gave birth to another invention, the streamer came to dominate this project. Eventually the collage didn’t get the attention I originally intended for it. It won’t get much more here either. This chapter takes a break from the *thinking with video* theme to document the genesis of the streamer as a way to think more about why the streamer works and how it came about. This is a small collection of notes charting how the streamer formed and describing some criteria for interface design. It doesn’t break new ground, but I think it is helpful to reinforce design guidelines by outlining the process and the criteria used.

⁴⁵ From Steinbeck’s introduction to *The Log from the Sea of Cortez*, in which he tells the story of a biological expedition he participated in in 1940.

7.1 Genesis of the Streamer

The following is a sparse chronology of how I went about developing the video streamer:

A) Buffer and Streamline

Capturing from a source stream into the collage needs to be as transparent as possible. The viewer's train of thought can easily be derailed by having to deal with setting up digitizing and then waiting for the system to capture the video. I began working on what eventually became the streamer just by trying to streamline the transfer from the source image to the collage. The way we normally go about introducing analog video to the digital domain excessively hinders fluid viewing. Normally, when you see something you want to work with you have to put your thoughts on hold while you digitize it. The first step towards the streamer was to have a digital wrap-around buffer of time that the analog source fills as you watch it (see section 3.3). So, I set aside lots of RAM and had the video digitizer loop through it with a continual feed.

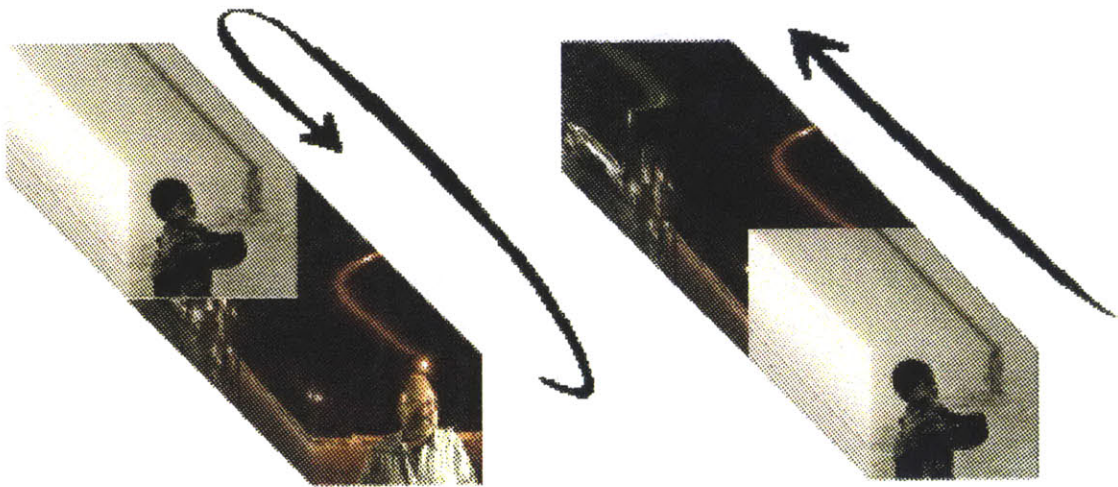
B) What does it feel like? What does it look like?

We always control video by turning a knob or pushing a button, with our hands. I sometimes wonder what a stream of motion images would feel like if we could touch it directly. Is video soft or hard, light or heavy, fluid or solid, etc.? This question of feel leads to a question of form. What shapes might motion images take?

Once I had a buffer for the incoming video source, I still had no notion of the streamer as it is now. My first step towards the streamer interface was to imagine what this buffer would look like if I could pop the lid off the computer and peak at the buffer in RAM, as if it were a physical object. I was probably also subconsciously primed about shape by flip books, by Vision and Modeling's XYT renderings, and even by some of my own ramblings with stacking frames of digital video a few years back (section 3.2). It just seemed natural that this video buffer was in the shape of a shoe box. It was obviously a good way to fit 250 video frames into the screen space for 25. When I actually stacked the frames up on-screen I was surprised to see that so much of time is visible in the streaks on the stacked edges, especially when you see the original motion picture create those streaks as it loads the buffer.

C) Going with the Flow

I originally had the digitizer paint a diagonal swath across the screen from the upper left to the lower right. When it got to the end, it would resume from the upper left, painting over the great 3D looking shape it had just streaked. That didn't look good. So I animated the buffer to flow out of the source image rather than sweeping the source image through the buffer the way it was doing. This probably came from some image in my mind of tape streaming past a record head, or perhaps an image of video feedback trailing older frames into the distance, but at the time it was an *aha!* sort of development.



D) Moving Images *Want* to Move

Once you have a 3D buffer full of pictures it's obvious that you merely stroke its edges to play back its contents, owing to flip books again. But when I paused the stream and didn't stroke it the streamer looked too static, more like a box of individual still pictures than a container of motion images. That led me to have the frame viewer continually roll through the buffer by default, allowing the motion images to stay in motion as often as possible. Users can always override to scroll through the buffer themselves, but if they don't, the pictures will play themselves back in their original motion, just as motion pictures want to do.

7.2 Design Approach

I outlined a variety of design challenges back in section 2.4. Now, reflecting back on how the streamer evolved, it's useful to summarize the design approach that helped to meet some of those challenges.

Does vs. Should - Let's begin with paragraph **B** above. The key question shaping the video buffer into the streamer was "What *does* it look like?" rather than "What *should* it look like?". Somehow this simple difference in how a question is posed makes a drastic difference in how free imagination can be. *Does* frees the imagination to find a solution, while *should* seems to impose constraints.

Let Elements Speak For Themselves, Direct Control - In paragraph **D** I imagined motion pictures to *want* to stay in motion. Of course they should, right? Again, the medium we are working with here is motion pictures. That medium should be the most prominent thing on the screen. The most immediate connection with that medium is through direct control. Granted, even the current streamer requires control via mouse, but after that the pictures are their own controllers. Minimizing buttons and sliders has kept the images visually prominent. Imagining an active nature for video has helped to imagine tighter controls.

"Advanced Image Processing" - The tools implemented for this thesis don't delve into sophisticated image processing or machine vision on the part of a computer. Hopefully, they do support more sophisticated image stream "processing" on the part of viewers. The machine has almost no sense of the signal's content. Its only inkling is the pacing of editing indicated by the shot parser. However, taking advantage of the malleability of digital images, we can use the machine to present alternative perspectives of things like camera motions and shot transitions as streaks on the sides of the streamer so that we can more clearly see and understand time dynamics. Partly to limit the scope of the project and partly to pursue better ways of rendering motion, this project has emphasized an interface where the only machine assistance has been to present images in such a way as to allow us to think differently about them.

RISC for User - Reduced Instruction Set Computing, RISC, is a paradigm for designing CPU architectures that speeds processing by eliminating extraneous specialized instructions that can be reimplemented through a combination of instructions from a small but efficient set. One of the design challenges in section 2.4 was to avoid creeping featurism. RISC and creeping featurism are concerned with

similar problems. I've tried to tend towards a small set of flexible but basic tools rather than a complex set of complex tools.

