Director's Workshop: Semantic Video Logging with Intelligent Icons

Bibliographic Reference:

Marc Davis. "Director's Workshop: Semantic Video Logging with Intelligent Icons." In: *Proceedings of AAAI-91 Workshop on Intelligent Multimedia Interfaces in Anaheim, California*, ed. Mark Maybury, AAAI Press, 122-132, 1991.

Director's Workshop:
Semantic Video Logging with Intelligent Icons

Position Paper For
AAAI-91 Intelligent Multimedia
Interfaces Workshop

Marc Eliot Davis
Media Laboratory
Massachusetts Institute of Technology
E15-488, 20 Ames Street
Cambridge, Massachusetts 02139
(617) 253-7674
mdavis@media-lab.media.mit.edu

Abstract - The Director's Workshop is an iconic interface for the logging of multimedia information in domain-independent, multimedia archives. The interface makes use of a temporally indexed, semantically structured representation of the content for multimedia information. Descriptive icons are quickly selected by cascading through icon hierarchies, and then organized into palettes for use in the logging process. When logging, icons are dragged from the icon palettes and dropped onto descriptive layers of the Media Time Line which displays the various content streams of the multimedia information in a vertical hierarchical organization along a horizontal temporal axis. The representation and interface are designed to make the logging process easier and more productive. The goal is to help users adequately describe multimedia information for later retrieval and resequencing by automatic presentation systems.

1. Introduction: The Need for Representations of Multimedia Content

If multimedia is going to become a new medium to think with, if we are going to be able to compose and communicate in multiple modalities (image, sound, text, etc.), an efficient and intuitive means of representing the content of multimedia information must be developed. We especially need ways of encoding our knowledge about the content of temporal dynamic media, like video and audio, which will allow human and computational agents to parse, process, and manipulate information in these (and hybrid) media forms. This paper outlines an ongoing research effort to create a representation of multimedia content, with a current focus on the problem of "data entry" or "video logging." In video logging the problems and complexities of creating a representation for multimedia become especially pressing: the potential tedium of the logging process functions as an acid test for the robustness, intuitiveness, and efficiency of any possible multimedia representation or interface.

For computers and video to be able to work together they must share a common representation of audio-video information. There is a wide range of applications which combine computers and video for which such a representation does not yet exist. Examples include: news archives, stock footage databases, video editing systems, multimedia learning environments, and film theory educational aids.

As an illustration of the need for representations of multimedia information, imagine that we are making a documentary about education in America. We are looking for a stock shot of a crowd of students walking into the front entrance of a high school building. The stock footage archive needs to be able to retrieve a shot which has specific content as well as certain formal characteristics (time of day, camera angle, lighting, etc.). The representation of the information must be able to address all aspects of the process: logging the footage into the archive, displaying information about it, retrieving it, and finally inserting it into its new context, the documentary to be made. Whether the final presentation is created by a human or automatically by a computer the representation problem remains the same. The question we must ask here is what information is required to describe multimedia content such that we can retrieve and resequence it.

2. Existing Paradigms: Logging and Editing

One may think to look to current practice in which multimedia information is logged for later recombination as in video and film editing. What one finds, however, is that the representations of information content rely heavily on the memories of the human agents involved in the process and that little more than in and out points, length, and a title or short description for a segment are recorded past the immediate needs of a given edit. To log and archive audio-video information for intelligent retrieval and recombination requires encoding the type of knowledge which editors use but do not explicitly represent in their practice. In logging the content of audio-video information certain criteria emerge:

- The representations of content must be a "temporally indexed", i.e., each bit of content information must be clearly marked as to its temporal extent correlated to the frame extents of the temporal medium. (e.g., "The crowd is in the frame" is true from frame 1023 to frame 1113).

- Semantic information must be detailed enough to allow for syntactic recombination, i.e., it should provide sufficient hooks to facilitate the embedding of the multimedia information in new contexts and sequences.

- Relations between segments -- containment, overlap, abutment -- must be represented to support automated inferencing and inheritance of semantic properties for intelligent retrieval and composition.

The type of basic level continuity information which human editors make use of in their editing decisions and which automated presentation designers will need to have access to includes: who (characters), what (objects), when (temporal location), where (spatial location), dialogue transcription, sound (dialogue, ambient, soundtrack, MIDI score), cinematographic information (camera motion and lighting), a textual description of the action, the gaze vectors of actors, and the screen positions (the 2-dimensional composition of the frame) and motions of actors and objects.* By associating these descriptive properties with segments indexed both temporally and semantically automated inference functions at proper levels of granularity. The question is still open as to precisely how much content information is enough (and in how much detail) to enable automatic presentation systems to retrieve and sequence multimedia segments. Through future work and repeated practical applications the adequate set of descriptors for multimedia content will become clear.

3. Multimedia Archives: Search, Retrieval, and Logging

The type of representation one needs for multimedia information will in large part be dictated by the applications in which that information will be used. Until now, multimedia archives have required relatively simple representations of multimedia information because their primary applications have been search and retrieval within small databases which treat video and audio information as the termination of a retrieval process rather than as the beginning of a composition process.

In the main, video, audio, and animations have been archived and retrieved as if they were non-temporal pieces of information which could be adequately represented by "keywords." A good example of this approach can be seen in Apple Computer's *Visual Almanac* which describes and accesses the contents of its archive by use of "keywords" and "image keys" (Apple Multimedia Lab, 1989). This technique is successful in retrieving matches in a fairly underspecified search but lacks the level of granularity and descriptive richness necessary for computer-assisted and automatic presentation design. The keyword approach is inadequate for representing multimedia content for the following reasons:

- Keywords do not describe the complex temporal structure of video and audio information.

- Keywords are not a semantic representation. They do not support inheritance and inference between multimedia segments.

Consider again our stock footage example. We want to find a shot of a crowd of students walking into the entrance of a high school. The first problem we encounter is that we may retrieve hundreds of shots which match a keyword query. How do we find the exact type of shot we are looking for? With keywords we have no way of distinguishing shots according to their temporal properties. Even with shots of the same length which are all continuous, smooth pans, the keyword approach cannot tell us which of these segments begins, at three seconds into the shot, to combine the pan with a zoom which ends with a close-up of the high school entrance door. Moreover, keywords are not a semantic representation of multimedia content.

