The Stratification System
A Design Environment for Random Access Video

Thomas G. Aguierre Smith
SM Media Arts and Sciences

Glorianna Davenport
Assistant Professor of Media Technology
Director of Interactive Cinema Group

The Media Lab
Massachusetts Institute of Technology
20 Ames Street
Cambridge, MA 02139

thomas@media.mit.edu and gid@media.mit.edu

Abstract

Content of a movie is produced in two different types of design environments. The first is the design environment of shooting where a camera is used to capture what is happening at a particular place and time. The second is the design environment of editing where the rushes are interpreted relative to a movie maker's intent. Annotation of the video stream allows the movie maker to make decisions based on specific content of video and in the best case enables a machine to help in that process.

Stratification is a context-based layered annotation method which treats descriptions of video content as objects. Stratification offers an graphical representation of the content of a video stream and enables movie makers to quickly query and view descriptions for any chunk of video. Stratification supports the development of complementary or even contradictory descriptions which result when different researchers access video source material which is made available on a common workstation or over a network.

The Stratification System was implemented on a DECstation 5000 UNIX workstation in Motif. The system was developed under the direction of Glorianna Davenport at the Interactive Cinema Group of the MIT Media Laboratory with partial support from British Telecommunications, Pioneer Corporation, and Asahi Broadcasting.

Introduction

A random access video database system is a challenging information management problem because the way that a particular sequence of images is described effects on how a maker will retrieve it and incorporate it into a movie.

Researchers in Interactive Cinema work with real footage of their own choosing and build systems that demonstrate the relationship between movie making and computational concepts. Stratification focuses on the idea that the movie maker's interpretation of what unfolds in front of a camera is dynamic.

The intention shifts from the shooting environment where the leading question is "What is happening and how can I best represent it with a recorder?" to the editing environment where content is re-evaluated and becomes "How can I arrange moving images with the purpose of communicating something to an outsider?"

Computation becomes a characteristic of the design environment when motion picture editing becomes digital. Digital movie making requires an *environment* that supports how makers use their knowledge of working and interacting with a medium to make content.

Motion picture content is produced in two different types of design environments each having its own set of constraints. Content first emerges in the *Design Environment for Shooting* and then is transformed into the final motion picture in the *Design Environment for Editing*. Between these two processes a log of the content of the video is produced that helps the movie maker remember what was shot and to guide the production of the final movie.

Background

A relational database of keywords (who, what, when, where) was used to describe the content of shots in Glorianna Davenport's *A City in Transition* New Orleans, 1983-86 (1987). A story generation engine relied on complex keyword queries to present the viewer with constrained views of the material that was available on the system. Ricki Goldman Segall in her thesis *Learning Constellations* (1990) edited ethnographic video observations into thematically coherent chunks. In both early experiments, the process of breaking the video into discrete chunks inhibited the dynamic between the maker and the user of the video material.

The development of a computer representation for the content of video called *stratification* represents a first step where a database of descriptions reflects how a camera records with in the context of an environment. Stratification was first developed as theoretical model for how multiple annotations can be applied to a stream of video (Davenport, et al. 1991) and was tested during an ethnographic video production in the state of Chiapas Mexico (Aguierre Smith 1992).

From Discrete to Stream-based Annotations

Most video tape logs consist of content markers which are scribbled on a piece of paper as the movie maker reviews the raw footage. The log depicted in figure 1 represents 26,310 contiguous frames (14 minutes) that were recorded in the Municipio of San Juan Chamula, Chiapas, Mexico while visiting the home of Carmelino Santiz Ruiz. Carmelino shows the video maker a medicinal plant garden and then invites the video maker into the house. Upon entering, the video maker notices five cases of Pepsi bottles stacked in the corner. As the Pepsi cases are video taped, Carmelino begins to pray at the altar in the room. After this, the video maker follows Carmelino into the kitchen where he is asked about the cases of Pepsi bottles.

