To appear in the proceedings of the USENIX Summer 1991 Technical Conference and Exhibition: "Multimedia -For Now and the Future."

Parsing Movies in Context

Thomas G. Aguierre Smith
thomas@media-lab.media.mit.edu

Natalio C. Pincever
che@media-lab.media.mit.edu

Interactive Cinema Group,
The Media Lab
Massachusetts Institute of Technology

Abstract

Traditional approaches in Multimedia systems force the user to segment video material into simple descriptions, which usually include endpoints and a brief text description in the form of keywords. We propose to segment contextual information into chunks rather than segmenting contiguous frames.

The computer can help us organize sets of descriptions that are related to the recorded moving image: it can help us remember what we have shot. Such descriptions can overlap, be contained in, and even encompass a multitude of other descriptions. Parsing moving image sequences is reduced to simply parsing the contextual information which forms the description. Our approach, Stratification, also has important ramifications in terms of an elastic representation of moving images. Ambient sound can provide us with important contextual clues as to what is going on within the frames. Using audio to find patterns of content is an important step towards the eventual automatization of the logging process.

1. Approaches for Production: Segmentation versus Stratification

Segmentation forces the user to break down the raw footage into segments denoted by the begin and end point and assign some sort of textual description to them. In a video database application, these textual descriptions are searched and the associated video is retrieved. In our research we have found some problems with this approach.

On a multimedia system in which it is possible to access individual frames randomly, each frame needs to be described independently. The user has to represent the content of each chunk: the part is forsaken for the whole. In terms of granularity, the chunk of video that a database application can retrieve for a user is predetermined during the logging process. The computer representation of the video footage only encompasses a begin and end point and a text description. Sub samples or finer grained search criteria / representation have to be made independently. In terms of elasticity, if you have a segment which contains 30 frames; the application will retrieve all 30 frames when queried. It has no representation of any set of 10 or 20 frames which form a subset of the base frames. One would get 30 frames and then select a sub sample of these frames, describe them independently so that they can be called up later to satisfy a more finely grained search criteria.

Furthermore while segmentation is necessary when editing a video, it imposes a specific intentionality of the person who is placing the material in a structured context. Although this is desirable for a linear editing system, it can seriously impede a group of people who need to use and have access to the same video resources from an archive over a network. If the video material is segmented from the start, how then can the descriptive structures support other users who may need to access the same material for a different purpose?

Our solution is to segment contextual information into chunks rather than segmenting contiguous frames. This new context based approach is called Stratification (Aguierre Smith 1991).

Begin and end frames are used to segment contextual descriptions for a contiguous set of frames. These descriptions are called strata: a shot begins and ends; a character enters and sits down; the camera zooms in. Each represents a stratum. Any frame can have a variable number of strata associated with it or with a part of it (pixel). The content for any set of frames can be derived by examining the union of all the contextual descriptions that are associated with it. Before we can begin to think about parsing, we need a good representation of moving image content Stratification is a way to structure video content information that will allow us the greatest latitude in terms of parsing.

Stratification is a descriptive methodology which generates rich multi-layered descriptions that can be parsed by other applications.

1.1. The Management of Multimedia Resources

The movie maker knows a lot about the images as she is recording them. Knowledge about the content of the moving image is at its maximum while being recorded and drops to a minimum while being viewed. The computer can help us organize sets of descriptions that are related to the recorded moving image: it can help us remember what we have shot Successive shots share a set of descriptive attributes that result from their proximity. These attributes are the linear context of the moving image. During recording, the linear context of the frames and the environmental context of the camera coincide. The environmental context is the "where," "who," "what," "when," "why," and "how" which relate to the scene: it's the physical space in which the recording takes place, giving it a unique identity. 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.

Clearly such descriptions can overlap, be contained in, and even encompass a multitude of other descriptions. Each descriptive attribute is an important contextual element: the union of several of these attributes produces the meaning or content for that piece of film. Moreover, each additional descriptive layer is automatically situated within the descriptive strata in which it already exits. In this way, rich descriptions of content can be built. On the other hand, segmentation, which is conventionally used in computer systems which allow the user to log video material, forces the user to break down raw footage into segments denoted by begin and end points: such a divide-and-conquer method forsakes the whole for the part. Coarser descriptions have to be included at this level of specification in order to describe one frame independently. If the unit represented by in and out points is as small as an individual frame, its description, in order to be independent, must encompass larger descriptive units. The granularity of description of sets of frames is directly related to the size of the image units being represented. In segmentation, the granularity of description is inversely proportional to the size of the image unit in question. The inverse relationship arises out of the need to describe each image unit 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 (Davenport, Aguierre Smith, Pincever 1990). The edited linear sequence may bear no resemblance to the ambient reality that was in effect during recording. During the process of conventional editing, the original rushes become separated from the final motion picture. In order to create a motion picture with a new meaning which is distinct from the raw footage, a new physical object must be created -- an edited print for film, or an edited master for video.

