Next Generation Network and Operating System Requirements for Continuous time Media

Scott M. Stevens

Software Engineering Institute
Carnegie Mellon University

Abstract

Accessing massive multimedia databases will require multiple representations of those databases. Initial access may be through visual representations of the database. However, traversing numerous levels of tree-like structures will quickly find the user lost. Simple database queries may overwhelm users with information.

To overcome these problems the Advanced Learning Technologies Project at Carnegie Mellon University's Software Engineering Institute embeds in multimedia objects the knowledge of the content of those objects over several dimensions. With this model, variable granularity knowledge about the domain, content, image structure, and the appropriate use of content and image is embedded with the object. In ALT, a rule base acts as a visual director, making a judgement on what image to display and how to manipulate it. This provides the ability to present disparate text, audio, images, and video, intelligently in response to users needs.

It is difficult to move through information that has an intrinsic and essentially fixed temporal element such as video. While detailed indexing of video can help, users often wish to peruse video much as they flip through the pages of a book. Two techniques developed for this project will facilitate such searches. First, detailed, embedded knowledge of the video information will allow for scans by various views, such as by content area or depth of information. Second, partitioning multimedia data into smaller objects reducing bandwidth problems associated with accessing central data in large video files. Concatenation of logically contiguous files allows for seamless, continuous play of long sequences

A. Introduction

The volume of information becoming available to the user through continuous time media is such that it can no longer be assumed that the human-computer system is composed of an intelligent user accessing tractable amounts of static information. The new model must be one of intelligent-dynamic information aiding intelligent users in entertainment, learning, and working tasks.

The Object Lens and Athena Muse projects at the Massachusetts Institute of Technology, the Information Visualizer project at Xerox Palo Alto Research Center, the Course Processor project at Carnegie Mellon University, and the User Interface Designer's Assistant and the Advanced Learning Technologies projects at CMU's Software Engineering Institute have investigated solutions to the problems of storing, accessing, and visualizing multimedia information. Tasks that involve huge information spaces overwhelm both users and today's electronic workspaces. Accessing massive multimedia databases will require multiple representations of those databases. Initial access may be through visual representations of the database. However, traversing numerous levels of tree-like structures will quickly find the user lost. Simple database queries may overwhelm users with information.

To overcome these problems we embed in multimedia objects the knowledge of the content of those objects over several dimensions. Much current work treats multimedia objects, especially video, more as text with a temporal dimension [1,2]. The Advanced Learning Technologies (ALT) project at CMU's Software Engineering Institute has developed a multidimensional model of multimedia objects. With this model, variable granularity knowledge about the domain, content, image structure, and the appropriate use of content and image is embedded with the object. These often orthogonal descriptions of an information database promote both usability and accessibility. For example, in ALT an expert system acts as a visual director, behaving intelligently in the presentation of images. Based on a history of current interactions (input and output) the system makes a judgement on what image to display and how to manipulate it. This provides, for the first time, the ability to present disparate text, audio, images, and video, intelligently in response to users' needs.

B. Constant Rate Continuous Time (CoReCT) Media

Many researchers have noted the unique nature of motion video. Differences between motion video and other media, such as text, are typically attributed to the temporal nature of video. Every medium has a temporal nature. It takes time to read (process) a text document or a still image. However, each user does this at his or her own rate. Often it is possible to assimilate visual information holistically, that is to come to an understanding of complex information all at once. The creative process is similar, and need not be restricted to the visual domain. Mozart said that he conceived of his compositions not successively but in their entirety [3]. Clearly time is an intrinsic part of music. While Mozart may have "seen" his whole composition at once, the temporal aspect of a piece must have been present in that same instant. Subjective and real time may be similarly contracted or expanded, by the user, while reading text or viewing an image.

If one likens the scrolling of text to viewing a motion video sequence we see the real difference between video and audio and other media. For example, if a system fixed the scroll rate of text for a mythical average reader, say at 400 words per minute, a reader that read at even 401 words per minute would soon be out of synch with the text. Even a 400 word per minute reader would undoubtedly find passages that, because of complexity or interest, require a much slower rate. Obviously, no one would argue that the scroll rate should be fixed. But video and audio must be played at a constant rate, the rate at which they were recorded, to have almost any meaning at all. A user might accept video and audio played back at 1.5 times normal speed for one ore two minutes. Even if digital signal processing is used to maintain a near normal audio tone, it is unlikely that users would accept long period of such playback rates. Moreover, the information transfer rate would still be principally controlled by the system.