Figure 1: Video Log of Video Shot at Carmelino's house

```
tape07 | 84793 | and this one also
tape07 | 85163 | for burns ?
tape07 | 85879 | fry it in a comal
tape07 | 86805 | grind it like this
tape07 | 87557 | applied like this
tape07 | 88050 | after four days
tape07 | 88823 | name of burn plant?
tape07 | 89667 | Carmelino's house
tape07 | 90035 | interior w bike
tape07 | 92263 | Pepsi bottles
tape07 | 92957 | lighting candles
tape07 | 93947 | lighting first one
tape07 | 94260 | translation 52.23
tape07 | 94800 | translation 52.41
tape07 | 94847 | beginning the prayer
tape07 | 96720 | squatting and praying
tape07 | 97913 | end zoom INRI cross
tape07 | 99061 | This is the kitchen
tape07 | 99449 | thanks to Dr. Berlin
tape07 | 99819 | hay luz tambien
tape07 | 100615 | grinding corn
tape07 | 103319 | drinking Pepsi
tape07 | 103757 | we only have Pepsi here
tape07 | 104231 | close up of Pepsi cola
tape07 | 107563 | everyone drinks Pepsi
tape07 | 111103 | women walking home
```

Each record in the database consists of the tape name; the frame number; a free text description. The effectiveness of these traditional logs is wholly dependent on the linearity of the medium. For example, when locating "Pepsi bottles" (frame 92263) the video editor uses the shuttle knob on the editing deck to fast forward to that location. In doing so, all the material that has been recorded up to that point appears on the monitor. The editor sees the context that the "Pepsi bottles" -- he sees the shots that come before and after it. He notes that it is taking place in Chamula; in Carmelino's garden; now we are in his house; and in a moment he will begin to pray. When sorted by frame number the free text descriptions provide the context for each content marker. For example, "Pepsi bottles" on the diagram above appears in chronological order between "interior with bike" and "lighting candles."

But when searching the database for the word, "Pepsi", the list of content markers returned cannot provide the context. In Figure 2, a database search for the words "Pepsi" among all the video annotations does not provide the needed context. With a computerized database of content markers one can rapidly find that items that they are interested in. But this is not as useful as expected. The maker needs to see the surrounding annotations.

Figure 2: Database Search for Word "Pepsi" without Context

tape06	1707	Pepsi or coke
tape07	92263	Pepsi bottles
tape07	103319	drinking Pepsi
tape07	103757	solo hay Pepsi
tape07	104231	close up of Pepsi cola
tape07	107563	everyone drinks Pepsi
tape13	74721	Pepsi bottles
tape14	93083	Pepsi .. Fanta
tape15	28487	Pepsi and orange
tape11	106501	arranging Pepsi
tape23	96843	Pepsi at Antonio's house
tape23	108901	Pepsi delivery

In a random access system the linear integrity of the raw footage is erased. In turn the contextual information that relates to the environment where the video was shot is also destroyed. What is required is a method to record this contextual information so that is can be recovered and re-used at a later time.

Keywords and Context

Keywords provide a more generalized way to create consistent descriptions from one tape to the next. With keywords, one can consistently find related chunks of video among the 27 hours of video tape that were shot. As shown in Figure 3, the keywords of "Chamula" and "Carmelino" remain constant while content markers specifically identify what is happening at a given moment. Keywords provide the context for content markers. Sets of keyword descriptors remain constant while the other descriptions change and evolve.