Editing on a digital-movie-database system will be radically different. In a full digital system, the link between source material and final movie does not have to be broken. The shot in the final version of the movie being made will be the same chunk of video data as the source material. At this point there are two names for the image unit which must coexist: one reflects the context of the source material; the other is an annotation that is related to a playback time for a personalized movie script.

1.2 Stratosphere

We have developed an application called Stratosphere which runs on a UNIX workstation under X/Motif using the Galatea video server (Applebaum 1990). Stratosphere (see Diagram 1) consists of a logging module, a stratagraph to display contextual description of frames through time, an icon-based video annotating system called Infocon and soft-story templates.

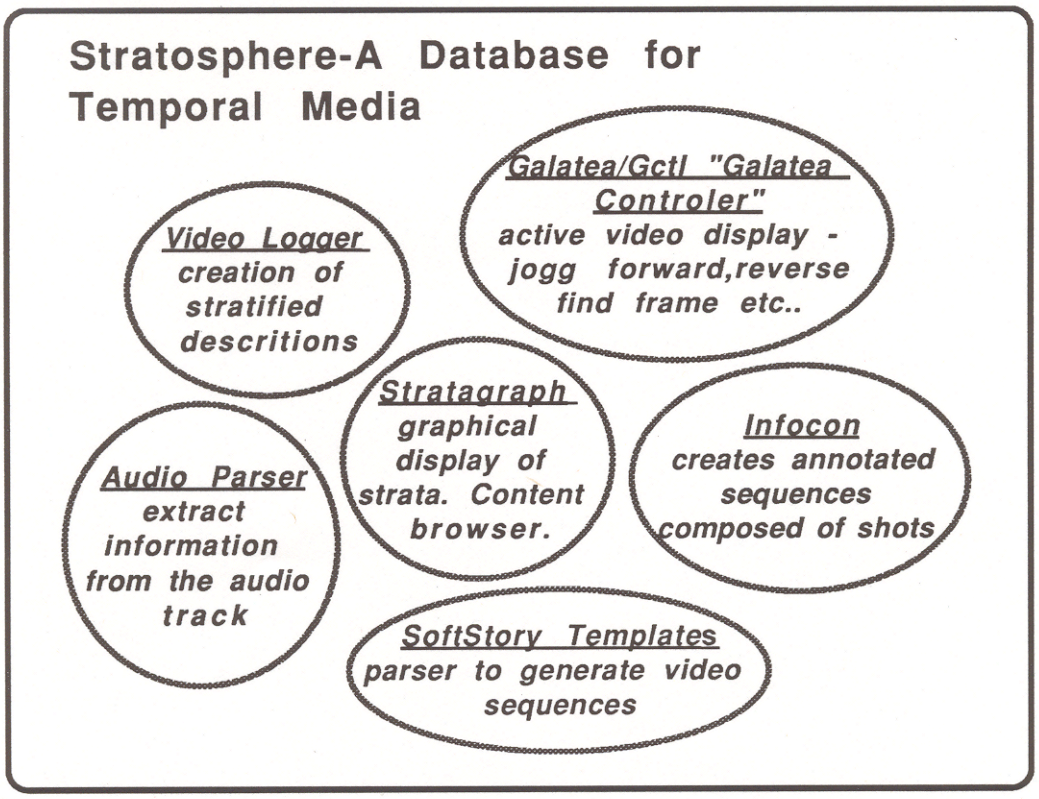

Diagram 1: Stratosphere

Stratosphere is an implementation of the Stratification method for representing moving image content. Stratification allows the user to keep track of contextual descriptions of video material and helps maintain the integrity of the contextual information during all phases of production.

In our UNIX video database environment, the distinction between source and final edit become so blurred that all video becomes a resource which can be used in different contexts (movies). Conceivably, one can edit the source material in the video database into a documentary film that will be played back on the computer. Moreover, these "edited-versions" can later be used by someone else to make another moving image production. The process of editing using Stratification becomes the process of creating context annotations, and storing them along with the initial descriptions made during recording. We are developing tools for video databases that will enable the user to semantically manipulate video resources which allows for alternative "readings/significations/edits" of the same material to co-exist on the system. Video resource can exist in many different contexts and in many personalized movie scripts.

When a user annotates a segment of video, s/he is creating a new stratum that is associated with a particular moving image in an edit. The old straw lines still exist in the database, but in editing a new meaning emerges in the re-juxtapositions of image units. This new meaning takes the form of an annotation and is displayed on the stratification graph (see figure 3). 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.

2. The Stratosphere Video Database Modules

As mentioned earlier, Stratosphere is composed of separate modules which allow the user to log, manipulate, manage and view chunks of video material which have been described using the Stratification method. In this iteration of Stratosphere, we are using a MicroVax II computer with a Parallax video card. UNIX IX Window System, Galatea video server, and Akai PO 1000 digital matrix patch bay which manages a system of six videodisc players.