The real difference is that video has a constant rate output that cannot be changed without negatively impacting the user's ability to extract information from the stream. Video is a Constant Rate Continuous Time (CoReCT) medium. Its temporal nature is constant due to the requirements of the viewer/listener. Text is a Variable Pace Continuous Time Medium (VaPid). Its temporal nature only comes to life in the hands of the users.

Documents are, hyperlinks aside, also continuous. The difference between a text document and motion video or audio is that the constraints on scheduling and synchrony are much more severe for video and audio. User's performance can be adversely affected when text is presented at less than fifteen characters per second [4]. With video, users will not accept a wait of over one thirtieth of a second (or, bowing to international pressures, one-twenty-fifth of a second). A user may accept a few second delay between the appearance of a page of text and an associated but separately stored table or image. If the synchrony between video and audio is violated by more than about a fifteenth of a second users will both note this difference and find it

unacceptable. But the continuity and synchrony differences between video plus audio and text are ones of scale, not of kind.

The differences between CoReCT and VaPid media become critical when a user is searching for information. The human visual system has an aptitude for quickly looking at an image or a page of text and finding a desired piece of information while ignoring unwanted information (noise). This is what makes flipping through the pages of a book a relatively efficient process. Even when the location of a piece of information is known a priori from an index, the final search of a page is aided by this ability. But the motion video analog of this process, fast forward, is not nearly so efficient.

It is difficult to move through information that has an intrinsic constant rate temporal element such as video. While detailed indexing of video can help, users often wish to peruse video much as they flip through the pages of a book. Current methods make this a difficult process. Analog videodisc scanning, jumping a set number of frames, may skip the target information completely. To be comprehensible, a simple scan such as a VCR's fast forward often takes too much time. Network and even bus bandwidth problems aside, displaying motion video at twenty times normal rate presents the information at an incomprehensible speed. And it would still take three minutes to scan through a one hour video!

Even if the visual system could make sense of such accelerated motion, a short two second shot would be presented in one tenth of a second. With human and system reaction times potentially adding to a second or more, significant overshoots will occur.

Audio is of no help. Beyond 1.5 or 2 times normal speed audio becomes incomprehensible. At faster playback rates frequency shifts make it inaudible. Even digital signal processing techniques to reduce the frequency shifts associated with high speed playback of audio fail at three or four times normal rate. With about 150 spoken words per minute one hour of video contains 9,000 words, which is about 15 pages of text. If the information being sought is from "talking head" video, such as a lecture, or worse yet from audio only, a comprehensible high playback rate of 3 to 4 times normal speed is a totally unacceptable search mechanism. Assuming the target information is on average half way through a one hour video file it would take 7.5 to 10 minutes to find. No user today would accept a system that took 10 minutes to find a word in 15 pages of text.

Three techniques are proposed for facilitating such searches. First, detailed, embedded knowledge of the video information will allow for scans by various views, such as by content area or depth of information. Second, video is partitioned into many separate small files. First pass searches retrieve a small segment of one to two minutes of video. Bandwidth problems associated with the transfer of large video segments are eliminated. Continuous play of a long sequence is accomplished by seamlessly concatenating logically contiguous files. Third, scans can be performed by changes in visual information, such as by scene change. Most compression schemes will permit easy analysis of scene change since major image transformations affect the compression and decompression algorithms. In cases where this is not efficient, embedded knowledge about the content of individual frames can substitute. Much like noticing chapter and section headings in a book while flipping pages, all three of these types of scans will permit information based scanning of CoReCT material.