With a semantic representation we are not restricted to finding exact matches to the labels used to describe content. We can also express relationships between semantic units in our query so that we could find, for example, a segment with a motor vehicle or an student population in the shot. In addition, we can use semantic inference mechanisms to help find shots matching our query. For example, in looking for a stock shot of a crowd of students entering a high school building a semantic representation of content could locate a shot of book-toting, high school age people entering a nondescript public building like a library or a museum. By making semantic inferences, the system could locate such a shot which is functionally equivalent to a shot of a crowd of students

* Gilles Bloch's slot-based representation of ACTION, SETTING, ACTORS, LOOKs, POSITIONs, and MOTION is a helpful first pass at this problem (Bloch, 1987).

entering a high school building. Keywords lack both the precision and the descriptive power of a temporally indexed, semantic representation of multimedia content.

Of course, text has the distinct advantage that it itself can be searched and even parsed. We do not currently have the corresponding ability to conduct "full-text" searches on, or to parse, multimedia information. If we are to have more sophisticated and finer grain search and retrieval in multimedia archives, and further, support for applications which automatically combine multimedia information into on-the-fly presentations, an equivalent of full-text searching will be required for multimedia information. In short, data structures and representation schemes for the content of temporal media are needed which would allow temporal media to be parsed, searched, retrieved, and recombined.

4. Formal Analysis: Content and Context

The reader many have noticed that in the second section's discussion of the content of audio-video information one of the journalist's 5 W's, "why," was left out. This omission is intentional and speaks to a larger question concerning automatic presentation systems and multimedia archives. The central metaphor for what an automatic presentation system does with its logged material is "bricolage." Bricolage is akin to bird's-nesting or the process by which objects are taken from their original contexts and used in new contexts for purposes other than those for which they were originally intended.

A multimedia presentation system will reuse and recycle information much in the same way that "stock footage" is used in documentary, news, educational, and avant-garde film. The function that a piece of footage performed in its original context is not insignificant to the function it may perform in new contexts, because knowledge of the original context contributes to the potential intertextual resonances which the information may evoke when embedded into a new context. However, it is important to emphasize that the function which a piece of audio-video information may have in a new context is not determined either by its original context or by any intrinsic meaning one may assign to it independent of the sequence into which it is embedded.

Early work in the analysis of context and order in film editing shows that what precedes and succeeds a piece of footage plays a dominant role in conditioning the meanings viewers will ascribe to it (Isenhour, 1975). Take for example a close-up shot of a man smiling. One might be tempted in semantically logging this shot to attach the label "happy" to it. Yet imagine that in a new context this shot is followed by a zoom-out shot revealing the smiling man to have a gun to his head, or alternately, a gun in his hand pointed at someone else. What is required, then, in the logging of semantic information is not a recording of the function of that content in its original context, but a detailed description of the formal characteristics of the information which condition its potential functions in new contexts.

Theorists of "reception-aesthetics" offer an extensive critical apparatus for the formal analysis of narrative texts (Iser, 1978) and film (Bordwell, 1985). The advantage of their approach is that they provide a theory of meaning creation and formal description which does not focus on what texts mean, but on how their structure guides the production of meaning in the text-reader/film-viewer relationship. By applying formal, structural analysis, as well as research in the semiotics of moving images (Metz, 1974), to the computational representation of the semantic content of multimedia information, we emphasize the descriptive features which enable the reusability and recombination of that content in new syntactic structures. Especially in the case of large, non-domain-specific multimedia archives, formal analysis provides the appropriate description of multimedia information such that segments become accessible chunks of content which a narrative generation system can retrieve and sequence in new contexts.

5. Multimedia Representation and Semantic Logging

In attempting to address the need for a temporally indexed, semantically rich description of multimedia content, our representation conceives of various "layers" of content information which allow for a detailed description of audio-video information over any given temporal extent. As in a musical score, information regarding various "parts" are horizontally organized and share the same temporal index in the vertical dimension. We can represent a given segment according to the following temporally organized and semantically structured descriptive layers:

Descriptive Layer	**Example Content**
temporal location	(late 1980's, Spring, afternoon)
spatial location	(Boston, high school building, outdoors in front of entrance)
characters	(crowd of male and female teenagers)
objects	(books, notebooks, bicycles)
action description	("A crowd of students are walking to class.")
dialogue transcription	(<crowd sounds>)
camera movement	(wide shot panning from screen right to screen left to entrance door)
lighting description	(ambient daylight on a clear, sunny day)
frame composition	(crowd in mid screen, high school building at screen left)

An advantage of this representation is that information is "good-until-cancelled": new information must be entered only when a state change occurs along the temporal axis. Levels of description inherit information hierarchically so that the system automatically knows that any information placed below, for example, a description of spatial location also occurs within that spatial location (see also Davenport, Pincever, and Aguierre Smith, 1991 for their discussion of "stratification"). Descriptive layers can be collapsed or expanded to hide or show levels of detail while preserving the temporal organization of the information. Additional descriptive layers may be added and integrated into the overall semantic structure of the other temporally indexed information creating an organized and detailed description of semantic content and the formal characteristics of the information to be logged.

6. Semantic Logging

6.1 The "Director's Workshop"

The logging interface for audio-video information uses descriptive icons which portray various aspects of the content of the multimedia information (all of the above-mentioned descriptive layers, except "action description" and "dialogue transcription", are iconically represented). The interface allows users to cascade through hierarchies of icons in order to locate icons representing specific content information. In cascading down through the levels of the hierarchy, the space of possible choices is narrowed such that a specific descriptive icon can be quickly selected (see Figs. 1 - 4).