Figure 3: Content Markers with Keywords Sorted by Frame Number

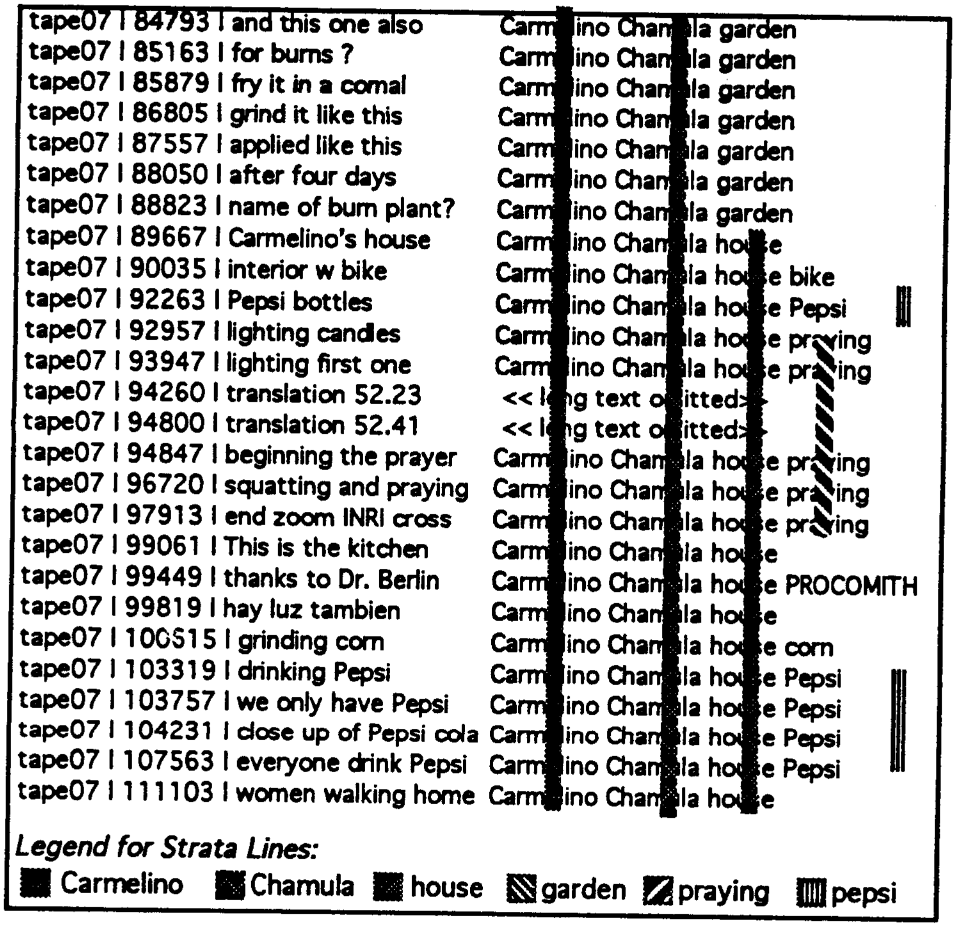
tape07	84793	and this one also	Carmelino Chamula garden
tape07	85163	for burns ?	Carmelino Chamula garden
tape07	85879	fry it in a comal	Carmelino Chamula garden
tape07	86805	grind it like this	Carmelino Chamula garden
tape07	87557	applied like this	Carmelino Chamula garden
tape07	88050	after four days	Carmelino Chamula garden
tape07	88823	name of burn plant?	Carmelino Chamula garden
tape07	89667	Carmelino's house	Carmelino Chamula house
tape07	90035	interior w bike	Carmelino Chamula house bike
tape07	92263	Pepsi bottles	Carmelino Chamula house Pepsi
tape07	92957	lighting candles	Carmelino Chamula house praying
tape07	93947	lighting first one	Carmelino Chamula house praying
tape07	94260	translation 52.23	<< long text omitted>>
tape07	94800	translation 52.41	<< long text omitted>>
tape07	94847	beginning the prayer	Carmelino Chamula house praying
tape07	96720	squatting and praying	Carmelino Chamula house praying
tape07	97913	end zoom INRI cross	Carmelino Chamula house praying
tape07	99061	This is the kitchen	Carmelino Chamula house
tape07	99449	thanks to Dr. Berlin	Carmelino Chamula house PROCOMITH
tape07	99819	hay luz tambien	Carmelino Chamula house
tape07	100615	grinding corn	Carmelino Chamula house corn
tape07	103319	drinking Pepsi	Carmelino Chamula house Pepsi
tape07	103757	we only have Pepsi	Carmelino Chamula house Pepsi
tape07	104231	close up of Pepsi cola	Carmelino Chamula house Pepsi
tape07	107563	everyone drink Pepsi	Carmelino Chamula house Pepsi
tape07	111103	women walking home	Carmelino Chamula house

Stratification

It becomes evident that descriptions of content have a lot to do with the linearity of a medium. In a random access system we can't rely on the linearity of the medium to provide us with a coherent description. Accordingly we need a new type of descriptive strategy that allows for the annotation of descriptions to have a temporal extent.

When sorted on frame number, the content markers become embedded in patterns of keywords. These patterns illustrate the contextual relationships among contiguously recorded video frames. It also illustrates how context is wedded to the linearity of the medium. We can now trace, in this pattern, what was shot and where. Sorted lists of content markers with keywords produce a layered representation of context. These layers are called *strata* (Figure 4).

