M A R C D A V I S

School of Information Management and Systems
University of California at Berkeley

P U B L I C A T I O N S

marc@sims.berkeley.edu
www.sims.berkeley.edu/~marc

Knowledge Representation for Video

Bibliographic Reference:

Marc Davis. "Knowledge Representation for Video." In: *Proceedings of Twelfth National Conference on Artificial Intelligence (AAAI-94) in Seattle, Washington*, AAAI Press, 120-127, 1994.

Knowledge Representation for Video

Marc Davis
Interval Research Corporation
1801-C Page Mill Road
Palo Alto, CA 94304
davis@interval.com

Abstract

Current computing systems are just beginning to enable the computational manipulation of temporal media like video and audio. Because of the opacity of these media they must be represented in order to be manipulable according to their contents. Knowledge representation techniques have been implicitly designed for representing the physical world and its textual representations. Temporal media pose unique problems and opportunities for knowledge representation which challenge many of its assumptions about the structure and function of what is represented. The semantics and syntax of temporal media require representational designs which employ fundamentally different conceptions of space, time, identity, and action. In particular, the effects of the syntax of video sequences on the semantics of video shots demands a representational design which can clearly articulate the differences between the context-dependent and context-independent semantics of video data. This paper outlines the theoretical foundations for designing representations of video, discusses *Media Streams*, an implemented system for video representation and retrieval, and critiques related efforts in this area.

Introduction

The central problem in the creation of robust and scalable systems for manipulating video information lies in representing video content. Currently, content providers possess large archives of film and video for which they lack sufficient tools for search and retrieval. For the types of applications that will be developed in the near future (interactive television, personalized news, video on demand, etc.) these archives will remain a largely untapped resource, unless we are able to access their contents. Without a way of accessing video information in terms of its content, a hundred hours of video is less useful than one.

Given the current state of the art in machine vision and signal processing, we cannot now, and probably will not be able to for a long time, have machines "watch" and understand the content of digital video archives for us. Unlike text, for which we have developed sophisticated parsing and indexing technologies, and which is accessible to processing in various structured forms (ASCII, RTF, PostScript, SGML, HTML), video is still largely opaque. Some headway has been made in this area. Algorithms for the automatic annotation of scene breaks are becoming more robust and enhanced to handle special cases such as fades [Otsuji, 1991 #1454; Zhang, 1993 #1537]. Work on camera motion detection is close to enabling reliable automatic classification of pans and zooms [Teodosio, 1992 #1458; Tonomura, 1993 #1543; Ueda, 1993 #1542]. Researchers are also making progress in the automatic segmentation and tagging of audio data by means of parsing the audio track for pauses and voice intensities [Arons, 1993 #2043], as well as specialized audio parsers for music, laughter, and other highly distinct acoustic phenomena [Hawley, 1993 #1616]. Advances in signal separation and speech recognition will also go a long way to automating the parsing of the content of the audio track. Yet this information alone does not enable the creation of a sufficient representation of video content to support content-based retrieval and manipulation. Signal-based parsing and segmentation technologies must be combined with representations of the higher level structure and function of video data in order to enable machines to make inferences about video content.

Why is video representation an important research area for AI? Besides the pragmatic value of this work for the information and entertainment industries, its relevance extends to the enabling of a broad-based shift in the media of human communication and knowledge. We are currently in a crucial phase of a second "Gutenberg shift" [McLuhan, 1962 #79] in which video is becoming a ubiquitous data type not only for viewing (i.e., reading) but for daily communication and composition (i.e., writing). This shift will only be possible when we can construct representations of video which enable us to parse, index, browse, search, retrieve, manipulate, and (re)sequence video according to representations of its content.

Video representation also requires the rethinking of traditional approaches to knowledge representation and story generation in AI. The generation problem has been framed as the problem of constructing a media independent engine for creating sequences of concepts or events which then guide synthesis processes in different media (usually text [Meehan, 1976 #8; Schank, 1981 #1386], occasionally graphics [Kahn, 1979 #86; Feiner,

1990 #1722]). With recorded video, the generation problem is recast as a representation and retrieval problem. The task, as in editing together found footage, is a matter of creating media specific representations of video which facilitate the retrieval and resequencing of exiting content. This difference in approach has fundamental ramifications for representational design. It is not merely a matter of adapting media independent representations to the specific properties of video, but of designing representations whose basic ontology and inference mechanisms capture the specific semantic and syntactic properties of video.

Therefore, the task which confronts artificial intellignce researchers in this area is to gather insights from disciplines that have studied the structure and function of video data and to use these insights in the design of new representations for video which are adequate to the task of representing the medium. Film analysis and theory have developed a useful repertoire of analytical strategies for describing the semantics and syntax of video data. These insights inform the following theoretical disussion and representational design.