Interface - There are two meanings for interface in regards to cars. On the low end, the clutch, the accelerator, and the steering wheel are the interface to the car. Driving, using this car interface, is supposedly a means to mobility and independence, a more intangible sort of interface. "Computer interface" commonly refers to the means of interaction between a person and some content residing on a computer, interface as the "front end". On a higher level, interface can mean how a system induces or enables a person to think in a certain way. An awareness of both these meanings has helped while developing the streamer. On a low level, the direct controls mentioned above are the interface to the video buffer. The transparency of direct controls helps the user stay in the realm of the other interface, thinking about motion image content by manipulating it.

A concern for the interface here has been a concern for how ways of thinking are induced and supported in a viewer, rather than the interface just being windows and knobs to what's going on in the computer. For this thesis, the difference between these two notions of interface is a difference between regarding video as something to be acted on or as something to be acted through. The moral is to pay attention to both notions of interface, as the higher guides the lower.

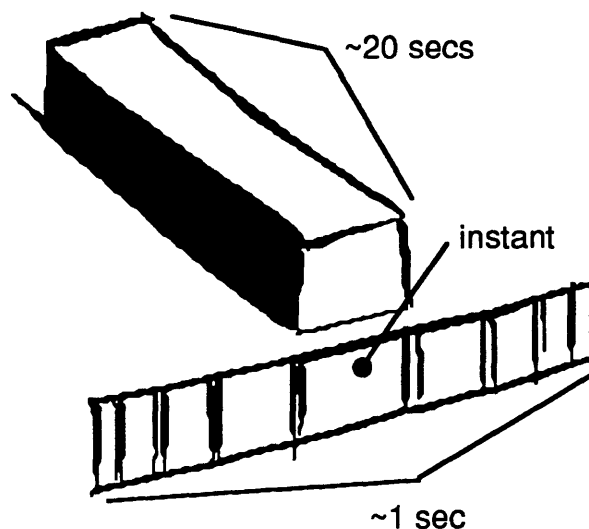
8 NEXT:

8.1 Continued Work

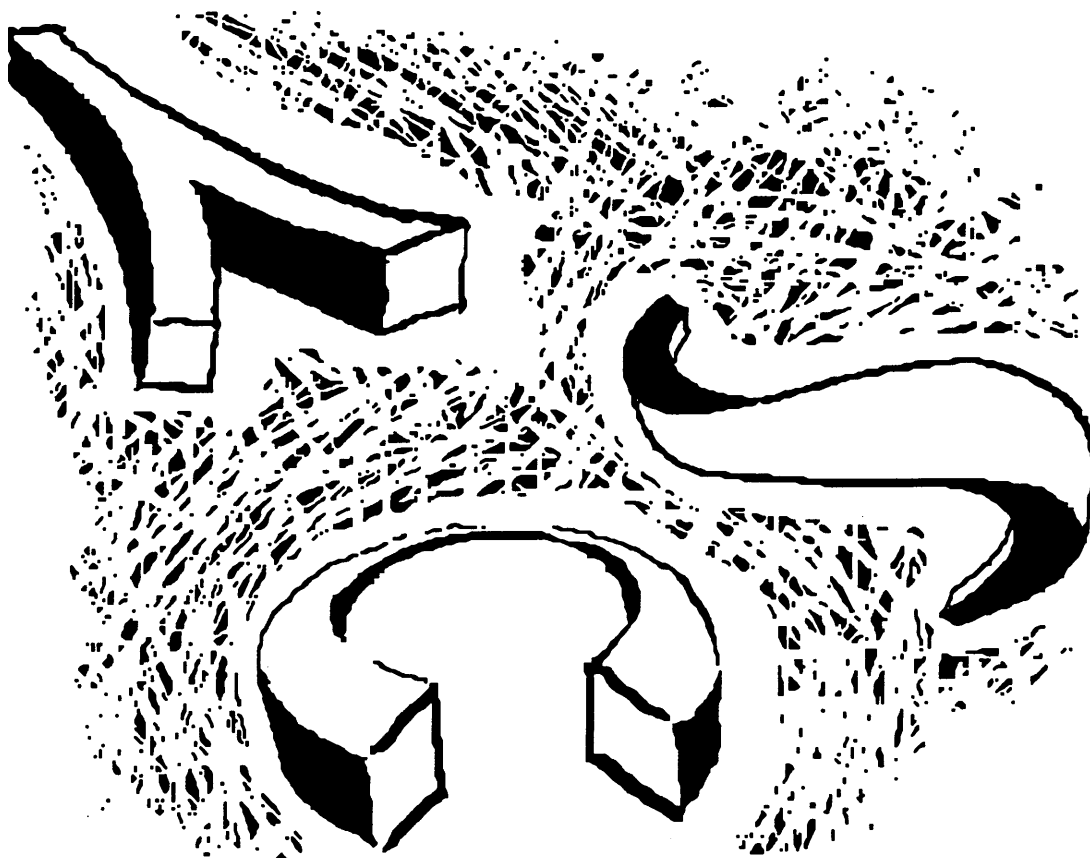
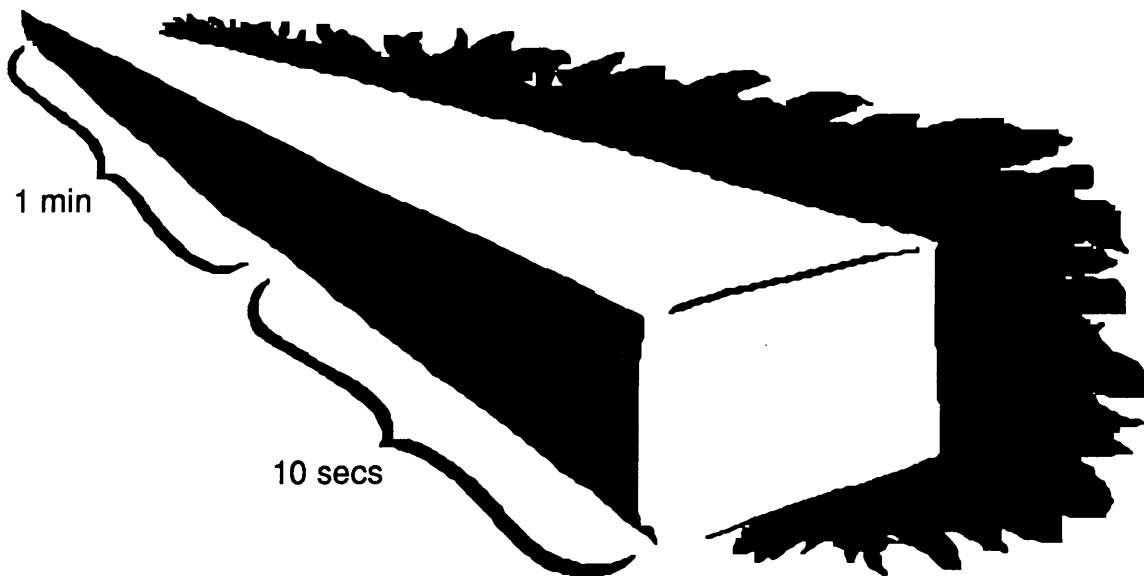
Neither the streamer nor the collage is complete. They both have open paths before them for continued work.