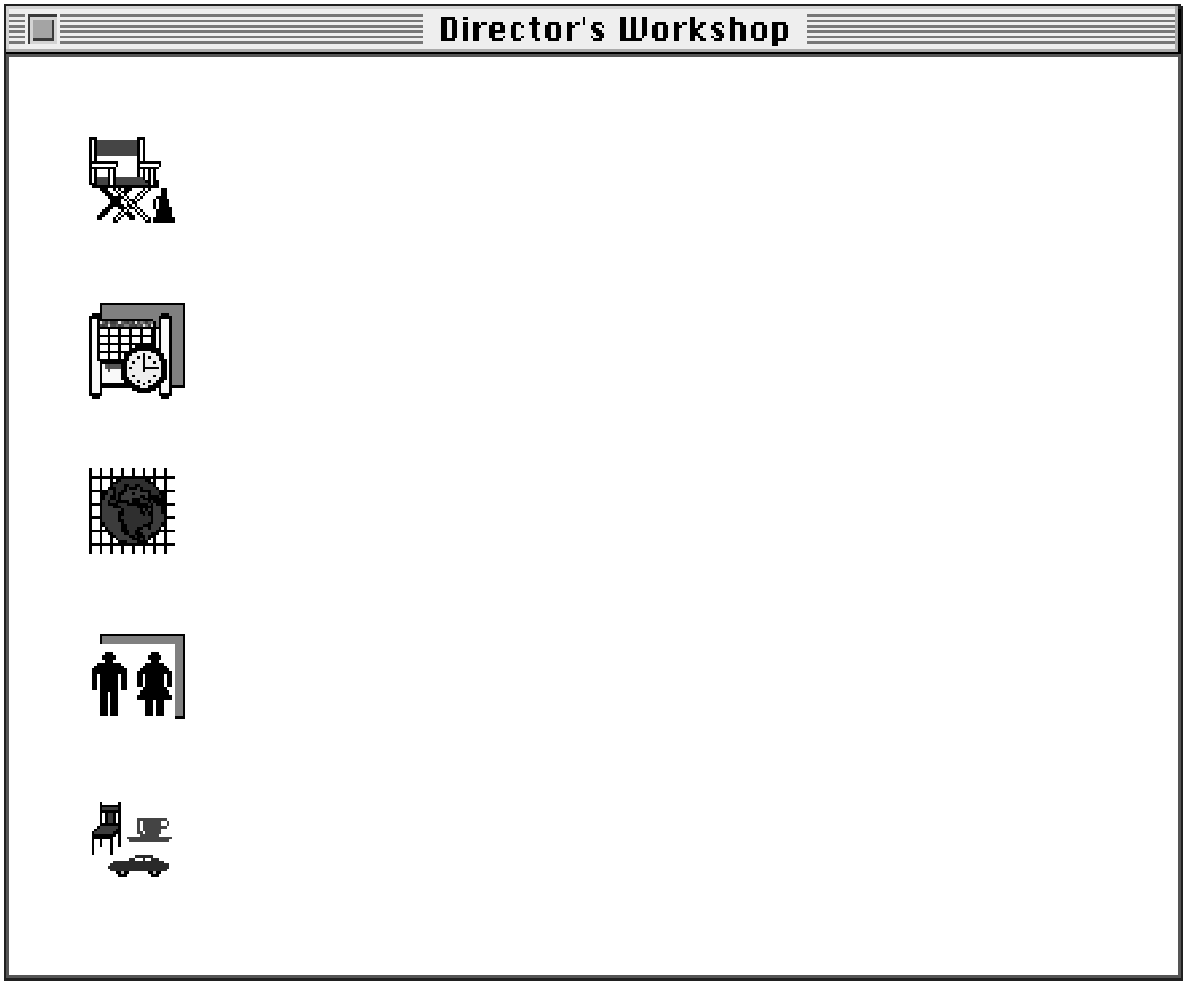


Fig. 1 - The initial screen of the Director's Workshop showing the icons for time, space, characters, and objects.