2.1 The Logger

The Video logger is a Motif window made up of buttons, menus and field widgets which allow the user to select and describe video material. Logging is coordinated with browsing the video material displayed in the "gctl" (Galatea control device) window. Using the mouse and keyboard the user can fast-forward, review, search for a particular frame number. Gctl also has a video slider which enables the user to quickly scroll through video at various speeds in both forward and backward directions with only a click of the mouse. When the mouse is released, the video is stopped within one or two frames. Once the desired in-point end point is found for a descriptive attribute, the user presses the button and the current frame number is displayed. Next, a title and a set of keywords are also entered for each stratum.

Describing video with keywords can be problematic. The choice of keywords is related the users intentions; they reflect the purposes and goals of a given multimedia project. We have addressed this problem by organizing keywords into classes. Keyword classes are sets of related words which can be created, stored and be reused for different projects. Each keyword class is stored as an ASCII text file. If desired, the user can edit the keyword class file with any UNIX text editor. When the user opens a class file, a button widget is created and displayed in the Logger window. Many different classes can be loaded during a logging session. When the user wants to associate a camera keyword with a strata line, s/he clicks on the camera class button and the all the keywords of this class are displayed in the keyword text widget. Double clicking on a particular keyword selects and enters it into the data file. Keywords classes can be combined to create customized keyword for sets. Keyword classes speed data entry. For example, the "camera" keyword class which contains, wide shot, medium shot, pan left, pan right, zoom, etc. is used over and over again.

2.2. Stratagraph and Strata Data Display

The Stratagraph is an interactive graphical display of the data generated by the Logger. It is a visual representation of the occurrence of strata though time. An example of a Stratagraph is show in Figure 1, each descriptive stratum partially describes a set of frames; the union of all the strata is the content The Stratagraph severs as a visual aid for logging video, since the user can log several descriptive attributes simultaneously.