8.1.1 Stream On

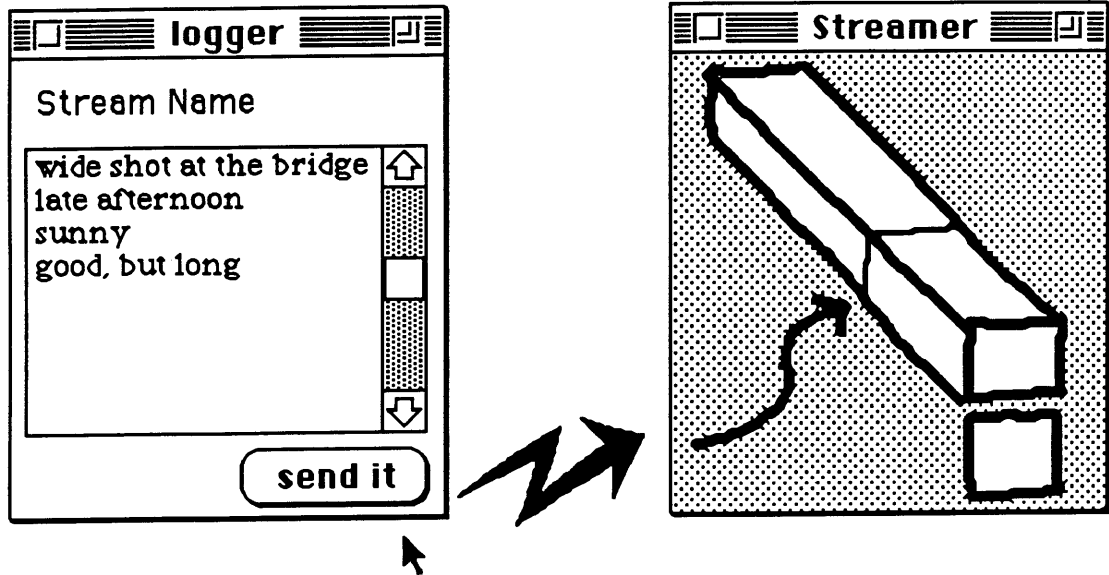
More Time - The streamer displays three different time scales: an approximately twenty second chunk that can hold a shot and sometimes strings of very short shots, the instant of time in a single frame, and about one second of context surrounding a frame. It needs to display larger time frames too, minutes or hours perhaps. And just as each time frame has a different set of characteristics associated to it (instant: framing of the shot; 1 sec: transitions; 20 secs: camera motions and shot lengths), there are different ways to render each time frame to make its characteristics clear.



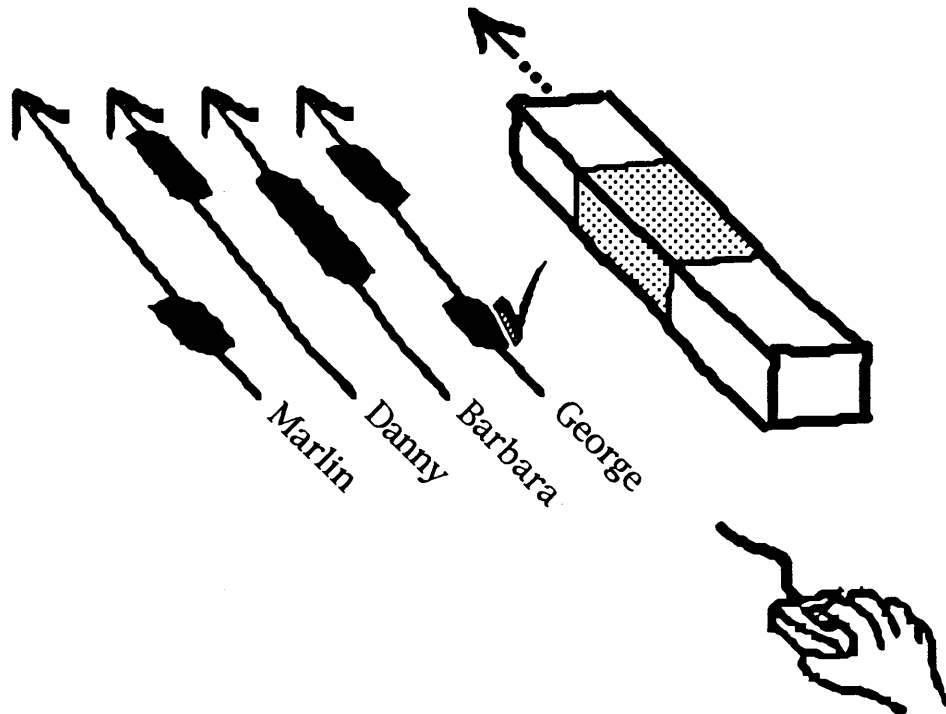
More shapes - One way of exploiting the streamer's 3D shape to show a larger chunk of time might be to render with perspective foreshortening. What other shapes would be useful?



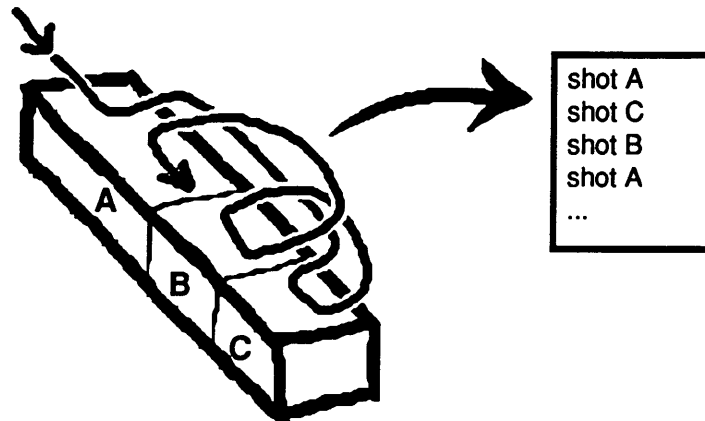
Annotation - Right now, the streamer serves both as a video capture utility and as a video store for reviewing video segments. I imagine two ways to extend it to handle annotations. First, it might receive textual annotations from another application and insert them into the stream at its current location.



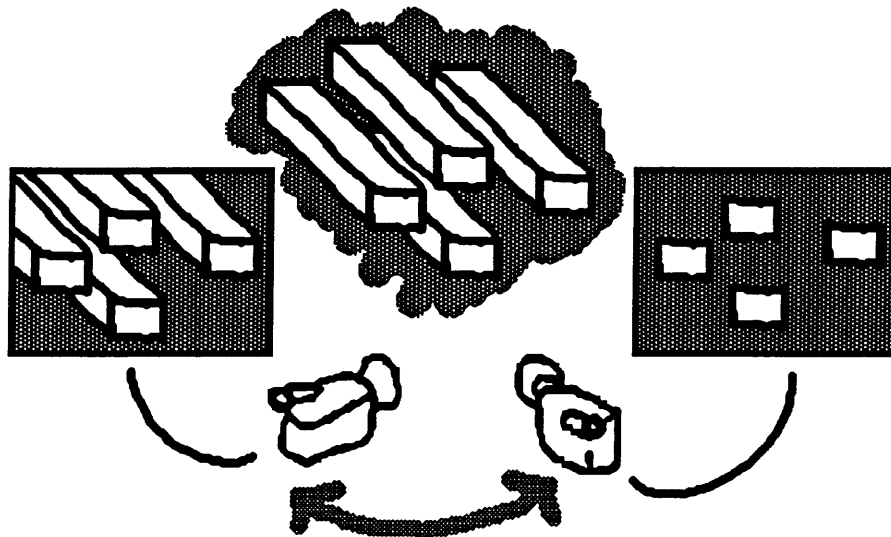
Second, it could have simple toggle tracks flowing beside the picture stream that someone logging the stream could quickly mark without pausing.