C. ALT Paradigm for Use of CoReCT Media

It is obvious that both the type and the volume of information accessible by computers is growing at an exponential rate. What is less obvious is that much of this information should be available from different viewpoints (from different information types). A reader of a Shakespearean play will glean a perspective unavailable to a viewer of a video of that same play. Likewise the viewer gains a different perspective than the reader, especially if the viewer can

compare more than one director's and actor's interpretation of that same play. What is even less obvious is that different descriptions of a multimedia database are necessary to promote usability and accessibility.

Current multimedia operating system descriptors are little different than for a graphics file. The information necessary for use of the data includes traditional "file control block" information such as the name (usually of the scene or story), the size of the file in bytes, the creation date, and the last modification date. The type of media should be available (e.g. audio, video, or video with audio). Lastly, media-specific descriptors must be recorded. For audio these include, sampling rate, filtering, attenuation, and compression algorithm and for video these include frame rate, position, hue, saturation, and compression algorithm used.

With this information an application that knows which video or audio segment is required can perform basic functions. For video with audio these would include, play, setting the position, clipping or scaling, if need be performing color corrections, and adjusting volume and tone. These functions form the complete set used by the typical multimedia application of today and also is the limit of most system support for digital multimedia.

Beginning in 1987 the Advanced Learning Technologies Project (ALT) at the Software Engineering Institute of Carnegie Mellon University began developing a radically new paradigm for interactive multimedia [5]. For more sophisticated applications such as ALT, additional descriptor requirements include, abstract representations of the data (as opposed to simply the physical representations of the data) and information identifying any procedural actions to take following access of the data (i.e., control information).

The goal of the ALT Project was to create a virtual workplace where a user could learn about and experience a software engineering technique called an inspection. Inspections are structured meeting where development workproducts are systematically examined to identify defects. Users are placed in a virtual building where they may walk to, amongst other rooms, a multimedia library, their own office, and a conference room where they may participate in a simulated inspection.

In the conference room the user sees three members of the inspection team enter the room, sit down, and begin the discussion. The user may talk to the video personae at any time through a natural language interface, or the user may simply sit back, look and listen. Of course, as in a real meeting, the personae will sooner or later ask the user to participate. This is a high fidelity simulation of a group discussion. The traditional multimedia/interactive video paradigm is play a video sequence, stop, wait for user input, branch. A more appropriate term for this in "interrupted video."

Consider a hypothetical simulation of a conflict between two parties, used to teach mediating skills. In the canonical interrupted video version the output data may consist of two conversations, one for each party. The student can see the two conversations played in sequence, or may see only one. This results in four possible permutations, as shown in figure 1. If on the other hand each of these conversations is broken up into smaller, meaningful pieces, then there are numerous possibilities to alter the information presented by the simulation according to the student's mediating actions (input). The conversations might be partitioned into expressions of emotion, position statements, monologues supporting a position, and conclusions. Depending on the student's effectiveness as a mediator, conflicts may be resolved and compromises made. The finer granularity provides for the ability to create a much richer, more highly interactive simulation than with a simple interrupted video paradigm.

Partitioning video into small files can facilitate the creation of high fidelity simulations. However, as the complexity of the simulation increases, there exists a greater need to augment audio and video building blocks with descriptive data to make them usable by the simulation. The ALT system uses this paradigm to implement a code inspection simulation.

The ALT system simulates four participants and the interactions between them in an inspection. ALT provides the ability for the user to take any role in the process with the system simulating the other three participants. This is achieved, in part through a rule-based expert system that was developed to model the participants. The expert system defines the "personalities" and controls the dialogue between the simulated members of the inspection team and the user. Throughout the simulation the expert system continuously composes the video and the audio.

The ALT system uses Intel's Digital Video Interactive (DVI). DVI permits up to seventy-two minutes of full screen, full motion video to be stored on one compact disc. The simulation of the inspection requires approximately ten hours of audio and two hours of motion video with audio. We are able to increase the apparent storage by saving fractional screen images and composing the full image from still images plus motion sequences. In analog video, images are fixed during production and post-production. For practical purposes, the image on a videodisc or video tape cannot be altered significantly during playback. If the information of interest takes up only a fraction of the screen one full frame of analog video storage must still be allocated. Since images are stored digitally in DVI neither of these limitations is true. We are able to compose the visual image during playback (see figure 2). Since we wished to store the images of the actors separately we save significant storage space by saving only the section of the images that is of interest.