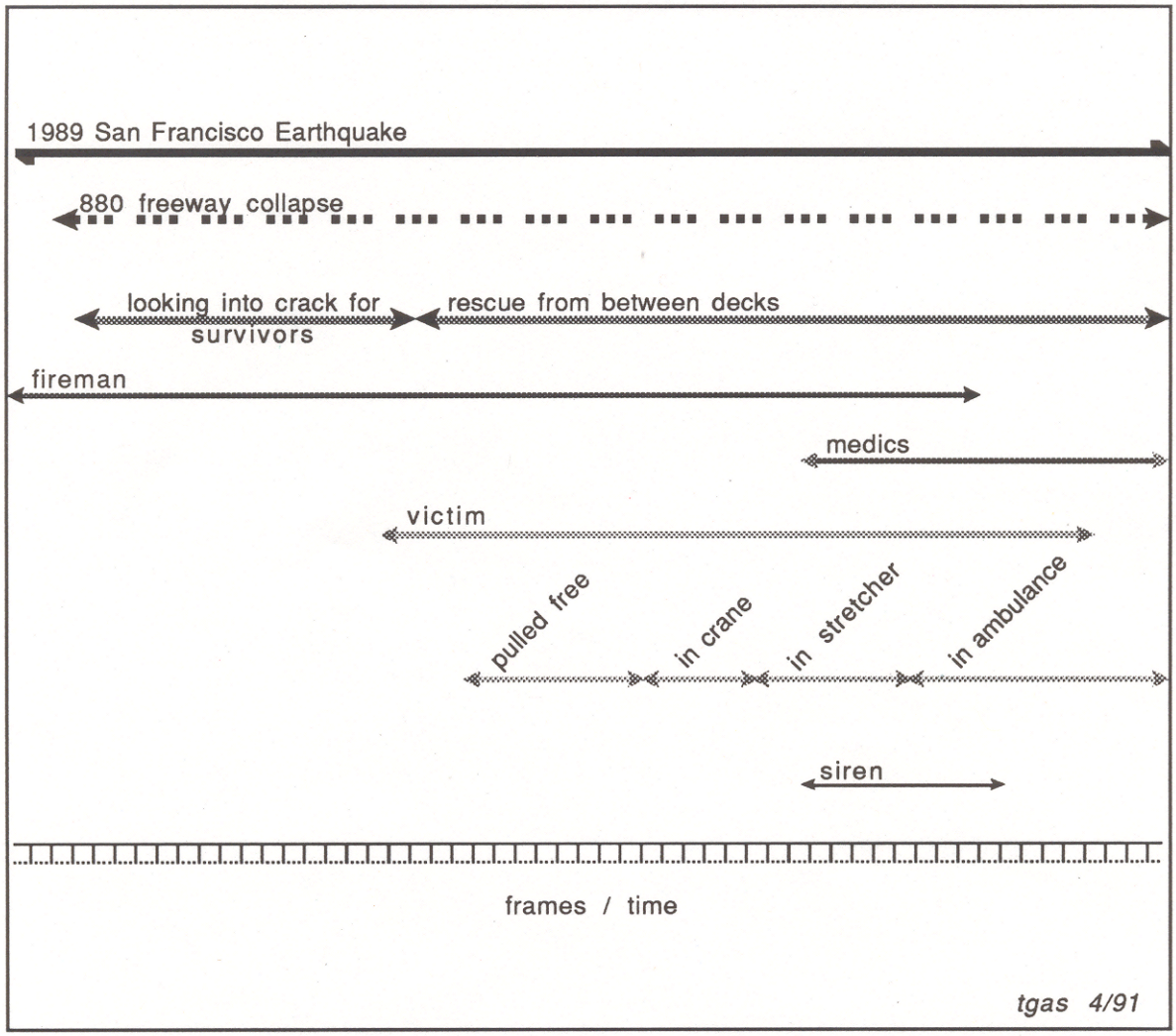

Figure 1: Rescue After the San Francisco Earthquake - Stratified Description

During the logging process, each descriptive attribute is first anchored to the Stratagraph with an in-point and a strata line is generated for each frame that follows. At this moment, the user creates a title for the stratum and selects keywords in the Logger window. When the descriptive attribute is no longer in effect, the user clicks the strata line to turn it off.

We are investigating how text and color can enhance the readability of such graphs. Two different x-axis scales are used to display the stratagraph: the real time scale and the "scale of difference." For the real-time display, the scale of the x-axis is time code (frame numbers for laserdiscs). This scale is well suited for browsing through video material. A current frame line (not shown) passes through the stratification graph in sync to the video displayed on the "gctl." The scale of difference is a compressed graphical representation of content where only changes in descriptive attributes are displayed.

When the user "rubs" the statagraph with the mouse, detailed contextual information for those frames are displayed in an associated Strata data window. A strata rub generates a report of all file strata that the frames are embedded in. The Stata data window is also used with the Infocon module.

2.3. Making Sequence Annotations: Infocon

During this phase of production, shots are delimited by in and out point and annotated so that they can be played independently. To create a new sequence the user inspects the Stratagraph for material with the desired content to select a shot, reviews the footage in GCTL and annotates these shots in the Infocon module. The Infocon (INFOrmation iCON) module consists of text widgets for annotations, buttons to grab frame numbers from Galatea, and video windows to display digitized in and out points. When Infocon retrieves the in- and out-frame number from Galatea, the associated frames are digitized and displayed in video windows on the Infocon module.

Although in and out point are important cues for visual continuity, they might not provide an adequate representation of the shot content. In addition to the in and out frame, a content frame is digitized and its associated frame number is also stored with the data recorded for the shot. The content frame is used to create an Infocon which serves as a visual representation of shot content. An Infocon is composed of a video window and a title bar which displays the annotation associated with the shot that it represents. Left-clicking in an Infocon plays the shot in the GCTL module. Middle clicking on an Infocon displays a pop-up menu which tells which volume name, in and out points, trim button, delete button and an exit button to hide the pop-up. The in and out frame for a given shot is only displayed in the Infocon module while the Infocon for each shot is displayed in its own window. Infocons for each shot are browsed and arranged from left to right on the screen. Ordered groups of Infocons are viewed as a sequence. Later, they can be reordered and played back. If needed, in and out points may be trimmed.

When a satisfactory sequence is arranged, annotations, volume-name, in and out frame numbers are saved as a record for each shot. These records are then saved to an ASCII text file. These sequence files can then be loaded into the Stratagraph and the strata for the annotations are displayed with the original logging data. In this version of the Stratosphere application, only annotations and their associated in and out and content frame numbers are stored. The digitized video images are generated on the fly, when a particular sequence file is loaded into Infocon.

In figures 2 and 3, we illustrate how a virtual annotated sequence is created. In figure 2, the user inspects the Stratagraph to select shot content. The in- and out-points are digitized by the Infocon

module and a content Icon is created. In figure 3, these shots are arranged to create a virtual sequence.