Scratch Editing - As one strokes a chunk of video in the streamer it plays its image back in the frame viewer. Stroking various chunks out of order essentially performs a quickie edit of the material. This could be useful for extremely quick *what-if* sequencing. Each stroke could be recorded in a list for a repeat playback or for refinement later.



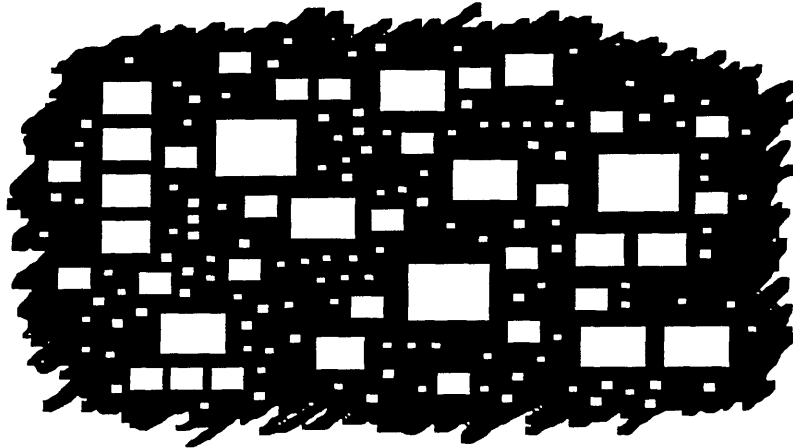
3D Video - One frequent request high on the list of possibilities is to be able to send footage back to the streamer from the collage. This suggests that it might be useful to view both the streamer and the collage in a 3D world so that sending video back and forth would merely be a matter of changing our view of video streams from on-end to a side view and back.



8.1.2 Collage+

More, More, More - Playback is still pretty jerky when the collage has more than a couple of clips playing at once. I think the sensation of watching and working with fifteen fluid clips will be a lot different from working with fifteen sporadic clips.

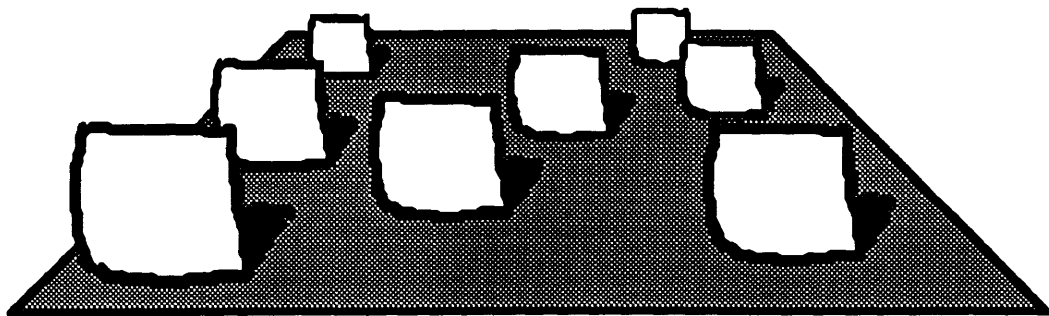
Watching fifteen clips play in the current collage is like watching for shooting stars, you never know where motion will appear next. My test pilots usually left most of their collage clips paused. I think the collage will appear to be a completely different beast when throughput improves to support twenty or more clips in full motion.



How many clips before we can make out constellations? Hundreds? Thousands?

Drawing Inferences - The collage lacks machine smarts. It has no way of entering or processing annotations. It doesn't even realize that when clips are near each other they might belong together as a group. Semantically loaded object oriented graphics tools might be useful for relating clips.

Other Playing Fields - The collage is a flat two dimensional page for collecting clips. Perhaps there are other spaces suitable for working with a collection of clips.



Relating Relations - Better transitions between collages might help to reveal relationships between different ways of organizing or thinking about the same collection of elements, especially how a particular element fits into different contexts of thinking. This may require more machine structure relating the elements in order for the machine to conduct a meaningful transition between collages. Perhaps an intermediate approach would be to have a collection of collages on screen. Would this aid in thinking about how collages relate in the same way a collage aids thinking about how clips relate?

8.2 Potential Applications

Potential applications for streaming video fall into at least three general realms: viewing, making, and analysis. Any particular activity may involve more than one of these. For example, editing involves both reviewing raw material and manipulating it to create more refined material.

In situations where one is viewing moving pictures, the video streamer and its accompanying tools can be useful for reviewing material and for getting a handle on things that change over time. Such situations are encountered when viewing hypermedia documents, while logging footage prior to editing or for entry into video databases.

Current video diagnostic and analysis tools are typically analog and look at a single frame at a time. Digital renderings of signal characteristics, such as the streaming color-histogram, can help to make temporal aspects of the signal more apparent. For example, slight horizontal shifts in video sync are quite noticeable in the picture when the video plays at normal speed, but examining the waveform or the video picture frame-by-frame it's difficult to spot the exact frame where the shift occurs. Buffering waveforms and stacking them in time could reveal the shift in the sync patterns at a glance. Also, given temporal renderings, displaying abstract signal data beside the picture stream helps to correlate the two.

8.2.1 In an Editing Toolbox

The most obvious application of the video streamer is in an editing environment. It might begin as a capture utility to feed non-linear edit systems. Eventually we should introduce editing into the streamer's three dimensional world. The streamer and collage are relevant to both novice and professional editing.

“Home” Editing - Few people grow their own video. Most of us consume our video preprocessed and packaged to sell. However, there is a growing population of video gardeners in the U.S. (more than 10 million people in the United States have camcorders), but most of their gardens wither on the vine as unedited and hardly viewed footage. Apparently shooting video is fun, but logging and editing are too dreary to make them worthwhile, or they require an expertise that only comes with years of dedicated experience.

Why do so many people shoot so much video yet never edit it? Some regard editing as a necessary evil on the path leading to a presentable story. Editing is considered undesirable because it requires extra equipment, because it takes years of experience before the mechanics of editing become second nature, and because story telling requires forethought and organization. It is hard. If we can look at *arranging* personal footage in a different light, maybe we can find ways to fit “editing” into the leisure time of home movies.