The rule base consists of over 8,000 lines of OPS/83 code implementing approximately 200 complex rules and associated procedures. It makes decisions in areas such as who should speak, the tone they should speak with, the content of what is to be spoken (context space search), and who is the persona speaking to. The rule base also controls the conversation and models the personalities of the participants.

The rule base uses the ten hours of audio, two hours of video with audio, and several thousand still images to dynamically compose the scene. The audio and video alone consist of over five thousand objects. All of these objects are organized in a multi-dimensional structure.

The rule base composes the visual and spoken dialogue from a four dimensional multimedia database (see figure 3). The first dimension is the topic under discussion (context space). The second dimension is the speaker (persona). The third dimension is the specificity within the topic. This is related to the temporal aspect being modeled in the conversation. (As people speak on a topic, they tend to get more specific and build on what was said previously, much as a text builds on previously chapters.) The fourth dimension is affect (emotion).

Abstract information about objects in ALT varies in granularity from scene headers describing information that is globally constant for a set of frames, to frame headers, information that is local to a single frame. Figure 4 suggests one digital file representation to implement both scene and frame headers for multimedia objects. In the current implementation, descriptions cannot be tagged to individual frames of motion video. This limitation was imposed by prior operating system limitations of the DVI environment. However, for still images and files of multiple still images that are animated to create a limited motion video, information at the frame level is available.

Figure 5 depicts the scene header information for the each of the video objects available to the rule base. Along with traditional "file control block" information, embedded in each object is information on camera angle, field of view (both objective points of view), character, topic, specificity within topic, tone, who the scene is addressed to, opinion, emotional subject, discussion resolution status, pointers to other topics, and gesticulation (all subjective points of view). The objects under the subjective points for both scene and frame (in the image database) header define the dialogue element's location in the four dimensional space. The rule base determines the points in this space which are needed to compose the appropriate dialogue.

An interesting part of the rule base is called Hitchcock, the visual director of the system. With digital video, we can manipulate the images when we play them back. In fact, for the ALT system we must in order to compose scenes with different personae in them. We want the expert system to behave intelligently in presenting these images, much like directors do today when shooting a scene, or when they edit it. But in ALT, the system does it during playback. In the simplest case, Hitchcock must determine who the speaker is addressing and display appropriate images (i.e. the actor must be looking at who he is talking to). More interestingly, if the user is dominating a conversation, the system may present a camera angle of the participants on the screen which is slightly lower. Years of experience and many studies have shown that images such as this tend to portray the viewer as more dominant [6]. Figure 6 three shows the scene and header information for each image.

In ALT there is a surrogate travel interface which allows the user to walk around the workplace and move into and out of a variety of rooms (see figure 7). The image sequences for this are contained in one scene, along with the header information shown in figure 8. With the addition of subjective information it becomes easier for the application to know what room the user is near and thus what rooms are accessible. This paradigm gives the application detailed information about the virtual world by associating that information with the data. Updates to the data, such as newly accessible space, are automatically passed to the controlling application.

Since the ALT Project began, other researchers have discussed related schemes, defining cinematic primitives for use by multimedia systems [7]. Unfortunately their current implementations use analog videodiscs as video sources [8]. In these systems, areas of interest are, at least in part, identified by frame number, even to the point where the user may need to use frame numbers for video access. With 54,000 frames on a single half hour videodisc it is hard to see how users can effectively interact with the video unless a form more abstract than frame number is used.

Having the video and audio in digital form is crucial to the ALT paradigm. The descriptions of an information database, being closely tied with the information, should remain with the information data rather than be implemented in an application working on the information. Abstract descriptions of multimedia greatly enhance usability for both applications and users.

D. Future Directions and Conclusions

The current implementation of ALT could be significantly improved with additional digital video operating system functionality. The principle added functionality needed is the ability to access multiple data streams from single or multiple sources. In our system, this capability would allow more than one reviewer to move at once and more than one audio stream to be played at the same time. With the addition of fully integrated scene and header information image databases could be stored as a series of motion video files. This header information should not be overly specified, e.g., a restrictive maximum size should not be enforced. These capabilities would provide better visual flow, improve flexibility, and save storage via video compression. The overall objective of these needs is to be able to manipulate multimedia data dynamically in different ways.