Representing Video

Current paradigms of video representation are drawn from practices which arose primarily out of "single-use" video applications. In single-use applications, video is shot, annotated, and edited for a given movie, video, or television program. Representations are created for one given use of the video data. There do exist certain cases today, like network news archives, film archives, and stock footage houses, in which video is used multiple times, but the level of granularity of the representation and the semantics of the representations do not support a wide reusability of video content. The challenge is to create representations which support "multi-use" applications of video. These are applications in which video may be dynamically resegmented, retrieved, and resequenced on the fly by a wide range of users *other than those who originally created the data*.

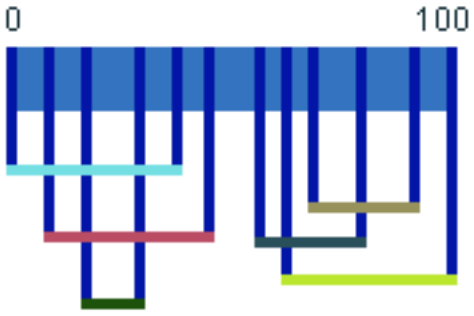
Most attempts to represent video content utilize representations developed for other media. Most commercially used representations apply techniques used for representing text (predominantly keywords or full text annotation); AI-influenced representations apply techniques developed for representing the physical world [Lenat, 1990 #2; Guha, 1994 #2042; Guha, 1994 #2066] or for representing abstract, supposedly media-independent concepts [Schank, 1974 #1884; Schank, 1993 #2013]. All of these attempts neglect to consider that video as a data type may have unique properties which may themselves need to explicitly represented and which may render techniques developed for other media inadequate.

Stream-Based Representation of Temporal Media

In designing a representation of video content we must think about the structure of what is being represented. A video camera produces a temporal stream of image and sound data represented as a stream of frames played back at a certain rate—normally 30 frames per second. This stream of frames has higher level structures of organization commonly referred to as follows: a stream of frames recorded between the time in which the recording device is turned on and turned off is a *shot*; a temporal concatenation of shots is referred to as a *sequence*; and a sequence of shots all sharing the same spatial location is often referred to as a *scene* [Bordwell, 1990 #1975].

In most representations of video content, a stream of video frames is segmented into units called *clips* whose boundaries often, but do not necessarily, coincide with shot or scene boundaries. Current tools for annotating video content used in film production, television production, and multimedia, add descriptors (often keywords) to clips. There is a significant problem with this approach. By taking an incoming video stream, segmenting it into various clips, and then representing the content of those clips, a clip-based representation imposes a *fixed segmentation* on the content of the video stream. To illustrate this point, imagine a camera recording a sequence of 100 frames. Traditionally, one or more parts of the stream of frames is segmented into clips which are then respectively annotated by attaching descriptors. The clip is a fixed segmentation of the video stream that is separated from its context of origin and enforces only one segmentation of the original data.

In a stream-based representation, the stream of frames is left intact and is represented by multi-layered annotations with precise time indexes (beginning and ending points in the video stream). The result is that this representation makes annotation pay off—the richer the annotation, the more numerous the possible segmentations of the video stream.



The Stream of 100 Frames of Video with 6 Annotations Resulting in *66* Possible Segmentations of the Stream

Clips change from being fixed segmentations of the video stream, to being the results of retrieval queries based on annotations of the video stream. In short, in addressing

the challenges of representing video *what we need are representations which make clips, not representations of clips.*

Video Syntax and Semantics

In attempting to create a representation of video content, an understanding of the semantics and syntax of video information is a primary concern. For video, it is essential to clearly distinguish between context-dependent and context-independent semantics. Syntax, the sequencing of individual video shots, creates new semantics which may not be present in any of the individual shots and which may supercede or contravene their existing semantics. This is evidenced by a basic property of the medium which enables not only the repurposing of video data (the resequencing of video shots taken from their original contexts and used to different ends in new contexts), but its basic syntactic functionality: the creation of meaningful sequences through concatenating visual and auditory representations of discontinuous times and discontiguous spaces. Eisenstein described this property as *montage* [Eisenstein, 1947 #38].

The early experimental evidence for effects of the syntax of shot combination on the semantics of individual shots was established by the Soviet cinematographer Lev Kuleshov early in this century [Kuleshov, 1974 #31; Isenhour, 1975 #40]. The classic example of the "Kuleshov Effect" was evidenced by the following experiment. The following sequence was shown to an audience: a long take in close-up of the Russian actor Mozhukin's expressionlessly neutral face — cut — a bowl of steaming soup — cut — the same face of the actor — cut — a woman in a coffin — cut — the same face of the actor — cut — a child playing with a toy bear— cut — the same face of the actor. When audience members were asked what they saw, they said, "Oh, he was hungry, then he was sad, then he was happy." The same exact image of the actor's face was used in each of the three short sequences. What the Kuleshov Effect reveals is that the semantics of video information is highly determined by what comes before and what comes after any given shot.