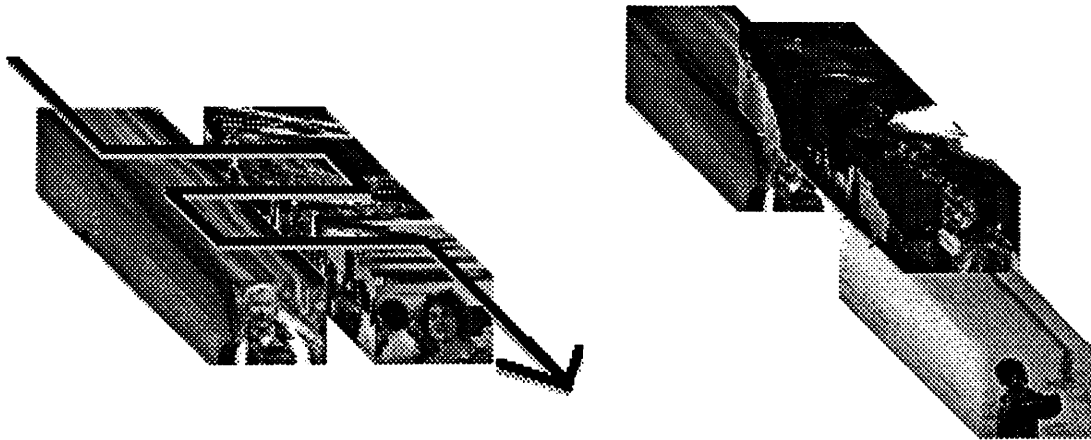
The collage suggests alternative forms for personal video. Glorianna Davenport sees the collage as a “video post-card”. The informal writing found on post-cards is usually a collection of concise and loosely connected statements. The collage mimics this style as an album for brief video clips that are loosely associated spatially. Constructing a collage is a type of editing that doesn’t depend on production planning and that doesn’t require a lot of experience in order to create even the simplest meaningful piece. Constructing a collage may be a way of arranging motion images better suited to personal footage than traditional editing is.

“Professional” Editing - Some new tools alter the basic nature of a process just by improving efficiency. That is the current state of non-linear video editing. Digital non-linear edit systems eliminate the mechanics of VTR’s -- we no longer have to lose our train of thought shuttling linear video tapes around -- and therefore make it a more efficient process by speeding up access to footage.⁴⁶ But most non-linear edit systems persist in only one view of video, in a linear timeline or array. They don’t allow for non-linear *thinking* as well as they could. The collage might cater well to the early stages of editing where a web of shots and segments could allow more creative thinking outside the confines of linearland.

A video streamer with annotation capabilities could also be useful in the early stages of editing for logging.

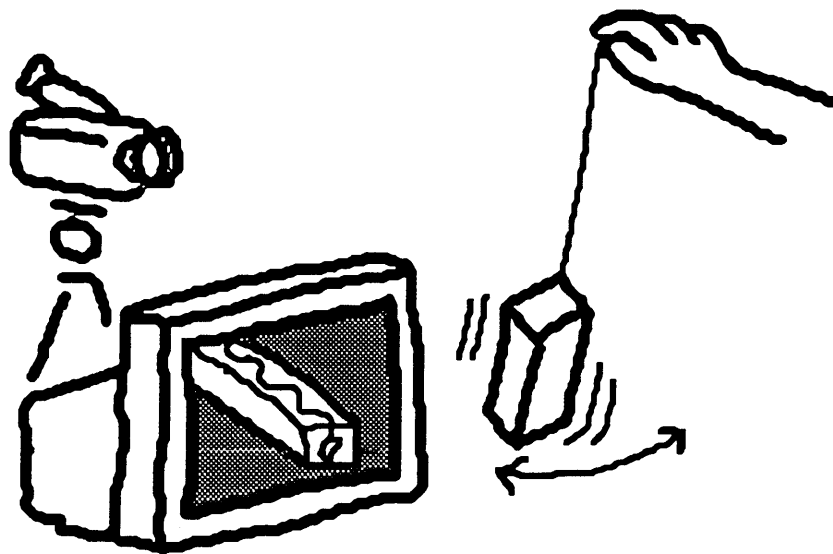
⁴⁶ However, non-linear edit sessions can run longer than conventional edit sessions because it is so easy to try more alternatives. Non-linear editors don’t have to commit to a preconceived plan for the excuse that making a minor change would take as long as re-editing the whole piece. So, while they improve the efficiency of the process in the creative sense, non-linear edit systems sometimes can actually lengthen edit sessions.

Ultimately, a hybrid of the collage and the streamer could offer new ways of editing in 3D-land. For example, we might line up two shots side by side and trace a thread through them to cross-cut between parallel actions. Or we can bridge two segments to indicate a segue.

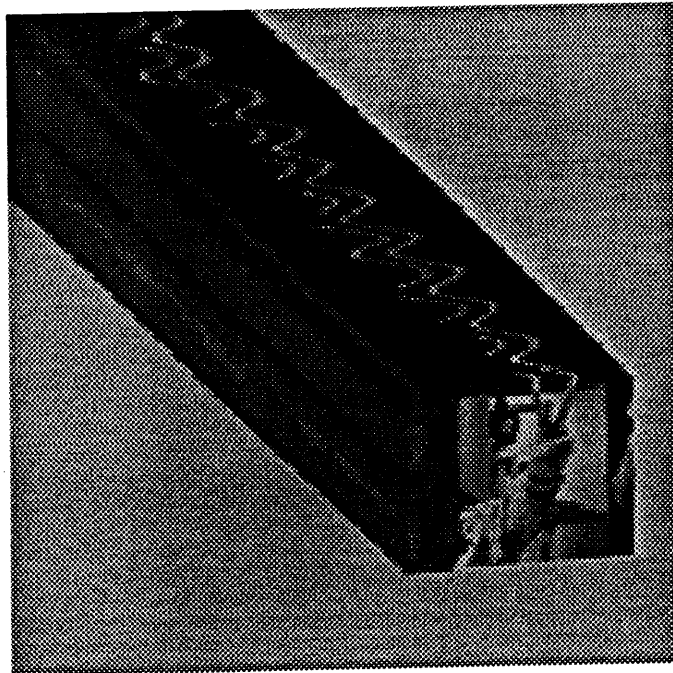


8.2.2 Interactive Installations

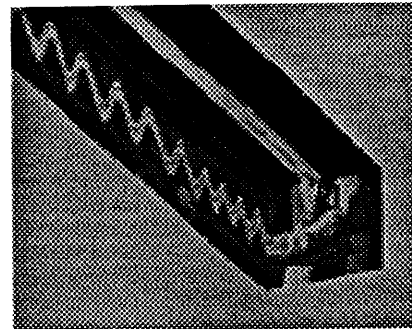
My ambition at the Sony Art Artist '92 exhibitions in Yokohama and Tokyo was to present the streamer as an occasion for people who do not ordinarily work with video to begin manipulating it in a playful way. Public settings like this offer the mutual benefits of exposing many people to tools and ways of thinking that are usually confined to professionals and of providing immense feedback regarding how easy the tools are grasped. Designing such an installation requires tools to be so obvious that anyone can use them and learn to use them quickly.



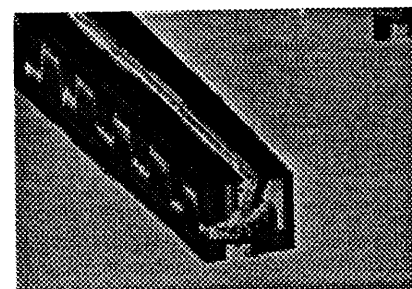
The streamer's portrayal of time might also find a home in a science museum as an exhibit that deals with patterns of motion. A number of people in Yokohama dangled the hanging streamer boxes in front of the camera to trail sine waves down the side of the streamer. Others experimented with drawing sine waves and square waves by waving their arms up and down at different speeds. One person jumped up and down in front of the camera and then analyzed his jump in the streamer ala Muybridge.