An improved simulation scenario might be the following:

> Suppose that in the middle of a comment one of the simulated personae, Andy, directs a caustic remark toward one of the other personae, Michelle. Andy finishes neutrally by asking for further comments from the other people in the meeting.

The rule base would like to:

- Play the audio for this dialogue.
- Show video images of the Andy looking frequently at the code but occasionally at the other reviewers.
- At the appropriate time (noted by audio scene's frame headers), show video images of the Andy looking at the Michelle and being aggressive.
- Simultaneously, show reactions of the Michelle.
- Conclude with video images of the Andy looking at each reviewer

Even without these added capabilities, the ALT paradigm provides for:

- The ability to associate common information to a set of frames via a scene header, and to differentiate frames according to information in a frame header
- The ability to access this header information quickly without having to process the actual scene or frame data associated with this information
- The ability to output selected frames from a given scene without any display artifacts from the unselected frames, and with minimal timing constraints

With ALT paradigm, variable granularity knowledge about a domain, content, image structure, and the appropriate use of content and image is embedded with the object. These often orthogonal descriptions of an information database are shown to promote both usability and accessibility. This provides, for the first time, the ability to present disparate text, audio, images, and video, intelligently in response to users' needs.

Acknowledgements: I would especially like to thank Michael Christel who is the software engineer on the Advanced Learning Technologies project and contributed significantly to every phase of the project. His review of this article and help in preparations of its figures was invaluable. This work was sponsored by the U.S. Department of Defense

References

1. "A Construction Set for Multimedia Applications," Matthew E. Hodges, Russell M. Sasnett, and Mark S. Ackerman, *IEEE Software,* January 1989
2. "Intermedia: The Concept and the Construction of a Seamless Information Environment," Nicole Yankelovich, Bernard J. Haan, Norman K. Meyrowitz, and Steven M. Drucker, Brown University, *IEE Computer*, January 1988
3. Hadamard, J., *The Psychology of Invention in the Mathematical Field.* Princeton University Press, 1945
4. "The Effect of VDU Text-Presentation Rate on Reading Comprehension and Reading Speed," Jo. W. Tombaugh, Michael D. Arkin, and Richard F. Dillon, Proceedings of ACM CHI '85 Conference of Human Factors in Computing Systems, 1985
5. "Intelligent Interactive Video Simulation of a Code Inspection," Scott M. Stevens, *Communications of the ACM*, July 1989 Volume 32 Number 7
6. Kraft, R. Mind and media: The psychological reality of cinematic principles. In *Images, Information & Interfaces: Directions for the 1990's*, D. Schultz and C.W. Moody, Eds. Human Factors Society, New York, 1988
7. "Cinematic Primitives for Multimedia," Glorianna Davenport, Thomas Aguierre Smith, and Natalio Pincever, *IEEE Computer Graphics & Applications*, July 1991
8. "Parsing Movies in Context," Thomas G. Aguierre Smith and Natalio C. Pincever, USENIX-Summer '91, Nashville, TN