Because of the impact of the syntax of video sequences on the semantics of video shots, any indexing or representational scheme for video content needs to explain how the semantics of video changes by resequencing and recombination. The challenge is then twofold: to describe what features or annotations survive recombination and to describe how the features which do not survive emerge from those which do.

The challenge of representing the syntax dependent and syntax independent semantic features of video content has a deep similarity to a core problem in knowledge representation: the frame problem [McCarthy, 1969 #2046]. The important difference between approaches to solving the frame problem in AI and the demands of creating a knowledge representation for video lie in the fact that video is itself a representation of the world with its own ontological properties and its own constraints on the construction and maintenance of continuity through the montage of shots. In a word, video has not only its own semantics and syntax, but its own "common sense" which previous approaches to common sense knowledge, temporal, and action representation have yet to address.

Ontological Issues in Video Representation

Space

Through sequencing of shots video enables the construction of many types of spaces: representations of spaces which have real world correlates (real spaces); spaces which do not but could exist in the physical world (artificial spaces); and even spaces which cannot exist in the physical world as we commonly experience it (impossible spaces). In thinking about the first two classes of spaces which can be constructed cinematically (real and artificial spaces) an important distinction can be made between three types of spatial locations: the actual spatial location of the recording of the video; the spatial location which the viewer of the video infers when the video is viewed independent of any other shots; and the spatial location which the viewer of the video infers when it is viewed in a given sequence.

For example, imagine a shot filmed in a dark alley in Paris on October 22, 1983 from 4:15 am to 4:17 am. The actual location of recording may be in a given street in a certain part of the city and could be expressed in terms of an exact longitude, latitude, and altitude. The shot we are imagining has no distinguishing features which mark it as a particular Parisian street or as a Parisian street at all. Independent of any sequence it appears as a "generic dark alley in a city." With the use of a preceding establishing shot, for example an aerial view of New York City at night, the shot now has the inferable spatial location of "a dark alley in New York City." Therefore, representations of the spatial location of a video must represent the difference between a video's actual recorded spatial location and its visually inferable ones.

The geometry of video spaces and the objects within them also have unique properties. The location of objects within the video frame can be represented by a hybrid 2 dimensional and 3 dimensional representation. Since video spaces can be constructed and concatenated into irreal geometries they have only a relational 3 dimensionality in which the geometry is best expressed in terms of *relative* as opposed to *absolute* positions. Therefore, 3 dimensional spatial relations are on the order of "in front of," or "on top of," etc. opposed to a given XYZ coordinate. Since the 3 dimensional world of the video is itself represented in a 2 dimensional projection, all objects in the 3 dimensional space of the recorded/constructed world have a location in the 2 dimensional plane of the screen. The 2 dimensional screen position of an object is a crucial aspect of its spatial

representation and composition which is used by filmmakers to create both aesthetic order (in terms of balanced compositions as in photography) and cognitive order (in terms of the "rules" of Western filmmaking for the construction of space through action, chief among them being the “180 degree rule” which results is the well known shot reverse shot of two person dialogue crosscutting).

Identity

Identity of persons and objects is complex in video. A considerable portion of the cinematic craft is devoted to the construction and maintenance of coherent identities for characters and locales. This is achieved thorough the discipline of “continuity.” Continuity is the process whereby salient details of a character’s and a locale’s appearance remain in continuity from shot to shot (i.e., remain constant when appropriate, change when appropriate). For example, if an actor is wearing a black hat in one shot and not in the next, if there is no inferable explanation for the absence of the hat “continuity” is said to have been broken. The effort to maintain continuity is deeply related to the frame problem in AI. But because video is not the physical world, but a systematic representation of it, continuity can be established and maintained by inferences not found in common sense reasoning.

Interesting examples center around techniques for maintaining the continuity of the identity of a character in a narrative film. A character can literally be “assembled” out of the parts of other characters at various levels of granularity. Kuleshov is well known for constructing a woman character by editing together shots of different body parts of several different women. The identity of a character between shots may rely on any combination of: role (which is comprised of costume, action, and location) and actor. In a demo reel from the stock footage house Archive Films, scenes of several different actors are cut together to make up the central character of a business man traveling around on a busy workday [Archive Films, 1992 #2044] . Continuity of identity can cut across roles and be established by the continuity of the actor. Shots of the same actor taken from various performances of different characters can be edited together to form one character. Imagine, for example, a story about a killer cyborg who goes to Mars which could be created by editing together several of Arnold Schwarzenegger’s films (The Terminator and Total Recall).