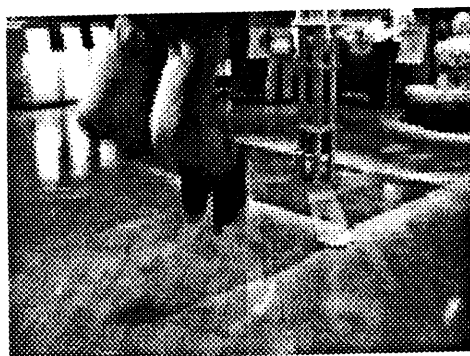
peace sign waves



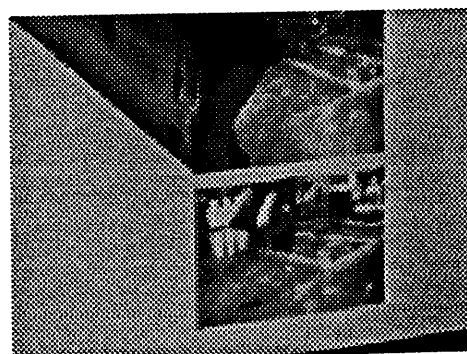
sine wave



square wave



jumping



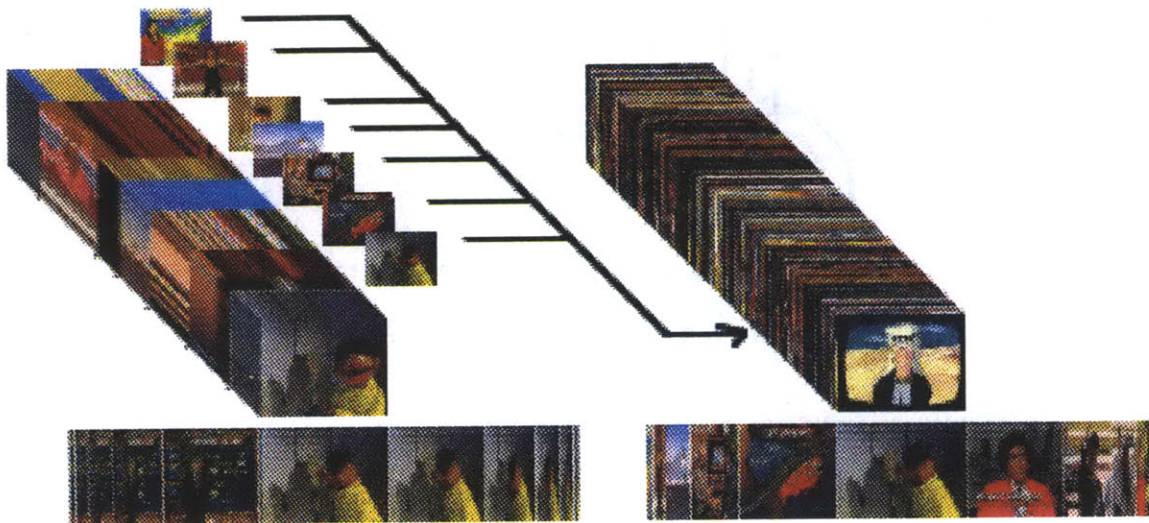
streamed jump

8.2.3 Video Scratch

This is a “just for the hell of it” application. Is there a visual analogy to rap music’s scratch?

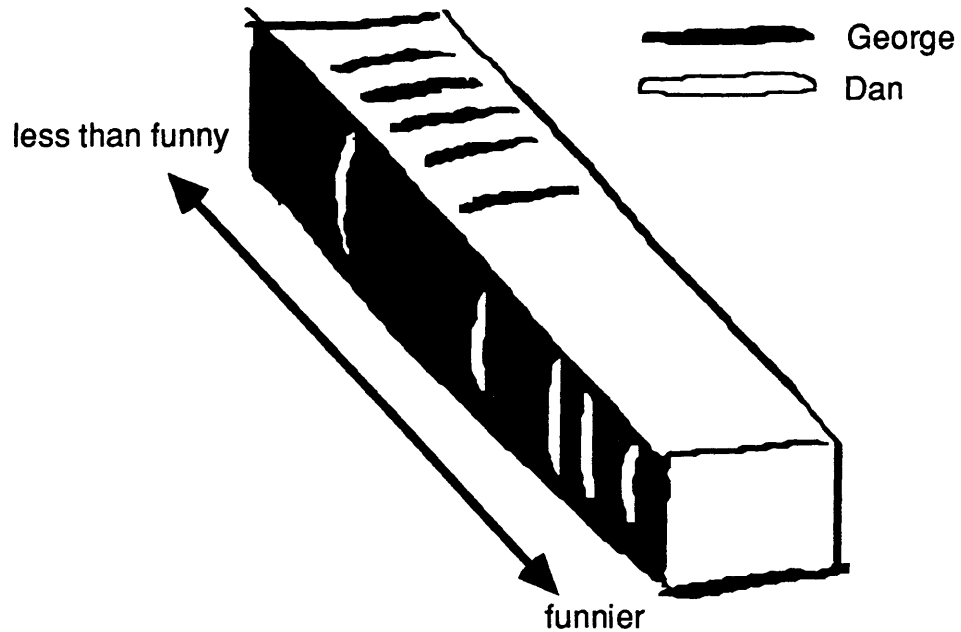
8.2.4 Browser

Abstracting a little, the streamer is a volume of sorted elements, like a library card catalog. The elements are video frames equally spaced in time, and sorted by time. I have taken a video clip of head-frames⁴⁷ selected by the parser during one pass through the streamer and fed that clip through the streamer in a second pass. In the second pass the streamer serves as a browser through discontinuous time. This is one way to see more time in the same space.



The streamer might also be used as a visual index to other things. Fed by a database query, the streamer could be used as a dense visual browser through elements sorted by how well they match some criteria. The visual elements in the streamer might refer to non-visual elements. Also, the edges could be color coded like tabs in a filing system by the retrieval software to indicate the distribution of secondary criteria.

⁴⁷ A head-frame is the first frame in a shot. The shot parser can save head-frames when it detects shot boundaries, forming a storyboard-like video clip.