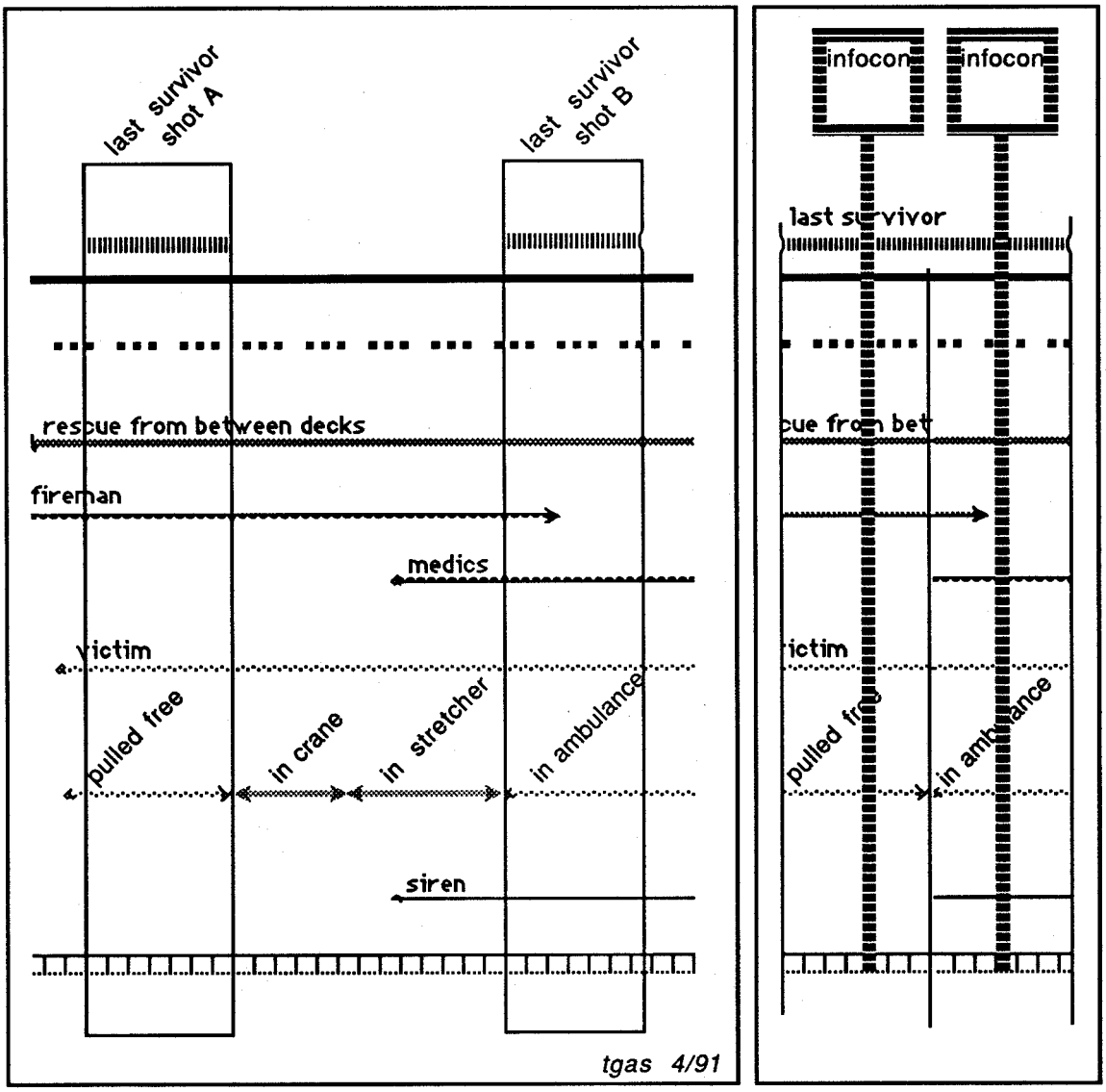

Figure 2: Shot selection for "Last Survivor" Sequence
Figure 3: Display of Virtual Annotated Sequence - Last Survivor with Infocons

The bold-dashed line in Figure 3 is an icon-line. The description of the shot can be derived by examining the strata that this line passes through.

When viewed, the virtual annotated sequence appears to be made up of adjacent shots; but as we can see in figure 2 this does not necessarily have to be the case. The stratagraph is used to graphically represent both logging information and annotated virtual sequences.

examining the strata that this line passes through (Table 1)

Table 1: Content for Infocons in Figure 3

Infocon A: Last Survivor A	Infocon B: Last Survivor B
1989 San Francisco Earthquake	1989 San Francisco Earthquake
880 freeway collapse	880 freeway collapse
rescue from between decks	rescue from between decks
fireman	medic
victim	victim
pulled free	in ambulance
	siren

In the scenario just described, the user manually selected the shot to be used in the "Last Survivor" sequence. Infocon can also generate icons on the fly for any set of strata provided that the data file is in the correct format. A file of annotations can be parsed using UNIX commands such as "grep," "sort," and "awk" and then sent to Infocon to generate icons.

2.4. Soft Story Templates

The ASCII data files created by the Logger and Infocon modules are represented on the stratagraph; they can also be "piped" through soft story templates to generate mini-video programs. An example of a soft story template is a news segment generator called ACE. ACE parses strata lines to find contiguous sets of frames which satisfy search and format requirements and then arranges them into sequences. ACE has a rule to construct a news sequence which consists of:

- a studio shot as an introduction;
- a predetermined number of action shots;
- a reporter in the field;
- a wrap up studio shot.

3. Audio Applications: Editing Assistance Through Cinematic Parsing

The shot, generated by the camera, is inherently associated with the circumstances and intentions which were in effect as it was being recorded. A similar case can be made for synchronous sound: as with the camera, the microphone/sound recorder mediates the environment in which it is situated.