Figure 4: Log of Content Markers with Strata.



```
tape07 | 84793 | and this one also        Carmelino Chamula garden
tape07 | 85163 | for burns ?              Carmelino Chamula garden
tape07 | 85879 | fry it in a comal        Carmelino Chamula garden
tape07 | 86805 | grind it like this       Carmelino Chamula garden
tape07 | 87557 | applied like this        Carmelino Chamula garden
tape07 | 88050 | after four days          Carmelino Chamula garden
tape07 | 88823 | name of burn plant?      Carmelino Chamula garden
tape07 | 89667 | Carmelino's house        Carmelino Chamula house
tape07 | 90035 | interior w bike          Carmelino Chamula house bike
tape07 | 92263 | Pepsi bottles            Carmelino Chamula house Pepsi
tape07 | 92957 | lighting candles         Carmelino Chamula house praying
tape07 | 93947 | lighting first one       Carmelino Chamula house praying
tape07 | 94260 | translation 52.23        << long text omitted>>
tape07 | 94800 | translation 52.41        << long text omitted>>
tape07 | 94847 | beginning the prayer     Carmelino Chamula house praying
tape07 | 96720 | squatting and praying    Carmelino Chamula house praying
tape07 | 97913 | end zoom INRI cross      Carmelino Chamula house praying
tape07 | 99061 | This is the kitchen      Carmelino Chamula house
tape07 | 99449 | thanks to Dr. Berlin     Carmelino Chamula house PROCOMITH
tape07 | 99819 | hay luz tambien          Carmelino Chamula house
tape07 | 100515 | grinding corn           Carmelino Chamula house corn
tape07 | 103319 | drinking Pepsi          Carmelino Chamula house Pepsi
tape07 | 103757 | we only have Pepsi      Carmelino Chamula house Pepsi
tape07 | 104231 | close up of Pepsi cola  Carmelino Chamula house Pepsi
tape07 | 107563 | everyone drink Pepsi    Carmelino Chamula house Pepsi
tape07 | 111103 | women walking home      Carmelino Chamula house
```

Legend for Strata Lines:

■ Carmelino ▩ Chamula ■ house ▧ garden ▨ praying ▥ pepsi

This methodological shift is subtle but critical. The linearity of the medium is preserved with stratification because each description has a temporal/linear extent. With other annotation methods, the linearity of the medium is destroyed because we have broken up the footage into *ad hoc* chunks. The keywords form strata that capture changes in descriptive state which the camera recorded.

If you know enough about the environment in which you are shooting you can derive a good description of the images that you have captured using stratification.

Any frame can have a variable number of strata associated with it. The content for *any* set of frames can be *derived* by examining the union of all the contextual descriptions that are associated with it.

Content can now be broken down into distinct descriptive threads or strata. One stratum constitutes a single descriptive attribute which has been derived from the shooting environment. When these descriptive threads are layered one on top of the other they produce descriptive strata from which inferences about the content of each frame can be derived. Stratification is an *elastic* representation of the content of a video stream because descriptions can be derived for chunk of video of any size.

Stratification is a method which produces layers of descriptions that can overlap, be contained in, and even encompass a multitude of other descriptions. Moreover, each additional descriptive layer is automatically situated within the descriptive strata that already exit. Users can create descriptions which are built upon each other rather then worrying about how to uniquely describe each frame independently.

In addition to logging, film makers need tools which will enable them to take segments of raw footage and arrange them to create meaningful sequences. Editing is the process of selecting chunks of footage and sound and rearranging them into a temporal linear sequence. The edited linear sequence may bear no resemblance to the environmental context that was in effect during recording. Conceivably, one can edit the source material in the video database into a documentary movie that will be played back on the computer. Moreover, these "edited-versions" can later be used by someone else to make another movie production.

When editing a sequence, important relationships in the raw footage come to the fore. In an edited sequence causal and temporal relationships in the raw footage are made explicit.

Implementation:

The Stratification system was intended as an experiment to test and work out the details of a stream based annotation system. Users should be allowed to use free text descriptions and more structured keyword type descriptions. A visual mapping of descriptions to a timeline was required to facilitate browsing. In addition the system needed to support multiple users on a network. Researchers needed to apply more than one type of analysis to any :hot of video. They might be interested in the transcript of the material in the field, linguistic analysis, narrative style, etc. Good source material should lend itself to be re-employed in different research environments and for different needs. And finally the format for the data files needed to be easily modifiable and accessible over the network. Knowledge about the content of video changes as time progresses. In many cases initial annotations need to be revised as new information becomes available.