8.2.5 Interactive Cinema

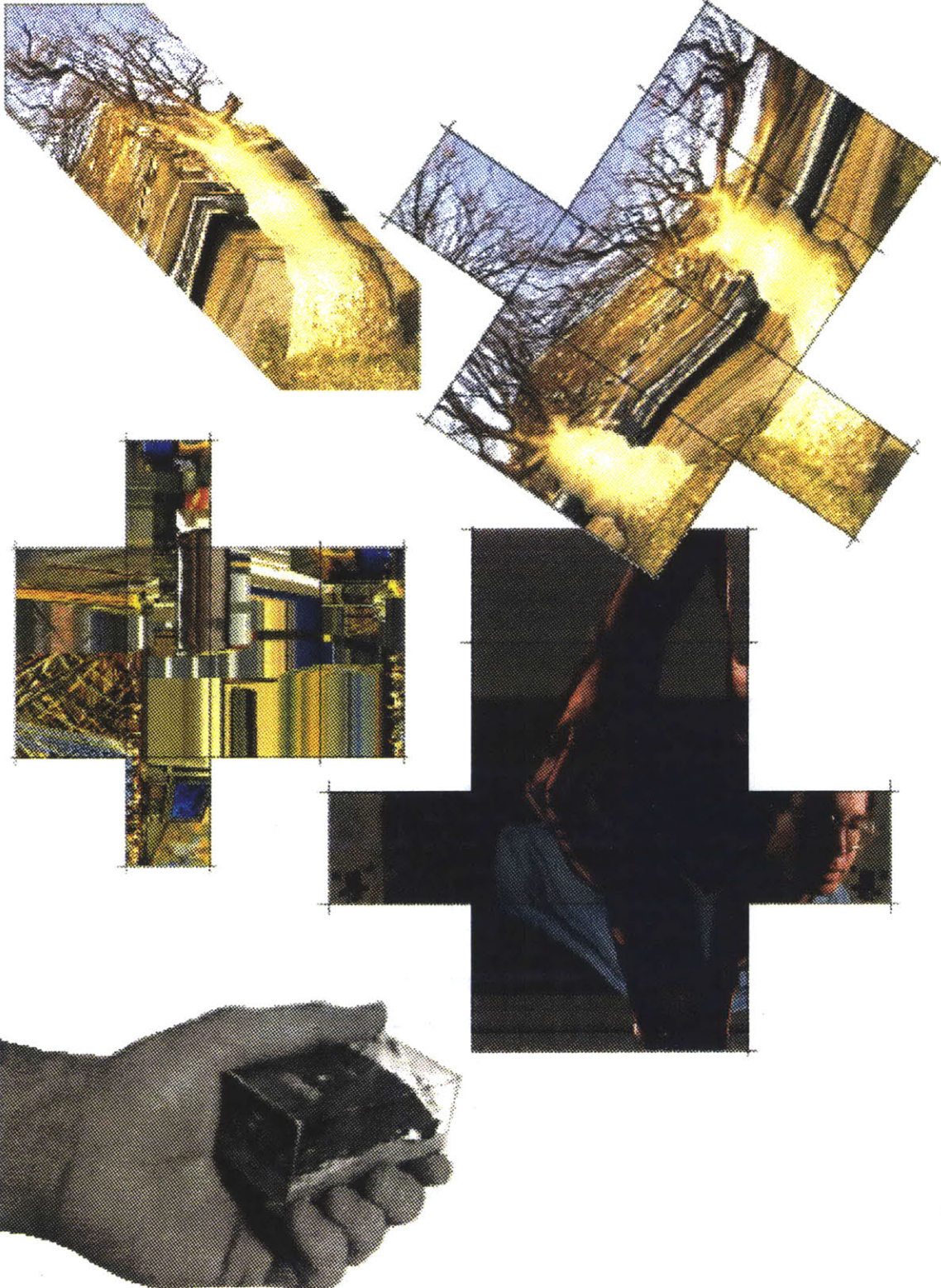
What would a *made for streaming* movie be like? How would you shoot and edit differently if you expect the viewer to explicitly select, save, and review certain shots? Perhaps you might use sweeping camera motions that streamed well to raise the streamer viewer's attention for key shots. For example, subtle clues in a murder mystery might be shot and edited to stream beautifully hoping for review and to be saved in the viewer's solution collage.

8.3 Merely Notions:

8.3.1 What does video feel like?

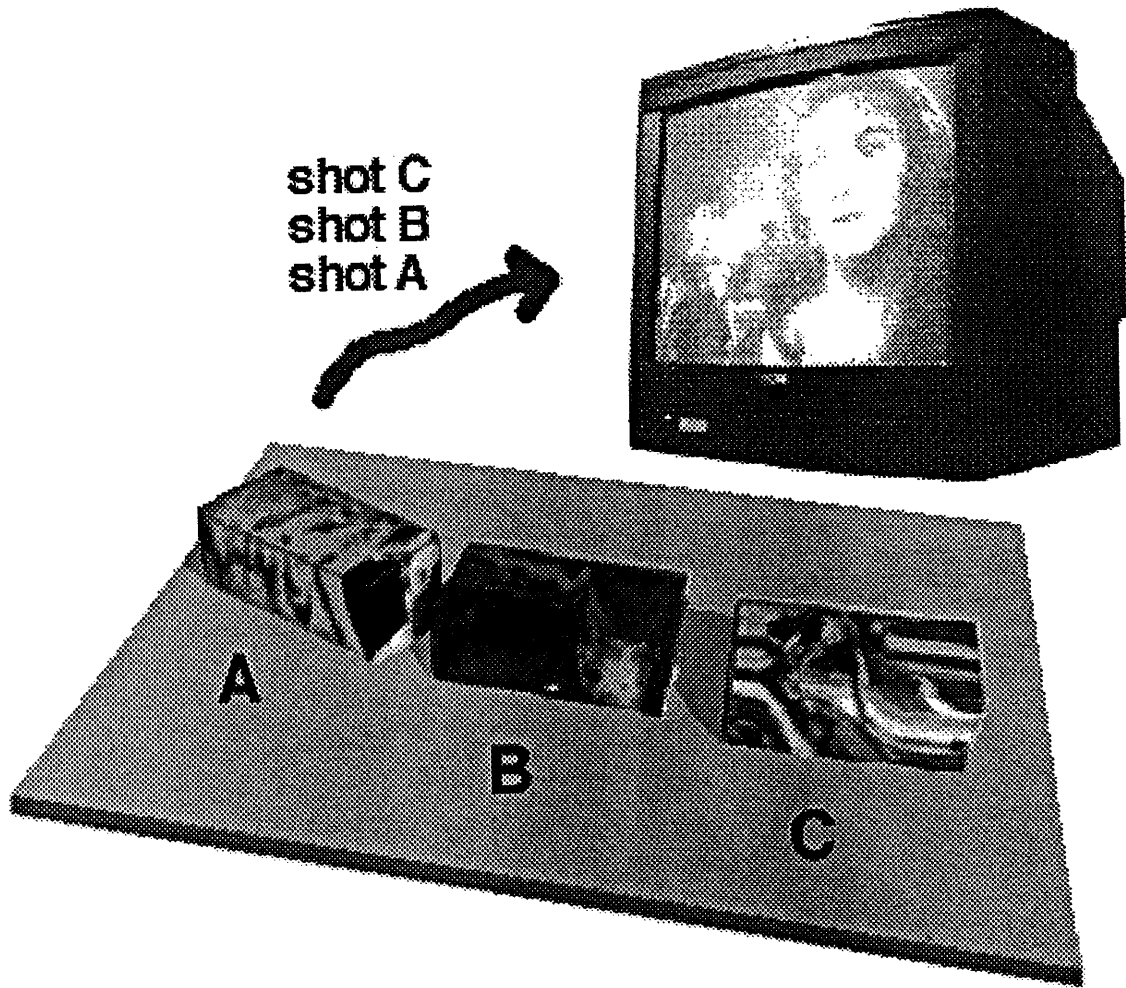
I noted earlier (section 6.3) that things are more tangible when we can touch them and hold them. Example: Just sitting in the driver's seat and holding the steering wheel changes the way we think when buying a new car. Flip books are motion pictures you can hold in your hand. How might we relate differently to video if we could hold it in our hand? Not a video cassette. Not a video watchman, which is 80% case and controls and 20% video. But a video object that is 100% motion images.