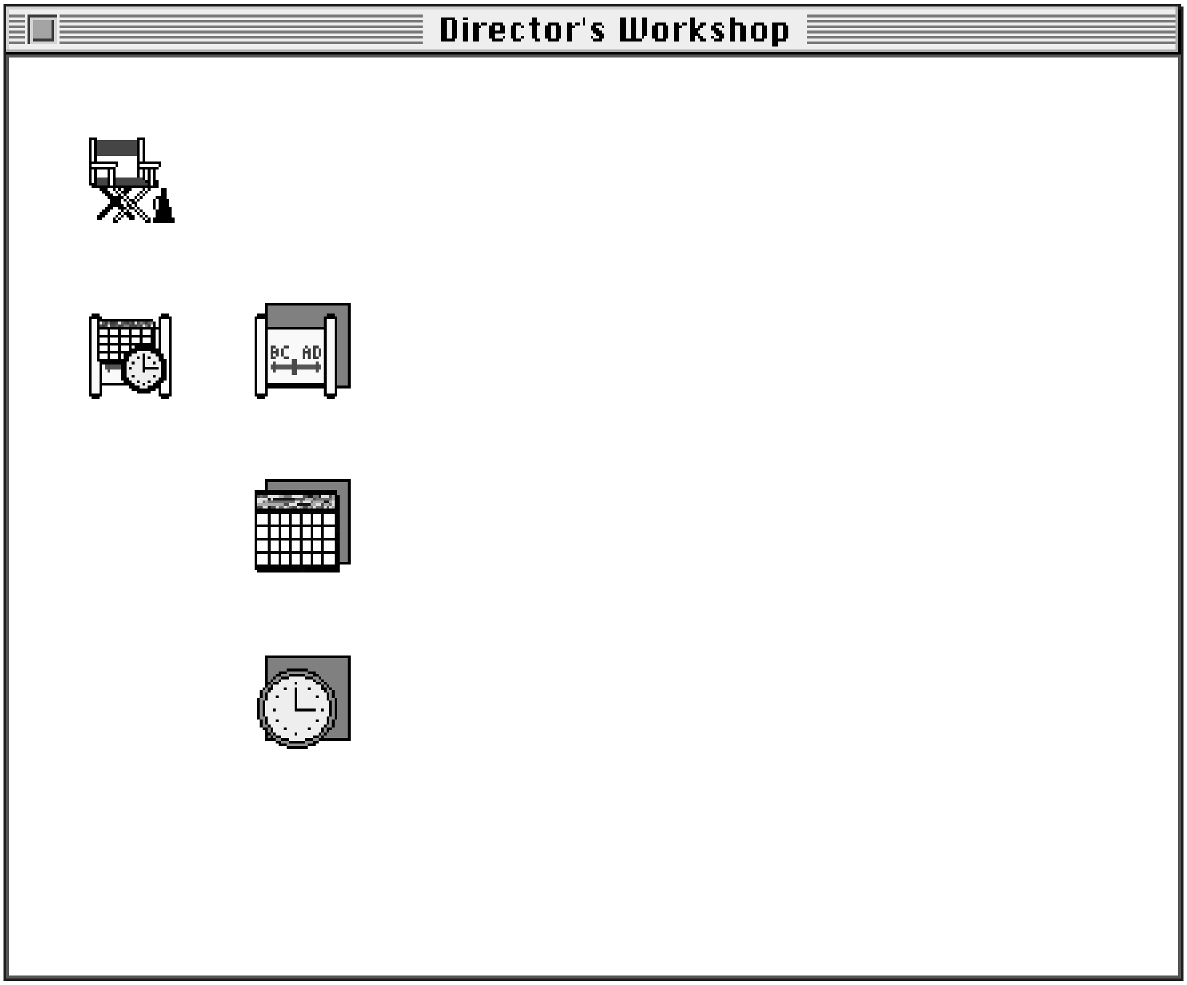


Fig. 2 - The time icon has been selected and displays its subordinate icons for historical period, times of the year, and times of day.

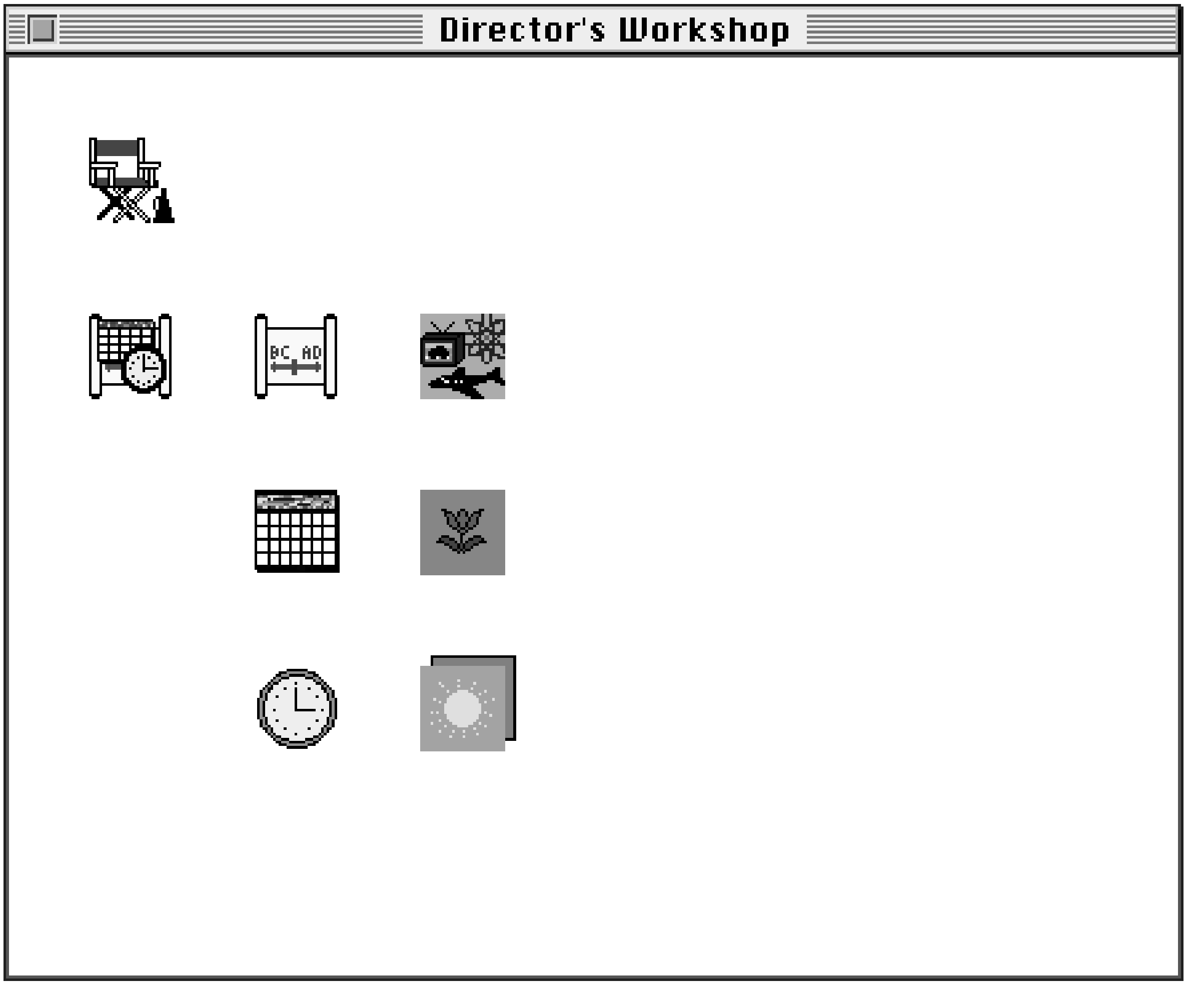


Fig. 3 - After cascading through and selecting subordinate icons from historical period, times of the year, and times of day, the icons have been selected for the twentieth century, Spring, and day.