Action

The central problem for representing temporal media is the representation of dynamic events. For video in particular, the challenge is to come up with techniques for representing and visualizing the complex structure of the actions of characters, objects, and cameras. A representation of cinematic action for video retrieval and repurposing needs to focus on the granularity, reusability, and semantics of its units. In representing the action of bodies in space, the representation needs to support the hierarchical decomposition of its units both spatially and temporally.

Spatial decomposition is supported by a representation that hierarchically orders the bodies and their parts which participate in an action. For example, in a complex action like driving an automobile, the arms, head, eyes, and legs all function independently. Human body motions are also categorizable in two ways: abstract physical motions and conventionalized physical motions. Abstract physical motions can be represented according to articulations and rotations of joints. There are however many commonly occurring, complex patterns of human motion which seem to have cross-cultural importance (e.g., walking, sitting, eating, talking, etc.). Conventionalized body motions compactly represent motions which may involve multiple abstract body motions.

Temporal decomposition is enabled by a hierarchical organization of units such that longer sequences of action can be broken down into their temporal subabstractions all the way down to their atomic units. In the representational design of the CYC system, Lenat points out the need for more than a purely temporal representation of events that would include semantically relevant atomic units organized into various temporal patterns (repeated cycles, scripts, etc.) [Lenat, 1990 #2]. For example, the atomic unit of “walking” would be “taking a step” which repeats cyclically. An atomic unit of “opening a jar” would be “turning the lid” (which itself could theoretically be broken down into smaller units—but much of the challenge of representing action is knowing what levels of granularity are useful).

In video, however, actions and their units do not have a fixed semantics because their meaning can shift as the video is recut and inserted into new sequences. For example, a shot of two people shaking hands, if positioned at the beginning of a sequence depicting a business meeting, could represent “greeting,” if positioned at the end, the same shot could represent “agreeing.” Video brings to our attention the effects of context and order on the meaning of represented action. In addition, the prospect of representing video for a global media archive brings forward an issue which traditional knowledge representation has largely ignored: cultural variance. The shot of two people shaking hands may signify greeting or agreeing in some cultures, but in others it does not. How are we to annotate shots of people bowing, shaking hands, waving hello and good-bye? The list goes on.

An answer to these issues is to represent the context-independent semantics of actions using physically-based description and to build up the representation of context-dependent semantics by creating a network of analogies between similar concrete action sequences which are themselves represented by physically-based descriptions.

Time

The representation of time in video requires the same distinction made for representing space: the difference between between actual recorded time and the two types of visually inferable time.

A further important distinction in narrative video must be made between three different types of temporal duration [Bordwell, 1990 #1975]:

- story duration (the duration of the events of the entire story as opposed to the particular story events selected for presentation in the video);
- plot duration (the duration of the particular events presented in the video);
- screen duration (the duration of the actual video as screened)

The full representation of these three types of duration is an open research problem.

Media Streams

Media Streams Overview

Over the past three years, members of the MIT Media Laboratory's Machine Understanding Group (Marc Davis with the assistance of Brian Williams and Golan Levin under the direction of Prof. Kenneth Haase) have been building a prototype for the representation and retrieval of video data. This system is called *Media Streams* [Davis, 1993 #2040; Davis, 1993 #1530]. Media Streams is written in Macintosh Common Lisp [Apple Computer, 1993 #1630] and FRAMER [Haase, 1994 #2045; Haase, 1993 #1557], a persistent framework for media annotation and description that supports cross-platform knowledge representation and database functionality. Media Streams runs on an Apple Macintosh Quadra 950 with two high resolution, accelerated 24-bit color displays and uses Apple's QuickTime digital video format [Apple Computer, 1993 #1538].

Media Streams utilizes a hierarchically structured semantic space of iconic primitives which are combined to form compound descriptors. These compound iconic descriptors are used to create multilayered, time indexed representations of video and audio data.

The Icon Space is the interface for the selection and compounding of the iconic descriptors in Media Streams (Fig. 1). To date there are over 2500 iconic primitives. Through compounding, the base set of primitives can produce millions of unique expressions. In the Icon Space, users can create palettes of iconic descriptors for use in annotation and search. By querying the space of descriptors, users can dynamically group related iconic descriptors on-the-fly. These icon palettes enable users to reuse the descriptive effort of others. When annotating video, users can make use of related icons that other users have already created and used to annotate similar pieces of video.

The Media Time Line is the core browser and viewer of Media Streams (Fig. 2). It enables users to visualize video at multiple timescales simultaneously, to read and write multi-layered iconic annotations, and provides one consistent interface for annotation, browsing, query, and editing of video and audio data.