We can unwrap the streamer's surface to present its six sides as an unfolded box. This can be printed and folded back up to form a hand held block of a frozen chunk of time. How long till hand held streamer boxes flow with motion images?



8.3.2 Desktop Video

Video objects can be as easily rearranged as building blocks on a table or photos in an album. That will be desktop video editing.



9 CONCLUSION

Usually, the best way to enjoy moving pictures is to plop down with your favorite snack in a comfortable seat before a large screen with friends and just absorb what you see. There are also occasions when it is useful or appropriate for viewing to be more actively engaging, even analytical and critical.

VIDEO TO HOLD

The original focus of this thesis project was to be a reconfigurable viewing environment for associative thinking about multiple streams. But the streamer sprouted mid-thesis and quickly blossomed, begging for more attention and leaving the collage less developed than I had originally anticipated. Still, I think the collage's multi-streams begin to suggest one way for cultivating understanding through working with motion images.

MOTION PICTURE OBJECTS TO THINK WITH

So, are the streamer and the collage effective at inducing reverie around creative manipulation of motion images? The streamer is off to a good start, but the collage has a way to go. The streamer presents a compelling way to acquire and review a chunk of footage. The collage's multi-streams offer a way of thinking with video with less allegiance to its linear origin. Together they suggest a cousin of editing, a viewing activity where one can tinker with motion images as a way of thinking about them.

THINK WATCHING

BIBLIOGRAPHY

Andrew, J. Dudley. *The Major Film Theories*. London: Oxford University Press, 1976.

Akutsu, Akihito, Yoshinobu Tonomura, Hideo Hashimoto, and Yuji Ohba. "Video Indexing Using Motion Vectors". Presented at SPIE 1992, November 1992.

Arnheim, Rudolf. *Visual Thinking*. Berkeley: University of California Press, 1969.

Bazin, Andre. *What is Cinema?* trans. Hugh Gray. Berkeley: University of California Press, 1962.

Bolt, Richard A. "Gaze-Orchestrated Dynamic Windows". *Computer Graphics*, Vol. 15, No. 3, August 1981.

Bolt, Richard A. *Spatial Data-Management*. MIT, 1979.

Bordwell, David, and Kristin Thompson. *Film Art*. New York: McGraw-Hill Publishing Company, 1990.

Bordwell, David. *Narration in the Fiction Film*. Madison, Wisconsin: The University of Wisconsin Press, 1985.

Bordwell, David. *Making Meaning*. Cambridge: Harvard University Press, 1989.

Brondmo, Hans Peter, and Glorianna Davenport. "Creating and Viewing the Elastic Charles - a Hypermedia Journal". *Hypertext II Conference Proceedings*, York, England, July 1989.

Bruckman, Amy. "The Electronic Scrapbook: Towards an Intelligent Home-Video Editing System." Master's thesis, MIT Media Lab, September 1991.

- Chalfen, Richard. "Cinéma Naïveté: A Study of Home Moviemaking as Visual Communication". In *Studies in the Anthropology of Visual Communication*, 2:2(87-103), 1975.
- Cubitt, Sean. *Timeshift: On Video Culture*. London: Routledge, 1991.
- Davenport, Glorianna, Thomas Aguiere Smith, and Natalio Pincever. "Cinematic Primitives for Multimedia". *IEEE Computer Graphics & Applications*, July 1991.
- Davenport, Glorianna, Ryan Evans, and Mark Halliday. "Orchestrating Digital Movies". Working paper, MIT Media Lab, 1992.
- Eisenstein, Sergei. *Film Form*. trans. Jay Leyda. New York: Harcourt, Brace and Company, 1949.
- Haber, Ralph Norman. "How We Remember What We See". *Scientific American*, 1970, pp. 104-112.
- Hindus, Debby. "Semi-Structured Capture and Display of Telephone Conversations". Master's thesis, MIT Media Lab, February 1992.
- Hoffman, Hans. *Search for the Real, and Other Essays*. edited by Sara T. Weeks and Bartlett H. Hayes, Jr., Cambridge, Massachusetts, The MIT Press, 1967.
- Kuleshov, Lev. *Kuleshov on Film: Writings by Lev Kuleshov*. translated and edited by Ronald Levaco, Berkeley: University of California Press, 1974.
- Liesegang, Franz Paul. *Dates and Sources*. translated and edited by Hermann Hecht, London: The Magic Lantern Society of Great Britain, 1986.
- Mills, Michael, Jonathan Cohen, and Yin Yin Wong. "A Magnifier Tool for Video Data". *Proceedings of CHI '92*, pp. 93-98.
- Minsky, Marvin. *The Society of Mind*. New York: Simon and Schuster, 1986.
- Monaco, James. *How to Read a Film*. New York: Oxford University Press, 1972.
- Norman, Donald A. *The Design of Everyday Things*. New York: Doubleday, 1988.
- Ohba, Akio. "Interactive Video Image Coding & Decoding, (Two new tools for video editing and computer animation)". Next Generation Human Interface Architecture Workshop '90, November 1990.

Oppenheimer, Frank. "Adult Play". *The Exploratorium*, special issue, March 1985.

Oppenheimer, Frank. *Working Prototypes: Exhibit Design at the Exploratorium*.

Otsuji, Kiyotaka, Yoshinobu Tonomura, and Yuji Ohba. "Video Browsing Using Brightness Data". *Visual Communications and Image Processing '91: Image Processing*, SPIE Vol. 1606, 1991.

Papert, Seymour. *Mindstorms*. New York: Basic Books, 1980.

Pincever, Natalio. "If You Could See What I Hear: Editing Assistance Through Cinematic Parsing". Master's Thesis, MIT Media Lab, June 1991.

Programmer's Guide to QuickTime. Alpha draft, Apple Computer, Fall 1991.

Rubin, Benjamin. "Constraint-Based Cinematic Editing". Master's thesis, MIT Media Lab, June 1989.

Sasnett, Russell. "Reconfigurable Video". Master's thesis, MIT Media Lab, February 1986.

Steinbeck, John. *The Log from the Sea of Cortez*. New York: Penguin Books, 1986.

Teodosio, Laura. "Salient Stills". Master's thesis, MIT Media Lab, June 1992.

Tonomura, Yoshinobu. "Video Handling Based on Structured Information for Hypermedia Systems". International Conference on Multimedia Information Systems '91.

Turk, Matthew. "Interactive-Time Vision: Face Recognition as a Visual Behavior". Ph.D. thesis, MIT Media Lab, September 1991.

Ueda, Hirotada, Takafumi Miyutake, and Satoshi Yoshizawa. "Impact: An Interactive Natural-Motion-Picture Dedicated Multimedia Authoring System". *CHI '91 Conference Proceedings*, pp. 343-350.