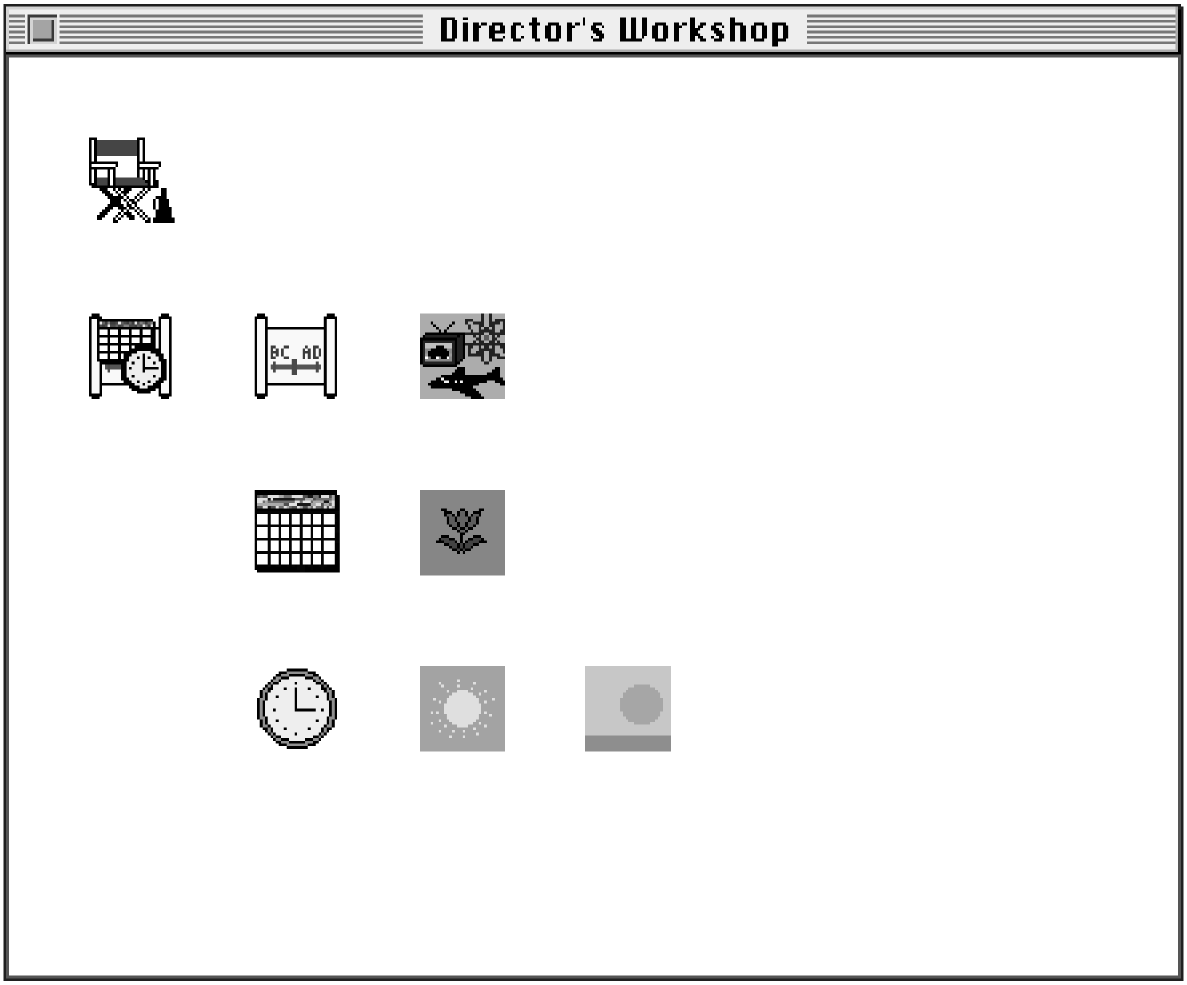
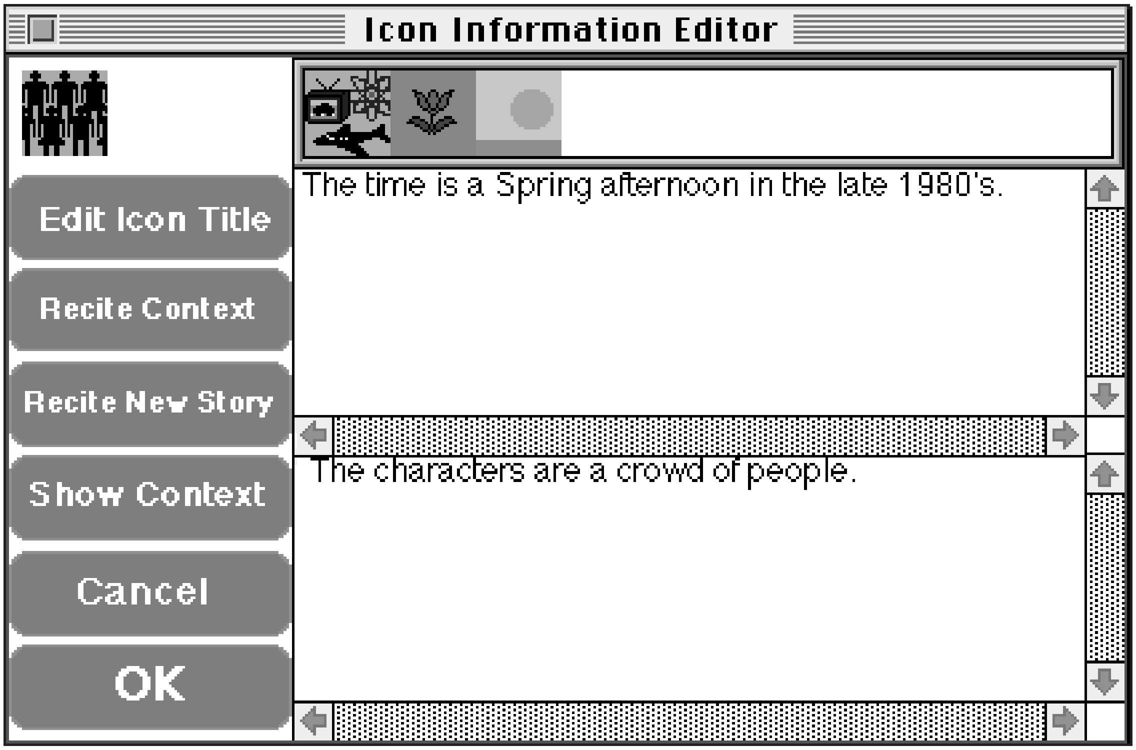


Fig. 4 - The icon for afternoon has been selected from the subordinate icons for day. The user has quickly specified a unique icon for an exact temporal location.