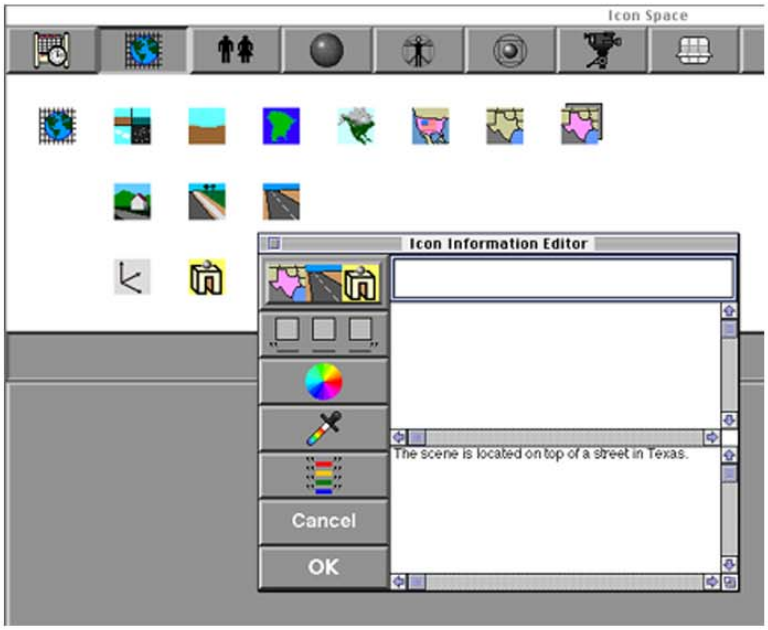


Figure 1: Icon Space (Detail)

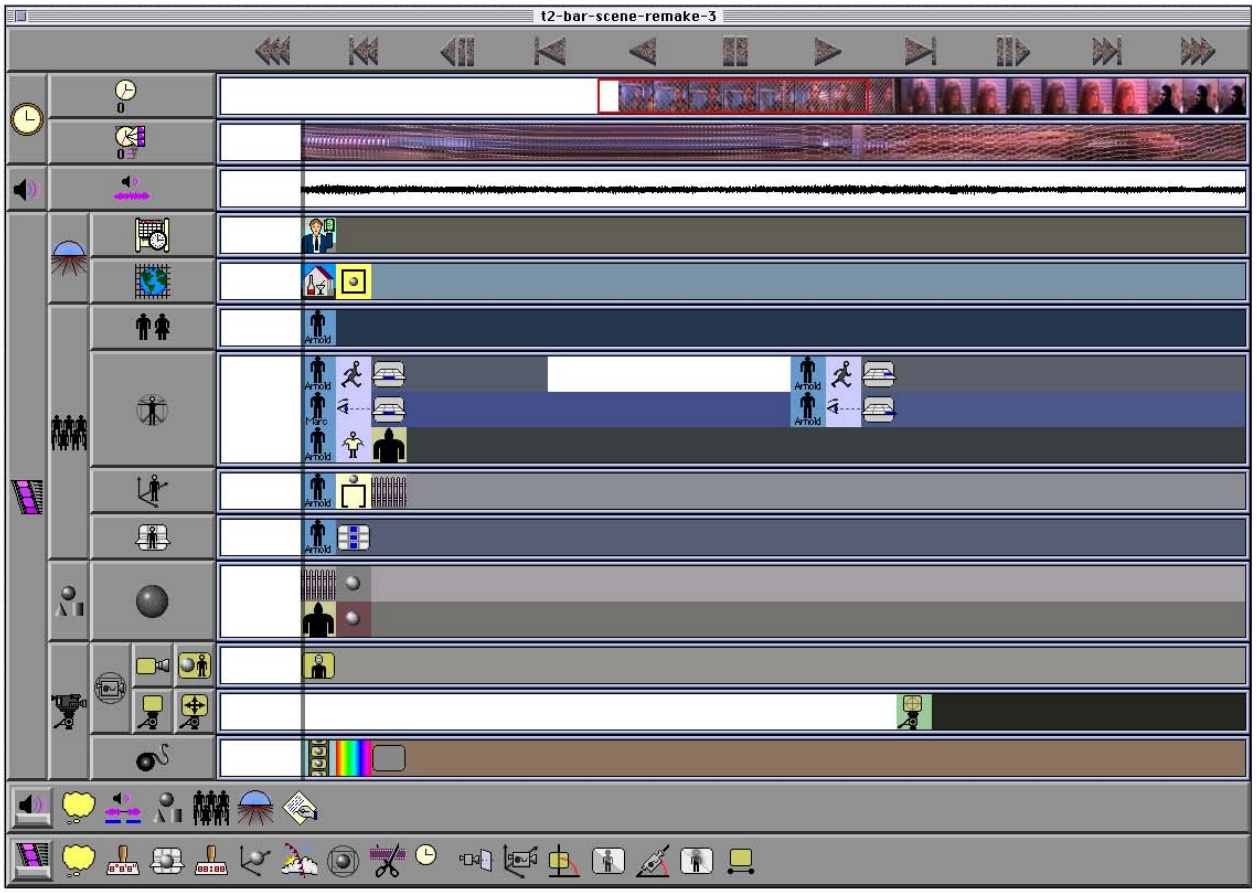


Figure 2: Media Time Line

Media Streams Representational Structures

The underlying representation of video in Media Streams combines two distinct representations: a semantically structured generalization space of atemporal categorical descriptors; and an episodically structured relational space of temporal analogical descriptions.