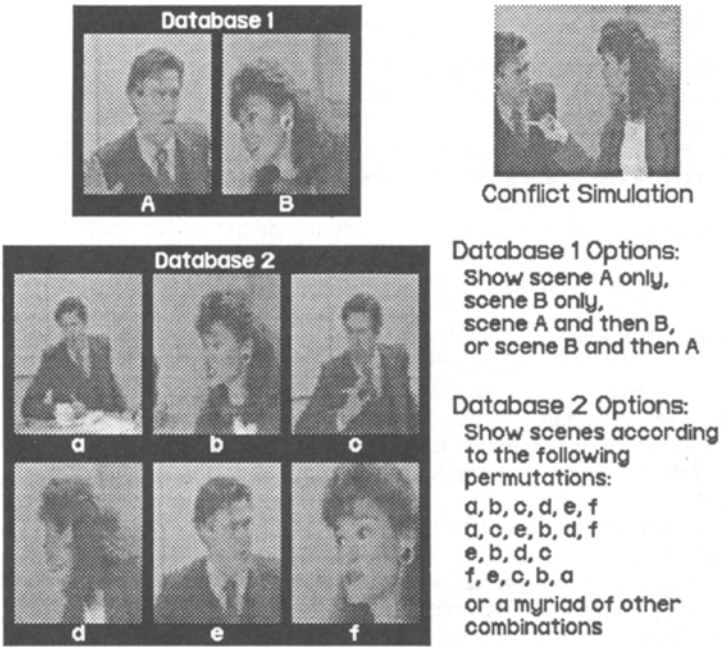


Figure 1. Effects of Granularity on Simulation Fidelity

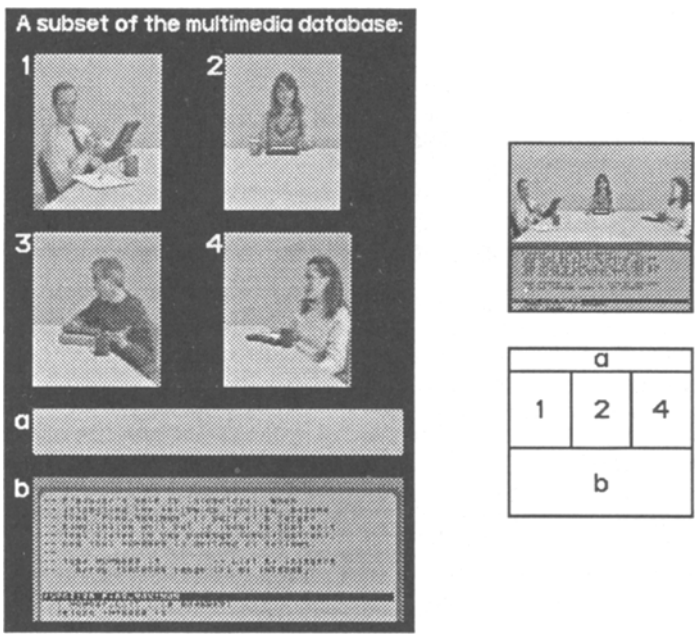
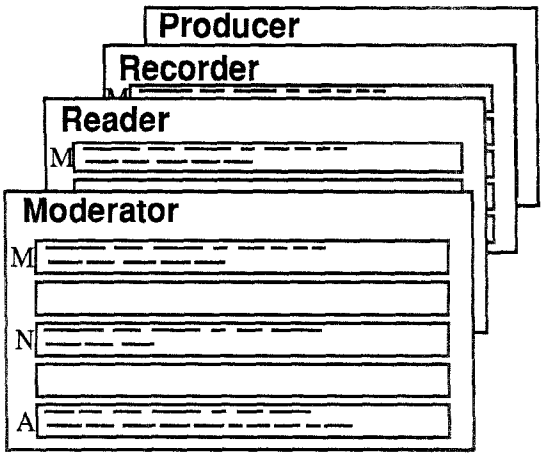


Figure 2. ALT Compositional Components



Individual cards represent the *speaker*, with areas within the cards indicating *affect*, e.g., mild, neutral, and aggressive. A stack of cards represents one temporal *level* of a topic, with each rectangular region below representing a *topic*.