Cascading icons solve the problem of locating specific semantic information in a visual representation of a large, complex knowledge base. For example, without cascading icons it would be cumbersome and time-consuming to locate the icon for "A spring afternoon in the late 1980's" or the icon for "The front entrance of a high school in Boston." Cascading icons are a vocabulary for navigating and specifying iconic representations of the semantic content of multimedia information.

The Director's Workshop utilizes the metaphor of a film director selecting locations, actors, and props for a given scene: the logger uses the Director's Workshop as a type of preprocess to select the icons needed for the representation of the content in the footage to be logged. Of course, semantic loggers don't actually "direct" footage but they do, in effect, analyze footage into the components which a director would have used in producing it. The system's interface metaphor can help make a potentially tedious process perhaps enjoyable as well as aid the user by remaining consistent throughout a user's interaction with multimedia information: through logging, browsing, editing, and sequencing.

By cascading through the hierarchy of icons, the logger creates compound icons representing specific semantic information. These icons can be further described and annotated in an "Icon Information Editor" which allows the user to browse and edit the semantic information represented by the compound icon (see Fig. 5). The "Icon Information Editor" incorporates speech generation (Macintalk) so that relevant icon information can, if the user desires, be read aloud so as to free the user's eyes to do other things. We plan to expand the role of speech generation in the system as well as add speech recognition functionality to the laserdisc controller so as to liberate the user's hands for semantic logging during the process of scanning and searching through video.



The time is a Spring afternoon in the late 1980's.

The characters are a crowd of people.

Fig. 5 - The "Icon Information Editor." The icon being edited is shown in the upper left hand corner, the upper panel contains the icons from which the icon to be edited inherits information. The inherited information is in the upper scrollable window, and the icon's default information is in the lower scrollable window. Text in both windows can be edited.

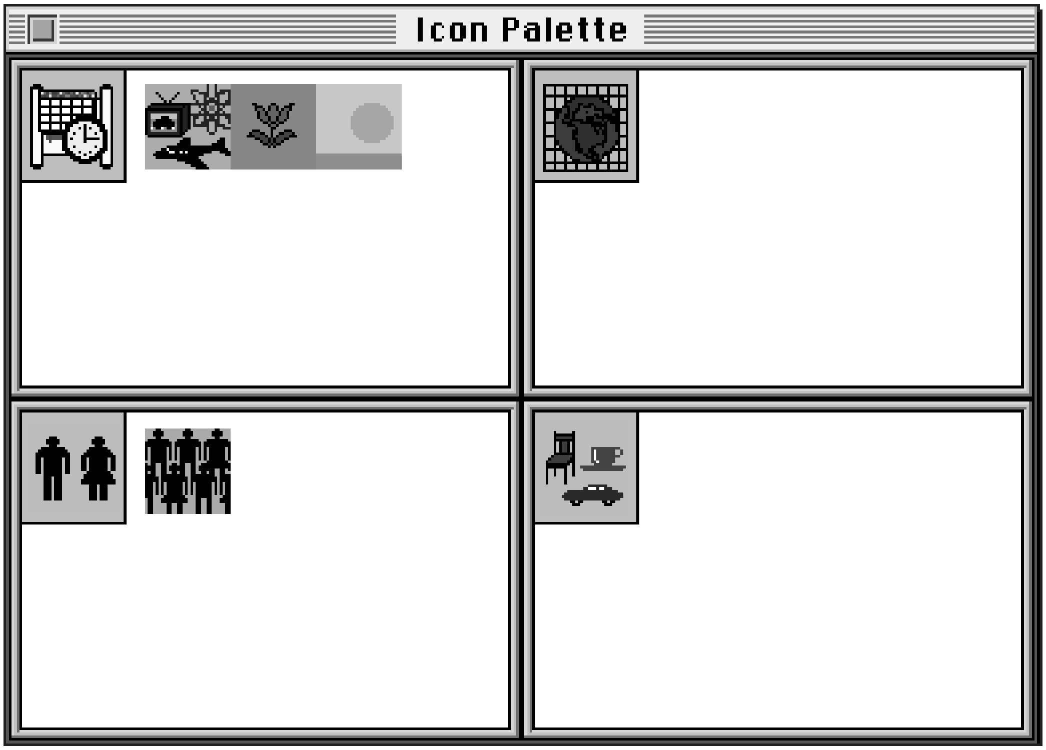


Fig. 6 - The "Icon Palette." Icons automatically place themselves in their corresponding subpalettes.

By using the "Director's Workshop", the logger creates a palette of icons applicable to the footage to be logged (see Fig. 6). Icon palettes can be stored and used as templates for other footage and recurring scenes. For example, footage of a political convention, or a Western for that matter, will have certain stereotypical content features which the icon palettes can supply as default locations, characters, props, etc.. Within the Director's Workshop, the logging process is assisted by the system's ability to make guesses about the content of given scenes based on their stereotypical components as well as the specific components which a logger has already specified in the logging process.

6.2 The "Media Time Line"

In the logging process, the icon palettes created in the "Director's Workshop" are used in annotating segments of audio-video information. The "Media Time Line" represents the continuous temporal extents of the audio, video, and semantic information. A videogram and audiogram represent the various media channels of the information and allow users to quickly scan and browse the representation without necessarily having to re-view the footage. The videogram, a representation of video which allows users to discern coarse content information and scene changes along the temporal axis, was developed by Ron MacNeil at the Media Laboratory's Visible Language Workshop (MacNeil, 1991). The additional descriptive layers temporally indexed below these media channels represent the semantic content of the multimedia information (see Fig. 7).