The semantic/episodic distinction was originated by researchers in human memory [Baddeley, 1984 #1612; Tulving, 1993 #1613] and made computational by Schank's work in dynamic memory [Schank, 1982 #84]. Semantic memory can be thought of as the categorical or definitional part of human memory: remembering what a thing is and what class or category it belongs to. Episodic memory can be thought of as the representation of a sequence of events, an episode. Semantic and episodic memory structures enable us to create a mixed representational system which can answer the fundamental problem of video retrieval systems: how do we determine the similarity of descriptors, of descriptions, of shots, and of sequences? Similarity needs to be context-sensitive and compare not just descriptors, but relations between them. The determination of similarity holds the key to retrieval, and due to the properties of video as a medium (especially its semantic and syntactic features discussed above) the semantic and episodic memory systems must work together using different similarity metrics in order to retrieve video based on its unique features.

Media Streams Retrieval Algorithms

Media Streams employs two different types of retrieval algorithms: atemporal semantically based retrieval of icons and video segments; and temporal analogically based retrieval of video segments and sequences. Both retrieval strategies can use each other and be interleaved.

These algorithms can be further distinguished by the objects they operate on and the criteria of similarity they employ. All retrieval algorithms operate on descriptors and relations between them. At the simplest level, retrieval can be based on the *identity* of components. A more semantically based retrieval utilizes the hierarchical tree structure of the Icon Space to match components based on *generalization* or *specialization*. The most sophisticated retrieval is that which takes into account the semantic and syntactic structure of the descriptions and the relations between them and thereby matches based on *analogical* similarity.

These retrieval algorithms are based on work done by Professor Kenneth Haase [Haase, 1991 #1556; Haase, 1993 #1566]. His analogical matching system called "Mnemosyne" (after the Greek goddess of memory who was also the mother of the nine muses) is a radically memory-based representational system in which analogical matching forms the core representation. The challenge which this memory-based representation addresses is the inflexibility and brittleness of most semantic or categorical representations. In knowledge representations where a fixed hierarchical semantic structure is not sufficient to allow flexibility of the representation, an episodic memory structure is needed so that the semantics of the descriptors used in the semantic memory are, in effect, represented by their differences and similarities to concrete examples of their use. Media Streams extends this work by utilizing two representational systems together (semantic and episodic) in order to facilitate memory-based representation and retrieval of video. Media Streams also adds the ability to represent and match on temporal relations. This extension is based on earlier work in temporal representation [Allen, 1985 #1450].

Related Work

The CYC Project: Representing the World

The goal of the CYC project is to overcome the brittleness and domain-specificity of all previous attempts at representing our common-sense knowledge about the world [Lenat, 1990 #2]. Since 1984 the CYC project has done extensive work in creating representations of objects, actions, and events. Recently the CYC project has begin to apply its large semantic knowledge base to the representation and retrieval of still images and video. Surprisingly, these attempts fall prey to exactly the same criticism which Lenat himself levied against efforts to represent the physical world by natural language. Lenat argued that natural language was an inadequate representational system for representing knowledge about the world because it is not a designed representation [Lenat, 1990 #2]. In other words, natural language is not designed in such a way so as to capture the salient features of the world which are amenable to computational representation. Nevertheless, the CYC project makes a methodological error in its efforts to represent stills and video: it applies its representation language (which is a representation of the world) to video without redesigning it for the representation of video. What Media Streams does in contrast is create a representation language for video, in other words, a representation of a representation of the world. According to Guha, CYC represents video as "information bearing objects with propositional content." Guha admits that this approach may break down due to the particular context-dependent and context-independent semantics of video data [Guha, 1994 #2042]. With video, editing and resequencing may change the given "propositional content" of any "information bearing object."

Schank: Conceptual Dependency and Case Based Reasoning

Conceptual dependency reduced all of human action into a small set of composable primitives [Schank, 1974 #1884]. This work has a certain appeal for its rigor and simplicity, yet it has an apparent deficit for application to video representation: the semantics of human action within video are not fixed and change on recombination. The challenge is not to reduce all video actions to unambiguous media-independent primitives, but to articulate a semantics of action which is conditioned by the properties of the medium.

Traditional case-based reasoning relies on the indexing of cases under predetermined abstractions. This approach presents two problems for video representation: the indexing must, as stated above, articulate the difference between context dependent and context independent aspects of video content; and then use this distinction in its indexing to support the reindexing of cases when video elements are resequenced.