Data Representations

The use of keyword classes and a special format for saving descriptions of video called Strata Data Format (SDF) are key features of the Stratification system. The implementation of keyword classes and SDF is designed to complement the file management and text processing utilities currently available in the UNIX operating system.

Each descriptive stratum consists of the source name, begin frame, end frame, free text description field, and keyword classes field. These descriptions are saved in delimited ASCII text files and stored in UNIX directories. SDF files are named in regard to a particular project and owned by an individual or group like any other UNIX file. SDF files can be combined and analyzed for associative browsing of content across projects.

Editing of these files can be accomplished using conventional text editors. In addition, easy to make UNIX shell scripts can be used to parse the files of stratified descriptions (figure 5).

Figure 5: Example of SDF Generated with the Stratification Method. Line 1 - 2 is the header which shows the context of the keyword classes that appear in the file. Line 3 - 11 are content marker annotations. Line 12 - 14 are class keyword strata lines.

```
/mas/ic/src/VIXen/Classes/Maya/places.class|
/mas/ic/src/VIXen/Classes/Maya/people.class
MayaMed|334|334|334|30|corn blowing in the wind
MayaMed|505|505|505|30|path to dominga's house
MayaMed|588|588|588|30|Laguana Pejte'
MayaMed|701|701|701|30|dominga walks down hill
MayaMed|1091|1091|1091|30|chicken's
MayaMed|1267|1267|1267|30|wacking kid
MayaMed|2042|2042|2042|30|pressing boys chest
MayaMed|2264|2264|2264|30|pressing arm
MayaMed|2484|2484|2484|30|throwing plants out
MayaMed|334|2904|1619|30| |places|Chamula
MayaMed|701|1090|895|30| |people|Dominga
MayaMed|1267|2904|2085|30| |people|Dominga
```

The UNIX file system is way to structure and organize annotations and even movie sequences. Ownership can be set for access. The place where a movie is stored can provide important contextual information about the content of a sequence. This of course requires that the user is somewhat rigorous about naming and creating directories. The additional effort pays off when tracing the use of a piece of footage in the system. A consistent format for both raw footage and edited footage enables the researcher to analyze how descriptions of raw and edited footage are built up through use.

Keyword Classes

Key words classes are organized into class hierarchies which are implemented as directory trees in UNIX. Each keyword class is stored as an ASCII text file. If desired, the user can edit a keyword class file with any UNIX text editor. The UNIX file system provides a simple yet useful way to structure different types of knowledge about a video resource. Researchers can have access to each other's keyword class files by setting the permissions on the files accordingly.

Successful perusal of the video database requires knowledge of the descriptive strategies that were used to describe the content. The choice of keywords is related the users intentions; they reflect the purposes and goals of a particular multimedia production. Keyword classes provide a flexible structure that allows for consistency in naming within a particular descriptive strategy. In the end, in the use of keywords consistency will help browsers.

The Stratagraph

The Stratagraph is an interactive graphical display of an annotated video stream. It is a visual representation of the occurrence of annotations of video though time. Keyword classes are displayed as buttons along the y-axis (figure 6).