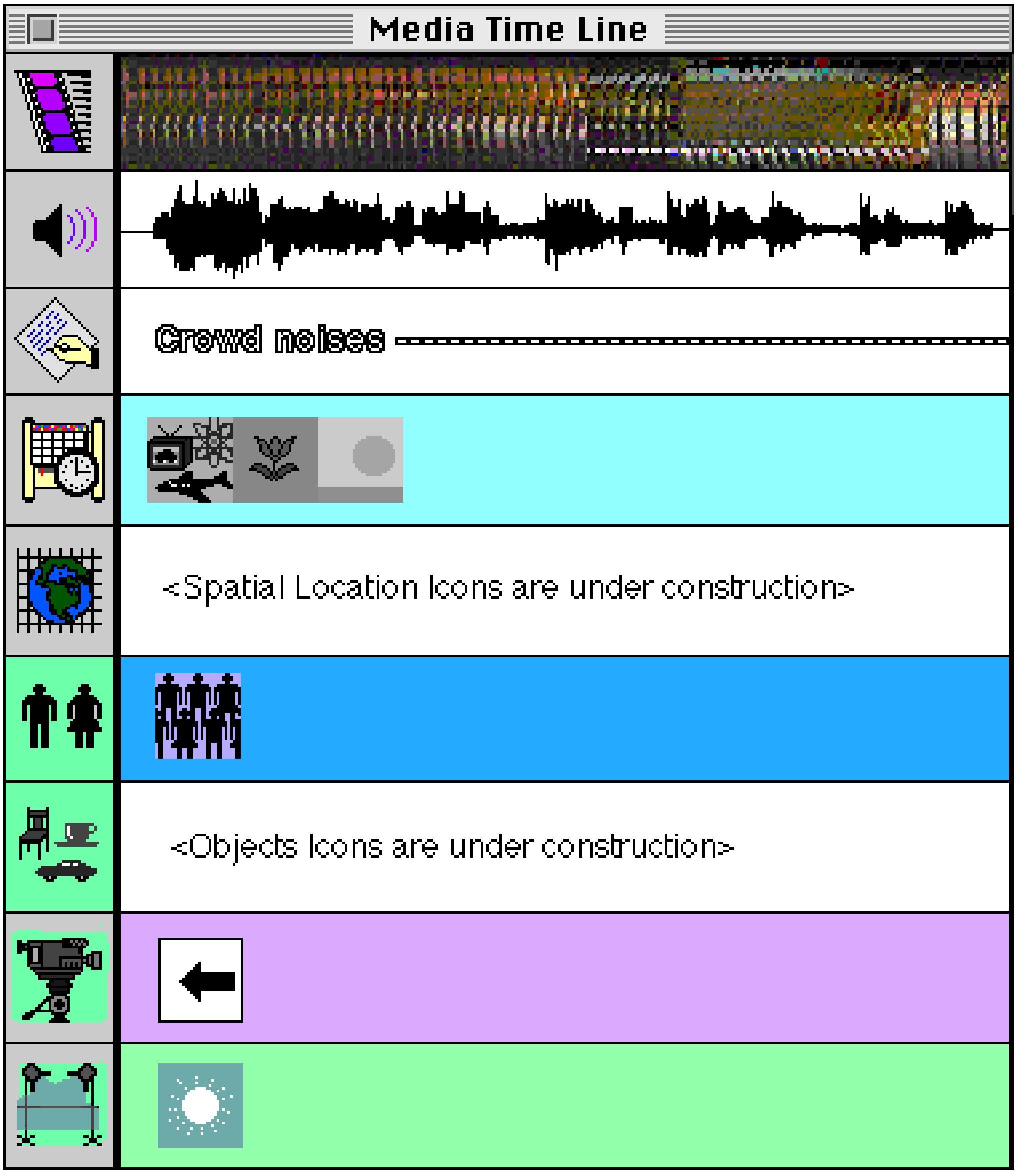

Media Time Line
Crowd noises
<Spatial Location Icons are under construction>
<Objects Icons are under construction>

Fig. 7 - The "Media Time Line."

While logging footage, appropriate icons are dragged onto the Media Time Line to record content information. Once dropped at the appropriate temporal index on the Media Time Line, the icons place themselves in their corresponding descriptive layers. The icons can be dragged off the Media Time Line when the information they represent is no longer valid. Thus the icons function both as a technology for data entry, manipulation, and display. They allow users to semantically describe, edit, and view the content information of temporal media. The layers of descriptive icons form a type of multi-layered story structure along the temporal axis of the logged information. Semantically logged footage can be easily incorporated into a story understanding and generation system because search and retrieval are aided by the system's ability to make inferences, respond to queries, and provide a fine-grained description of the semantic content and formal characteristics of the logged information.

As automated parsing of content improves, the descriptive icons will function less as a medium for data entry, and more as a means for data display. Further research and practical implementation will determine the efficacy of the paradigm of iconic representation of semantic content arrayed in descriptive layers as an idiom for the logging, display, and manipulation of semantically described, temporal multimedia.

6.3 Progress To Date

The system described above is written in Macintosh Common Lisp 2.0 running on a Mac IIx with a 24-bit color monitor. The "Director's Workshop" and "Icon Palettes" are fully implemented, the "Icon Information Editor" needs to be interfaced to the example-based, natural language story understanding and generation system under development at the Media Laboratory by Prof. Ken Haase. The "Media Time Line" interface has been designed and will be fully implemented by the end of this summer. The semantic logger's mechanisms for inference, inheritance, and smart default are currently written in the representation language language ARLOtje (Haase, 1990). This functionality will be expanded when the system is redesigned on the basis of the knowledge representation framework underlying the example-based story system currently in development. Finally, the full integration of video and audio digitization into a working prototype of the system is contingent on near term developments in video and audio compression and storage technology.

7. Automating the Logging Process

Ideas about automating the logging process for video and audio tend to focus on recording information during the production process or on automatically parsing recorded footage. Currently, all information produced at the time of recording video and audio remains inaccessible to users and is often discarded after its production. Storyboards, production stills, continuity logs, editing logs, on location comments and annotations all help to record to the semantic content of video and audio information. A system for computerizing and centrally recording such information has been developed to facilitate the production process (Lasky, 1990) and could be extended to help automate the logging process as well. In addition, the physical situation of the recording devices themselves within their production environment can provide useful information about the scene being recorded. A "data camera" which encoded along with the video signal a time stamp, spatial coordinates, and semantic content annotations recorded on location could pre-log information useful in parsing the content of video and audio (Davenport, Pincever, and Aguierre Smith, 1991).