Schank and his students have recently applied their efforts to video representation. They are conducting a large scale project to develop a video database for interactive corporate training applications. In this work video is represented as if it were just text, or a fortiori, ideas. The video data is treated as if it were fully transparent and one need only represent the ideas behind it in order to fully represent its contents. Schank does concede that this approach is designed for the needs of his current project and that it may prove inadequate for representing video which will be resegmented and/or repurposed [Schank, 1993 #2013].

Bloch: AI and Video Representation

The mots promising prior work done in knowledge represntation for video is the research of Gilles Bloch [Bloch, 1987 #41]. In his short unpublished paper he outlines the issues involved in applying Schank's conceptual dependency representation to video segments. He also discusses using Noël Burch's categories for transitions, and mentions the importance of gaze vectors in video [Burch, 1969 #4]. His prototype system supposedly was able to construct simple video sequences using Schankian scripts. His work did not address the issue of how these representations are written (annotation) or read (browsing) and the extent to which they supported repurposability and resegmentation of the content is unclear. Unfortunately, Bloch's untimely death cut off this fruitful early path of research in applying artificial intelligence techniques to the problems of video representation.

Conclusion and Future Work

This paper is a first attempt to articulate the challenge of creating robust representations of video within artificial intelligence which will support the description, retrieval, and resequencing of video according to its content. Work in the representation of video content requires a fundamental analysis of the structure and function of video. The implications for designing media specific representations for video results in the representation of unique semantic, syntactic, and ontological properties of the representational system of video. Media Streams is a research effort in video annotation and retrieval which has begun to develop these types of representations. Much research remains to be done especially in the area of the representation of time, transitions, and the higher level structures of sequences, scenes, and stories.

Acknowledgements

The research discussed above was conducted at the MIT Media Laboratory and Interval Research Corporation. The support of the Laboratory and its sponsors is gratefully acknowledged. I want to thank Brian Williams and Golan Levin for their continually creative and Herculean efforts and my advisor, Prof. Kenneth Haase, for his insight, inspiration, and support. Thanks also to Warren Sack, David Levitt, and Wendy Buffett for editorial and moral support.

References

Allen, J.F., *Maintaining Knowledge about Temporal Intervals,* in *Readings In Knowledge Representation,* R.J. Brachman and H.J. Levesque, Editor. Morgan Kaufmann Publishers, Inc.: San Mateo, California. p. 510-521. 1985.

Apple Computer, *Macintosh Common Lisp Reference*. Cupertino, California: Apple Computer. 1993.

Apple Computer, *QuickTime Developer's Guide*. Cupertino, California: Apple Computer. 1993.

Archive Films, *Archive Films Demo Reel*. Archive Films: New York. 1992.

Arndt, T. and S.-K. Chang. "Image Sequence Compression by Iconic Indexing." In: *Proceedings of 1989 IEEE Workshop on Visual Languages*. Rome, Italy: IEEE Computer Society Press. p. 177-182. 1989.

Arons, B., *Interactively Skimming Recorded Speech*. Massachusetts Institute of Technology: 1993.

Baddeley, A.D., *Memory Theory and Memory Therapy,* in *Clinical Management of Memory Problems,* B.A. Wilson and N. Moffat, Editor. Aspen Systems Corporation:

Rockville, Maryland. p. 5-27. 1984.

Bloch, G.R., *From Concepts to Film Sequences*. Yale University Department of Computer Science: 1987.

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

Bordwell, D. and K. Thompson, *Film Art - An Introduction*. third ed. McGraw-Hill Publishing Company. 1990.

Burch, N., *Theory of Film Practice*. Princeton: Princeton University Press. 1969.

Davis, M., "Media Streams: An Iconic Visual Language for Video Annotation." *Telektronikk*, 4.93: p. 59-71. 1993.

Davis, M. "Media Streams: An Iconic Visual Language for Video Annotation." In: *Proceedings of 1993 IEEE Symposium on Visual Languages*. Bergen, Norway: IEEE Computer Society Press. p. 196-202. 1993.

Del Bimbo, A., E. Vicario, and D. Zingoni. "A Spatio-Temporal Logic for Image Sequence Coding and Retrieval." In: *Proceedings of 1992 IEEE Workshop on Visual Languages*. Seattle, Washington: IEEE Computer Society Press. p. 228-230. 1992.

Del Bimbo, A., E. Vicario, and D. Zingoni. "Sequence Retrieval by Contents through Spatio Temporal Indexing." In: *Proceedings of 1993 IEEE Symposium on Visual Languages*. Bergen, Norway: IEEE Computer Society Press. p. 88-92. 1993.