Ambient sound offers us clues about the physical and geographic aspects of place. When parsed, the original synchronous sound track for any movie may help us segment the video/audio stream into shots or help us determine significant action points. Audio tracks of home movies are essentially composed of the same elements as all movies: ambient sound and meaningful dialog. Yet it is in the ambient sound where we find the most interesting distinction. In narrative movies, this element plays a secondary role: most ambient noise is undesirable, and is eventually eliminated. In fact, many re-shoots are caused by unwanted changes in ambient noise, like a plane flying overhead. This is mainly true due to the fact that such noises are extremely difficult

to edit. The only type of ambient noise generally used is what's called "room tone," which is used to create the mood of the space in which the action is supposed to occur.

We take another approach: ambient sound can provide us with important contextual clues as to what is going on in the frames. Ambient sound in raw footage has all its original components present, from cars going by to hand adjustments on the camera. Room tone becomes more than just ambiance: it becomes the specific attribute of the time and space in which the shooting is being done. Several audio techniques can be used to obtain the information needed to extract said attributes.

3.1 Techniques Used to Parse Audio Tracks

Statistical properties of the audio signal, like mean and mean-squared, can be computed for analysis purposes. They can tell us where changes in the power of the signal occur, which can clue to drastic changes in background noise levels. Using domain-restricted knowledge, such processing can give an indication of the type of changes happening: a rapid increase followed by a rapid decrease is usually a thump, caused by someone shutting a door, or dropping something heavy on the floor, or grandma falling off her chair.

We use several techniques to analyze the spectrum of the signal. Of interest are average spectrum and average amplitude (sum of mean-squared spectrum). These techniques give us information useful to determine changes in the power distribution and the spectral content of a signal. These techniques are used in the following way: we create a template with the spectral characteristics of a sample window, which is then compared to the following window. If there are considerable changes in the power distribution, then we assume that the spectral content of this second window is different, thus capturing changes in the nature of the signal, not changes in the amplitude. A combination of both time- and frequency domain analysis should suffice to capture enough changes in signal content to find the shot boundaries.

Using audio as an indication of patterns of content could prove quite useful. The audio track can provide information that otherwise would be extremely difficult to parse. Take, for example, a funny incident the family is gathered in the living -room, and grandma fell off her chair. As stated earlier, there's no way of telling whether this was just a thump, or was it something more interesting. One way to do this is to use laughter. If it is possible to parse laughter, then we can tell that something funny happened, thus it is not desirable to remove it from the track.

Another interesting possibility is the development of templates, or "spectral signatures" of different sounds, to keep for later comparison. For example, assume that the system has stored the spectral content of a police car passing by. By checking through the sound track for occurrences of this spectral pattern, the system can tell that within a certain segment, there's a police car going by. This could lead, provided there was enough memory for storing patterns and enough computing power to perform the comparisons, to the complete automation of the logging process. This ties in into the stratified representation approach. Once a specific attribute is found, it becomes a part of the stratified description of the shot being examined. In the last example, a police car passing by was found on a particular shot This could then be incorporated into the strata graph, as a shot description: " A car passes by."

The converse is also true: the strata graph can provide information that, used with the "spectral signature" of the particular sound, can create new templates. For example, assume that we have a car on Pismo Beach. Parsing the audio track, we've found that the "car" template is present on this shot. We also know from the strata graph that this particular car is on the beach. We can then use the audio corresponding to that particular segment as "car on the beach" and save it for fine-grained parsing.

3.2. The Parser: Implementation

It is in the parser that the concepts and methods discussed above are implemented. The parser has two different modules: the digital signal processor, and the rules module. The digital signal processor does the actual parsing required. To achieve this, it performs several tasks, such as Fast Fourier Transforms (FFTs), Linear Predictive Coding (LPC), and statistical analysis. The rules module consists of a set of heuristics, which make certain "guesses" as to how to interpret the information extracted by the signal processing module. The parsing itself was done in different layers, each giving different information that could not be extracted from the others.

The first layer consists of time-domain analysis of the sampled waveform. The sound file is divided into chunks, each being the equivalent of a frame of video. The number of samples representing a frame (1470) is obtained by dividing the sampling rate (44.1 KHz) by the number of frames in a second (30 for NTSC video). The amplitude mean is then obtained for each block. Using heuristics based on observation of raw footage, some preliminary conclusions are then drawn from the observation of these mean values.

This second layer consists of frequency domain analysis, after performing a Short- Time-Fourier Transform on the sound file. Analysis done at this level consist of standard signal processing techniques described in an earlier section. These two analysis schemes detect changes in the energy distribution on all frequency bands, thus capturing changes in the nature of the signal, not changes in the amplitude. This second-level processing is only performed for cases that are ambiguous at the first level, meaning that a clear distinction cannot be made simply with amplitude changes.