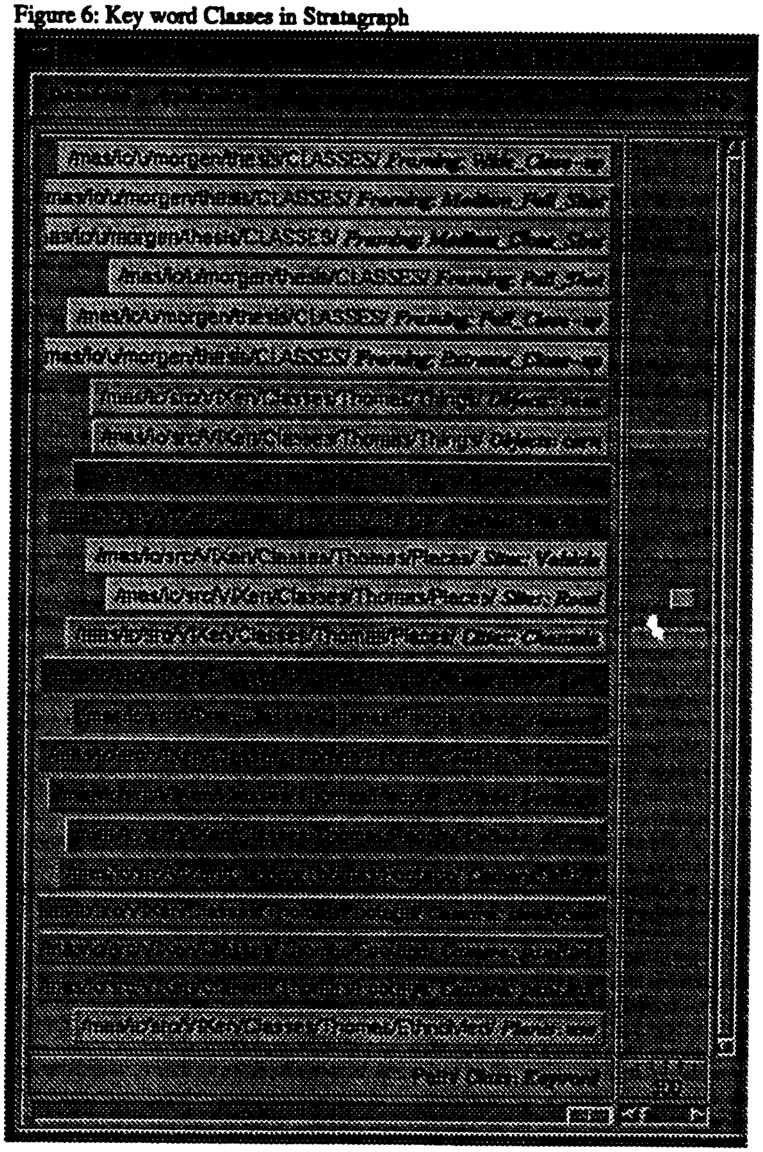
Each button shows the keywords path name in the UNIX file system. The path name indicates the context for each keyword class. On the y-axis one can inspect where a particular keyword is from and how it is related to other keywords.

These all can be included on the y-axis. For example as shown in figure 6, the first strata displayed belong to the user "morgen"while the others belong to thomas.

Each keyword class has its own color. The keyword classes on the vertical axis are also buttons. When pressed, the graph scrolls to display the first instance of that keyword and the video is cued to the in-point.

The units on the horizontal axis are time code (frame numbers for the laserdisc). Another type of interaction consists of clicking the horizontal axis with the mouse. If the user wants to know about the annotations that are associated with any particular frame. The user can click on a frame number (the horizontal axis) to create a "strata

Figure 6: Key word Classes in Stratagraph



line" that intersects all the strata that are layered on top of that particular frame. The laserdisc is cued to the frame number selected and a report showing all the descriptions that are associated with these strata lines is displayed in the "Strata Content" window.

The strata line can be extended for a chunk of video by clicking the right mouse. The left click and move and right click action is called a "strata rub" (figure 7). This rubbing action displays all descriptions which are in effect for that chunk of video in a "Strata Content" window(figure 8) while the laserdisc plays the shot .

Figure 7: Stratagraph showing strata rubs.