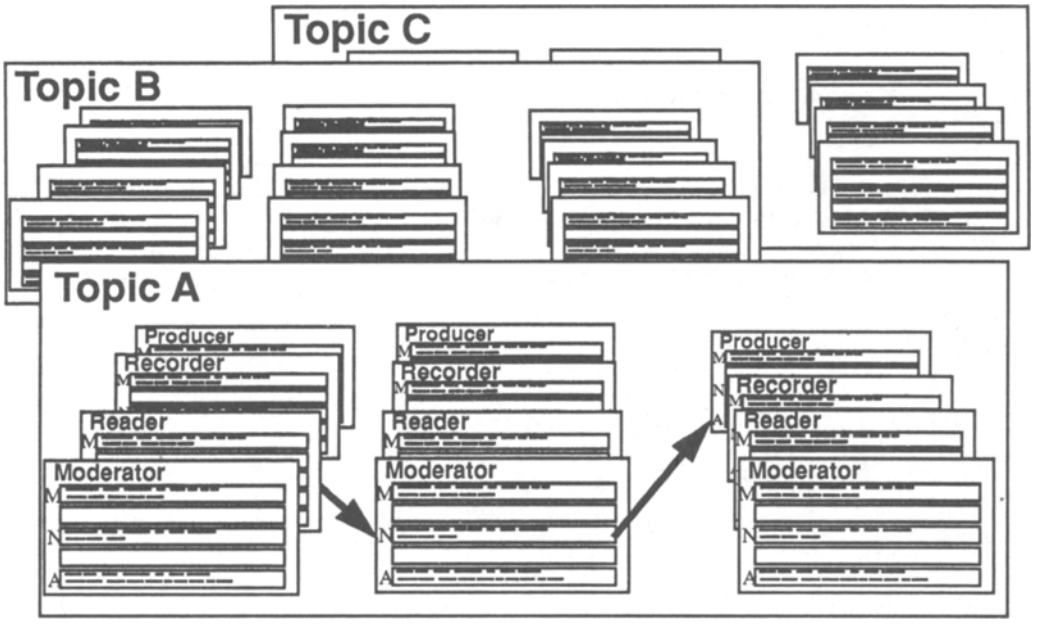


Figure 3. ALT's Four Dimensional Dialogue Structure.

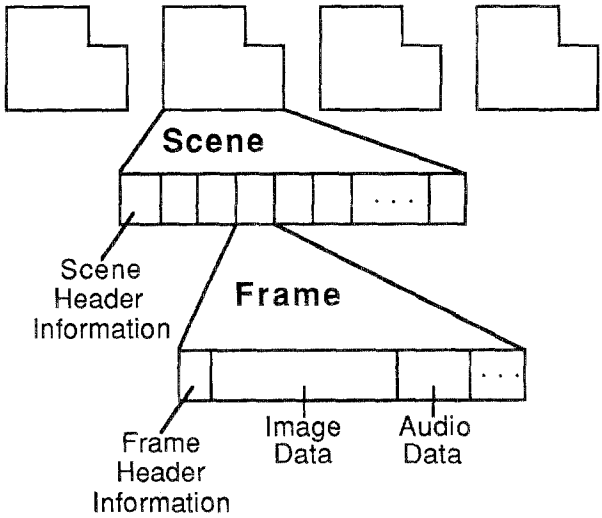


Figure 4. ALT file representations for scene and frame headers.

Scene Header for Inspection Video Dialogue:

- Traditional "file control block" information
  - Name of scene
  - Size in bytes of scene
  - Creation date of scene
- Media type (audio, video, audio and video)
- Media-specific descriptors
  - Audio (sampling rate, filtering, attenuation,...)
  - Video (frame rate, position, compression algorithm,...)
- Objective point of view
  - camera angle
  - field of view
- Subjective point of view
  - character
  - topic and specificity within the topic
  - tone
  - who the scene is addressed to
  - opinion
  - the personal focus (emotional subject)
  - resolved status
  - bridge to another topic
  - gesticulation

Figure 5. ALT video scene header information.

Scene and Frame Headers for Inspection Images

- **Scene header:**
  - Objective point of view
    - camera angle
    - absolute location
    - field of view
  - Subjective point of view
    - character
    - emotional level
- **Frame header:**
  - Objective point of view
    - hue and saturation
  - Subjective point of view
    - direction of gaze
    - gesticulation

Figure 6. ALT image scene and frame header information.

Figure 7. One still image from ALT surrogate travel interface.

Scene and Frame Headers for Surrogate Travel:

- **Scene header**:
  - Objective point of view
    - Field of view and depth of field (constant throughout the scene)
  - Subjective point of view
    - Geographic boundary information
- **Frame header**:
  - Objective point of view
    - Camera angle
    - Absolute location (x, y, z)
  - Subjective point of view
    - Near room(s)
    - Accessible room(s)
    - Indication of frame as a room entrance

Figure 8. Surrogate travel scene and frame headers.