Another interesting issue that comes up during this second layer of parsing is that of time duration of events, or granularity. How long does an event have to last in order to be considered a new unit, instead of a short occurrence within another? This is perhaps the most interesting feature of the parser: it takes the lowest average amplitude for a sample window as the background noise level. If this is only the background noise level, then no segment can have lower amplitude; it must be the absolute minimum. Thus, if the overall signal level eventually comes back down to the saved minimum, then the background noise level has not changed.

3.4. The Parser: Applications

For example, let's assume that the mean increases sharply, to then decrease as sharply as it increased after a short period of time (a "spike"). The heuristic in this case would be that this was a thump, caused by a slamming door, an accidental tap on the camera microphone, or a loud

knock. Another example would be a smooth increase followed by a smooth decrease. The heuristic for this case could yield several results: a passing car, an airplane flying overhead. None of these would qualify as a shot change, since they are transitory changes in background noise levels. There are other cases which could possibly be shot transitions, but the information available at this stage of processing would be insufficient. For example, a smooth increase in average amplitude. This could mean that a car drove into fu1me and stopped, that someone turned on a noisy device (like a fan or a computer), or it could also represent a smooth pan (which could mean a shot transition). In these cases, a second layer of processing would be required.

The following figure shows an actual segment of the audio track taken from a home movie. It is a simple amplitude vs. time representation.

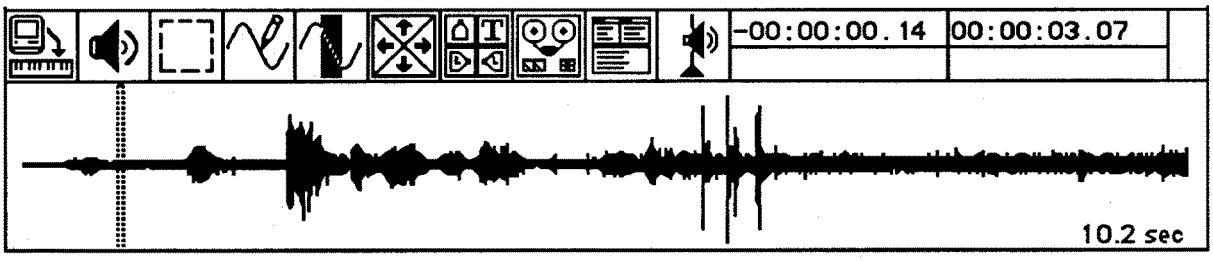

FIGURE 4: Soundfile representation- Amplitude vs. Time

The first four changes in amplitude are representations of human speech, while the four spikes towards the end represent hand adjustment after a shot change. The transition from one shot to the other is in between the two groups. The figure that follows shows how the parser interprets this audio segment.