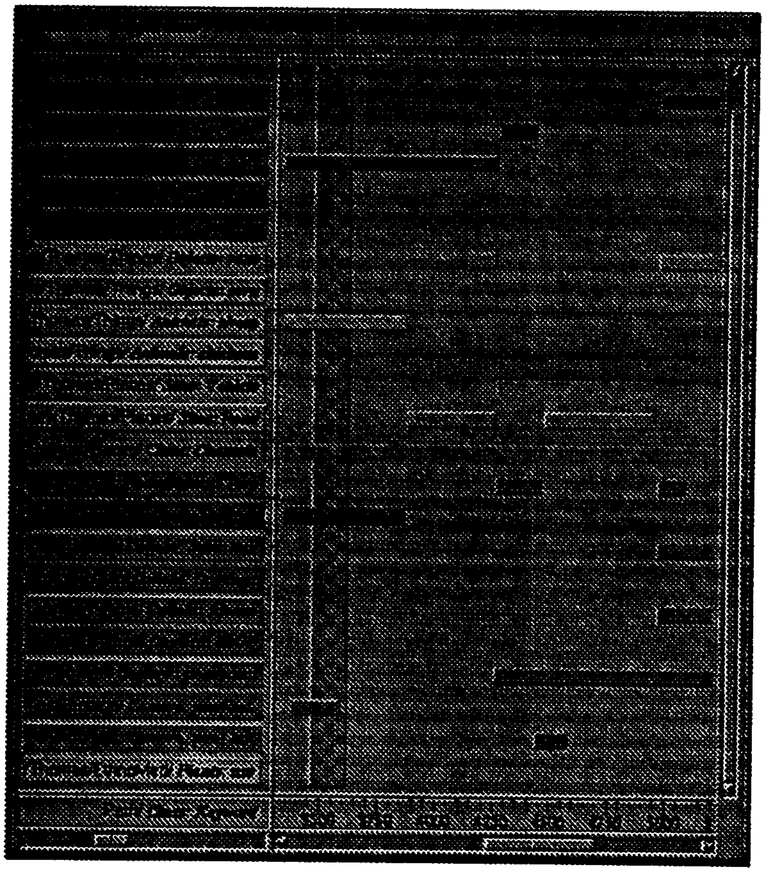
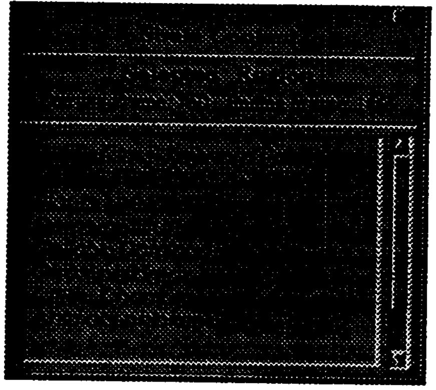


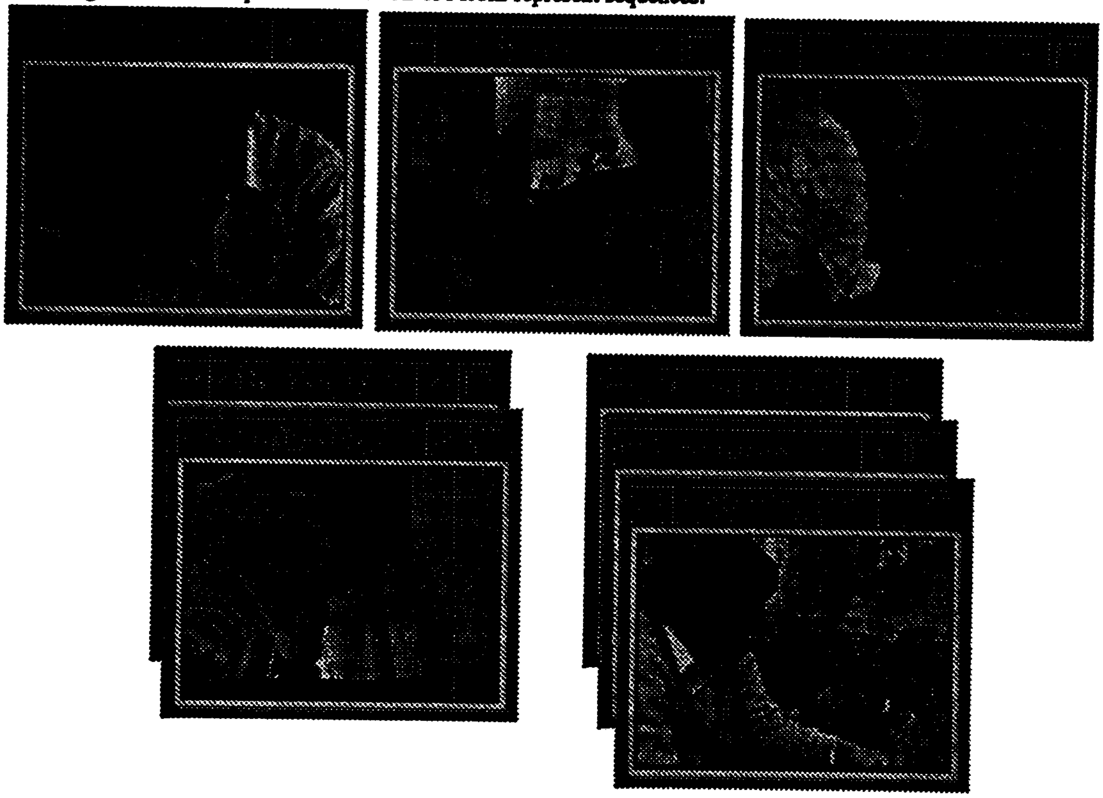
Figure 8: Strata content window.