Eisenstein, S.M., *The Film Sense*. San Diego: Harcourt Brace Jovanovich, Publishers. 1947.

Feiner, S.K. and K.R. McKeown. "Generating Coordinated Multimedia Explanations." In: *Proceedings of Sixth IEEE Conference on Artificial Intelligence Applications*. Santa Barbara: 1990.

Guha, R.V., Personal Communication. January 24, 1994.

Haase, K. "Making Clouds from Cement: Building Abstractions out of Concrete Examples." In: *Proceedings of US-Japan Workshop on Integrated Comprehension and Generation in Perceptually Grounded Environments*. Japan: 1991.

Haase, K., *FRAMER: A Persistent Portable Representation Library*. Internal Document. MIT Media Laboratory: 1993.

Haase, K., *Integrating Analogical and Case-Based Reasoning in a Dynamic Memory*. Internal Document. MIT Media Laboratory: 1993.

Haase, K., *FRAMER Programming Manual*. Internal Document. MIT Media Laboratory: Cambridge, Massachusetts. 1994.

Hawley, M., *Structure out of Sound*. Massachusetts Insitute of Technology: 1993.

Isenhour, J.P., "The Effects of Context and Order in Film Editing." *AV Communications Review*, 23(1): p. 69-80. 1975.

Kahn, K., *Creation of Computer Animations from Story Descriptions*. Massachusetts Institute of Technology Artificial Intelligence Laboratory: 1979.

Karp, P. and S. Feiner. "Issues in the Automated Generation of Animated Presentations." In: *Proceedings of Graphics Interface '90*. Halifax: p. 39-48. 1990.

Lenat, D.B., Forthcoming October 18, 1993.

Lenat, D.B. and R.V. Guha, *Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project*. Reading, Massachusetts: Addison-Wesley Publishing Company, Inc. 1990.

Levaco, R., ed. *Kuleshov on Film: Writings by Lev Kuleshov*. Berkeley, ed. R. Levaco. University of California Press: Berkeley. 1974.

McCarthy, J. and P. Hayes, *Some Philosophical Problems from the Standpoint of Artificial Intelligence,* in *Machine Intelligence 4*. Endinburgh University Press: Endinburgh. 1969.

McLuhan, M., *The Gutenberg Galaxy: The Making of Typographic Man*. Toronto: University of Toronto Press. 1962.

Meehan, J., *The Metanovel: Writing Stories by Computer*. Yale University: 1976.

Nagasaka, A. and Y. Tanaka, *Automatic Video Indexing and Full Video Search for Object Appearances,* in *IFIP Transactions, Visual Database Systems II,* K. Wegner, Editor. Elvesier Publishers: 1992.

Otsuji, K., Y. Tonomura, and Y. Ohba, "Video Browsing Using Brightness Data." *SPIE Visual Communications and Image Processing '91: Image Processing*, SPIE 1606: p. 980-989. 1991.

Pincever, N.C., *If You Could See What I Hear: Editing Assistance Through Cinematic Parsing*. Massachusetts Institute of Technology: 1990.

Schank, R.C., *Dynamic Memory: A Theory of Reminding*

and Learning in Computers and People. Cambridge: Cambridge University Press. 1982.

Schank, R.C., Personal Communication. October 18,1993.

Schank, R.C. and C.J. Rieger III, "Inference and the Computer Understanding of Natural Language." *Artificial Intelligence*, 5(4): p. 373-412. 1974.

Schank, R.C. and C. Riesbeck, *Inside Computer Understanding: Five Programs Plus Miniatures*. The Artficial Intelligence Series, ed. R.C. Schank. Hillsdale, New Jersey: Lawrence Erlbaum Associates. 1981.

Teodosio, L., *Salient Stills*. Massachusetts Institute of Technology Media Laboratory: 1992.

Tonomura, Y., *et al*. "VideoMAP and VideoSpaceIcon: Tools for Anatomizing Content." In: *Proceedings of INTERCHI'93 Conference on Human Factors in Computing Systems*. Amsterdam, The Netherlands: ACM. p. 131-136. 1993.

Tulving, E., "What is Episodic Memory?" *Current Directions in Psychological Science*, 2(3): p. 67-70. 1993.

Ueda, H., *et al*. "Automatic Structure Visualization for Video Editing." In: *Proceedings of INTERCHI'93 Conference on Human Factors in Computing Systems*. Amsterdam, The Netherlands: ACM. p. 137-141. 1993.

Zabih, R., J. Woodfill, and M. Withgott. "A Real-Time System for Automatically Annotating Unstructured Image Sequences." In: *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics*. Le Touquet, France: IEEE Press. p. 345-350. 1993.

Zhang, H., A. Kankanhalli, and S.W. Smoliar, "Automatic Partitioning of Full-Motion Video." *Multimedia Systems*, 1: p. 10-28. 1993.