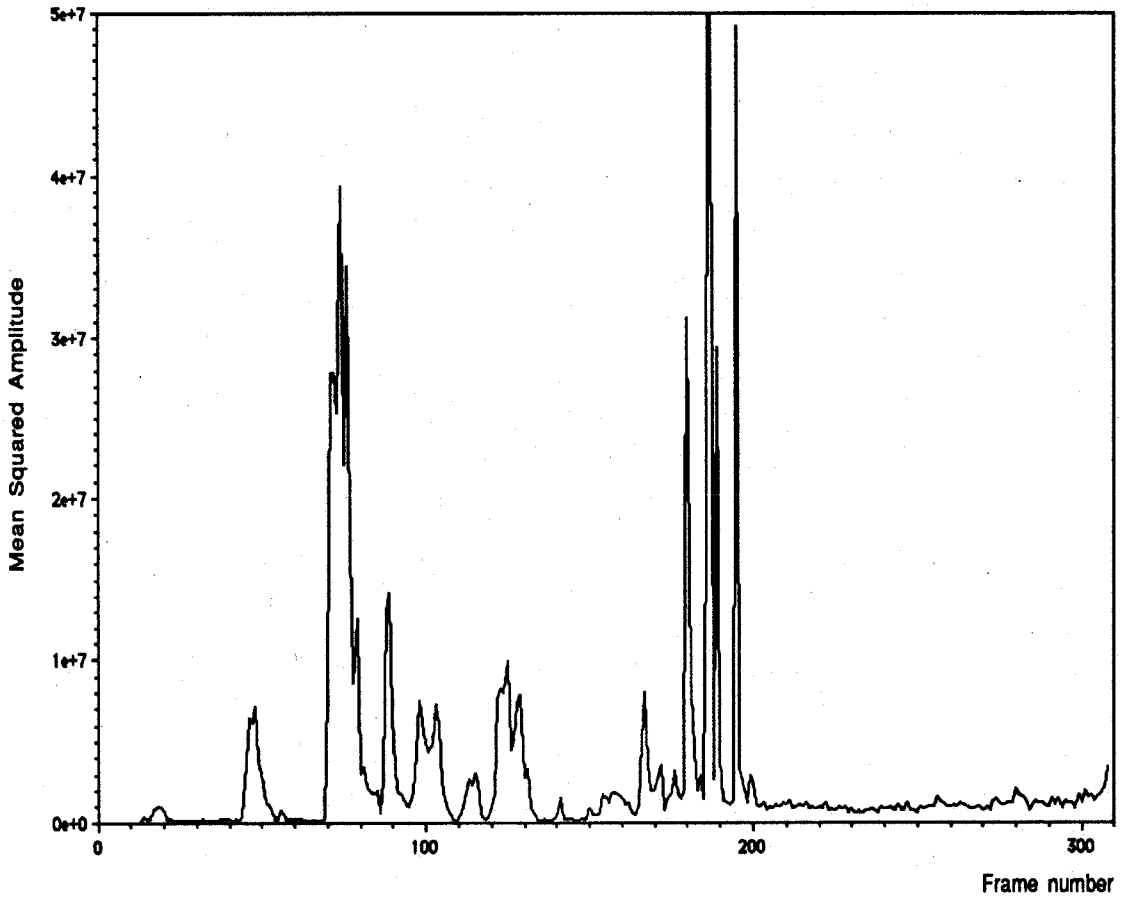

Figure 5: Parser Interpretation Shot Transition Soundfile

This shows the first pass at the parser: the amplitude layer. Mean-squared amplitude is shown on the vertical axis, while the fuune number is shown on the horizontal. The parser will eliminate the speech part, using the timing constraints as well as the fact that the amplitude eventually returns top it's original level. This will cause the parser to select the region between frames 160 and 170 as the one with a possible transition. Second level parsing will then show the change occurring at 165 fuunes, using average spectral amplitude differences.

4. Conclusions: Parsing Video In Context

Moving images, sound, and text are the base components of complex multimedia systems. We define these media as resources which can be used simultaneously in multiple contexts and applications.

A context-based representation is robust enough to describe video resource and sound resources that can co- exist in many different contexts. Without knowledge about the context, descriptions become ambiguous. Another problem caused by such description is the lack of an elastic representation. Elasticity pertains to the use of multimedia data: how can various individuals who will have access to the same set of multimedia data, satisfy the information requirements of their own needs? For example, some individuals will want to use all 5 minutes of a videotaped

interview while others will only want the last 20 seconds of the interview. Others still will only want one frame of the interview to use as a picture in a newsletter.

The Stratification methodology breaks down the environmental factors into distinct descriptive threads called stratum. When these threads are layered on top of one another they produce a descriptive strata from which inferences concerning the content of each frame can be derived Applications are being built that apply this methodology to logging and editing.

Contextual representations of audio are based on the ambient sound present in raw footage. This ambient sound becomes the specific attribute of the set of frames being examined. Through the use of heuristics, statistical analysis, and signal processing, changes in ambient sound can be parsed. Further application of these ideas can lead to the automation of the logging process through the use of templates, provided enough computing power was available.

Elasticity in description and searching is especially important when authoring a multimedia system. Our research points to new ways in which the user can interact with multimedia systems by varying the length of various chunks of video on demand. The representation gives the user maximum flexibility in manipulating multimedia resources. The value of each type of description, both sound and image, is enhanced by the other. This will lead to an integrated representation of the image and sound descriptions.

Acknowledgements

The authors would like to thank Joshua Holden, Kristine Ma, Lee Morgenroth, Carlos Reategui, and Derrick Yim for their assistance on implementation details; Hiroshi Ikeda, research affiliates from Asahi Broadcasting Corporation, for his work on Infocon; Mike Hawley, for DSP ideas and useful comments; and Glorianna Davenport, for comments, support, and guidance.

Bibliography

Aguierre Smith, Thomas G. "Stratification: Toward a Computer Representation of the Moving Image," working paper, Interactive Cinema Group, MIT Media Lab, January 1991.

Applebaum, Daniel. "The Galatea Network Video Device Control System, MIT, 1990.
Davenport, G. and W. Mackay. "Virtual video editing in interactive Multimedia Applications," Communications of the ACM, July 1989.

Davenport. Glorianna, Thomas G. Aguierre Smith and Natalio Pincever , "Cinematic Primitives for Multimedia: Toward a more profound intersection of cinematic knowledge and computer science representation." to appear in special issue of IEEE Computer Graphics and Applications on Multimedia, to be published Summer 1991.

Ellis, Dan. Some software resources for digital audio in UN/X, Music and Cognition Group technical document, MIT Media Lab, April 1991.

Handel, Stephen. Listening, MIT Press, Cambridge, 1989.

Ludwig, L. and D. Dunn. "Laboratory for Emulation and Study of Integrated and coordinated media communication." Proc. ACM SIGComm, 1987.

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

Purcell, P. and D. Applebaum. "Light table: an Interface to Visual Information Systems," in Electronic Design Studio, MIT Press, Cambridge, June 1990.

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