The shot needs to be named and made into an icon so that it can be arranged into sequences. When naming, the maker creates a mnemonic that is used to reference the shot a later time. Although in and out points are important cues for visual continuity, they might not provide an adequate representation of the shot content. A content frame is digitized and made into a Picon (picture icon) its associated frame number is stored as the content marker frame for the shot. A Picon has a title bar which displays the annotation (figure 9).

Arrays of Picons are displayed on the screen and arranged into sequences. When Picons are arranged in a cascade (left to right and up and down) they can be played back in the order that they appear. Ordered groups of PIcons compose a sequence.

Figure 9: Picons represent shots. Sets of Picons represent sequences.



When a satisfactory sequence is arranged, annotations, volume name, in and out frame numbers are saved as an ASCII SDF file with the name of the file as the sequence name (figure 10). Sequences can be saved as play lists and can be re-loaded back into the system at a later time.

Figure 10: SDF file for a sequence "Carmelino.work.movie"

```
Filename: Carmelino.work.movie
MayaMed|13706|13782|13777|30|face
MayaMed|14024|14294|14026|30|arranging
MayaMed|14899|15102|15000|30|the press
MayaMed|15555|16260|16000|30|the dryer
MayaMed|18012|18507|18400|30|the garden
MayaMed|18510|19140|19000|30|house
MayaMed|19157|19353|19200|30|with healer
```

A sequence file is a play list that shows how chunks of video are related in a new context. The first generation of annotations reflect the context of where the images were recorded, subsequent annotations reflect shifts in meaning that occur when the chunk appears in a virtual edited sequence. Logging video and assembling sequences collapse into the same process. The source material does not change per se, but the context of the material is what dynamically changes. Since cinematic context is inextricably linked to context, the significance of the source material becomes transformed and refined through use.

Conclusion:

Stratification is a first step which maps and creates a correspondence between the Intentionality which was manifest in shooting to a computational log. This log is designed to enhance the conversational exchange between human and machine via a graphical interface.

Stratification serves as a computer representation of the video stream on random access system computer system. The integrity of the context of where the footage was shot is maintained while additional contexts can be created during the process of sequence assembly.

In a sense, the content of a series of frames is defined during logging. Yet the significance of those frames gets refined and built up through use. This information is valuable for individuals who want to use the same video resources over a network in order to communicate with each other. The stratification method provides a way to represent alternative "readings/significations/edits" of the same video resource to coexist on the system.

References

Aguierre Smith (1992) <u>If You Could See What I Mean... Descriptions of Video in an Anthropologist's Video Notebook.</u> SM Thesis, Media Arts and Sciences, MIT, September 1992.

Davenport, G. (August 19, 1987). New Orleans in Transition, 1983 -1986: The Interactive Delivery of a Cinematic Case Study. The International Congress for Design Planning and Theory. Boston : Park Plaza Hotel

Davenport, G.,Aguierre Smith, T., & Pincever, N. (1991). Cinematic Primitives for Multimedia: Toward a more profound intersection of cinematic knowledge and computer science representation. IEEE Computer Graphics and Applications(July).

Goldman Segall, R. (1990). Learning Constellations: A Multimedia Ethnographic Research Environment Using Video Technology for Exploring Children's Thinking. Ph.D., Media Arts and Sciences, MIT, August 1990.

Acknowledgments

Special thanks to Erhhung Yuan and Joshua Holden the two undergraduate researchers who have diligently worked on the Stratification modules. We would like to thank Hiroshi Ikeda and Hiroaki Komatsu. This research has been supported by British Telecom, Asahi Broadcasting Corporation and Pioneer.