Many aspects of the content of video images can best be parsed using techniques of image processing and recognition. For the purposes of automatic editing of footage from disparate sources (especially in the case of doing "match cuts"), information about the position of objects within the 2-dimensional coordinate system of the frame, their "screen position," as well as the motion paths of objects from frame to frame require greater precision than semantic descriptions provide. Recently, significant headway has been made in automating the recognition of video content (Ueda, 1991). Researchers at Hitachi Central Research Lab can automatically detect scene boundaries (8 frames/second) and camera motion (0.3 frames/second), and with user assistance semi-automatically recognize objects and their motion paths in the video frame (0.3 frames/second). At the Media Laboratory, research is also being conducted in automatic logging of audio-video information by means of parsing the audio track for pauses, speaker changes, voice intensities, and other audio cues (Pincever, 1990).

This technology, and advances on it, will free human users from much of the drudgery of the logging process. By running an automated preprocess on the video information, the user can randomly access presegmented scenes similarly to tracks on a compact disc. Camera motion detection can classify and label segments of footage according to the types of pans, zooms, dolly shots, etc. which they exhibit. Finally, object detection can exactly locate and log the screen positions of semantically describable objects. With such an automation of the logging process, the potential tedium involved in recording semantic information with the "Director's Workshop" and the "Media Time Line" would be significantly alleviated allowing users to accomplish what automation currently cannot: the semantic description of content.

8. Conclusions and Future Work

The research outlined above is an attempt to solve some fundamental problems in the representation of multimedia information: by means of temporally indexed, semantically structured, cascading icons organized into hierarchical descriptive layers, the system helps users log, manipulate, and display specific content information essential to the representation of complex, temporal media. In the near term we are looking to incorporate recent advances in automated content recognition and audio-video parsing. In addition, the entire problem of the audio domain needs to be substantively addresses as so much of current thinking on multimedia (this paper included) focuses primarily on its visual aspects. Sound is perhaps the greatest determinant of affective context in multimedia and its role needs to be further explored.

The representation outlined in this paper is a step toward automated, intelligent resequencing of multimedia information. Much of the hardest work lies ahead of us in the problem domain of automated composition. An interesting project is underway which applies case-based reasoning techniques to the problem of video story creation by means of archetypical story models in a restricted subject domain (Bruckman, 1991). This and other models of automated resequencing can build on the representation and interface technologies we are developing.

With the advent of digital video and advanced object recognition technology the limitations of a frame-based representation will become apparent. When characters, backdrops, props, and even camera angle and lighting become manipulable objects which can themselves be cut, copied, pasted, and recombined into new scenes and sequences, the representation will have to move beyond the limits of the frame as the basic organizing unit of composition and temporality and become more "object-oriented." The semantic representation of video content and the tools of the "Director's Workshop" and the "Media Time Line" can provide useful metaphors for manipulating objects in digital multimedia because of the common semantic and syntactic organization which underlies the construction of multimedia sequences both in frame-based video and object-oriented digital multimedia.

Acknowledgements

The research discussed above was conducted at the MIT Media Laboratory. The support of the Laboratory and its funders is gratefully acknowledged. I would like to thank my advisor, Prof. Ken Haase, for his continual support and inspiration, and Prof. Glorianna Davenport for her insights and feedback. Amy Bruckman has been instrumental to this research in all its phases and offered valuable editorial assistance in the writing of this paper. I also want to thank my colleagues at the Media Laboratory for their help and conviviality: Edith Ackermann, Barry Arons, Michelle Fineblum, Ron MacNeil, Kevin McGee, David Rosenthal, Alan Ruttenberg, and Mike Travers.

References

Apple Multimedia Lab. Visual Almanac. Cupertino: Apple Computer, Inc., 1989.

Bloch, Gilles R. "From Concepts to Film Sequences." Unpublished Paper. New Haven: Yale University Department of Computer Science, 1987.

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

Bruckman, Amy. "The Electronic Scrapbook: Preliminary Results." In these Proceedings. Anaheim: AAAI-91 Intelligent Multimedia Interfaces Workshop, 1991.

Davenport, Glorianna, Pincever, Natalio, and Aguierre Smith, Thomas G. "Cinematic Primitives for Multimedia: Towards a more profound intersection of cinematic knowledge and computer science representation." Forthcoming in a special issue of IEEE on Multimedia, Summer 1991.

Haase, Ken. ARLOtje Internals Manual. Internal Document. Cambridge: MIT Media Laboratory, 1990.

Isenhour, John Preston. "The Effects of Context and Order in Film Editing." *AV Communication Review*, vol. 23, no. 1, Spring 1975, pp. 69-80.

Iser, Wolfgang. The Act of Reading: A Theory of Aesthetic Response. Baltimore: Johns Hopkins University Press, 1978.

Lasky, Alan. "SLIPSTREAM: A Data Rich Production Environment." Cambridge: MIT Media Laboratory Master's Thesis, 1990.

MacNeil, Ron. "Capturing Multimedia Design Knowledge Using TYRO, the Constraint Based Designer's Apprentice." Forthcoming in the Proceedings of the 1991 SPIE/SPSE Symposium on Electronic Imaging.

Metz, Christian. Film Language: A Semiotics of Cinema. Chicago: University of Chicago Press, 1974.

Pincever, Natalio. "IF YOU COULD SEE WHAT I HEAR: Editing Assistance through cinematic parsing." Master of Science Thesis Proposal. Cambridge: MIT Media Laboratory, 1990.

Ueda, Hirotada, Miayatake, Takafumi, and Yoshizawa, Satoshi. "IMPACT: An Interactive Natural-Motion-Picture Dedicated Multimedia Authoring System." CHI '91 Conference Proceedings (New Orleans, Louisiana, April 28 - May 2, 1991) ACM, New York, 1991, pp. 343 - 350.