Media Streams: An Iconic Visual Language for Video Representation

Bibliographic Reference:

Marc Davis. "Media Streams: An Iconic Visual Language for Video Representation." In: *Readings in Human-Computer Interaction: Toward the Year 2000*, eds. Ronald M. Baecker, Jonathan Grudin, William A. S. Buxton, and Saul Greenberg. 854-866. 2nd ed., San Francisco: Morgan Kaufmann Publishers, Inc., 1995.

Media Streams: An Iconic Visual Language for Video Representation

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

Abstract

In order to enable the search and retrieval of video from large archives, we need a representation language for video content. Although some aspects of video can be automatically parsed, a sufficient representation requires that video be annotated. We discuss the design of a video representation language with special attention to the issue of creating a global, reusable video archive. Our prototype system, Media Streams, enables users to create multi-layered, iconic annotations of streams of video data. Within Media Streams, the organization and categories of the Icon Space allow users to browse and compound over 3500 iconic primitives by means of a cascading hierarchical structure that supports compounding icons across branches of the hierarchy. A Media Time Line enables users to visualize, browse, annotate, and retrieve video content. The challenges of creating a representation of human action in video are discussed in detail, with focus on the effect of the syntax of video sequences on the semantics of video shots.

1 Introduction: The Need for Video Representation

Without content representation, the development of large-scale systems for manipulating video will not happen. Currently, content providers possess massive 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. With one hour of video, its content can be stored in human memory, but as we move up in orders of magnitude, we need to find ways of creating machine-readable and human-usable representations of video content. It is not simply a matter of cataloging reels or tapes, but of representing the content of video so as to facilitate the retrieval and repurposing of video according to these representations.

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 parse and understand the content of digital video archives for us. Unlike text, for which we have developed sophisticated parsing 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 shot breaks are becoming more robust and enhanced to handle special cases such as fades (Nagasaka and Tanaka 1992; Zhang and others 1993). Work on camera motion detection is close to enabling reliable automatic classification of pans and zooms (Teodosio 1992; Tonomura and others 1993; Ueda and others 1993). Problems which are still quite difficult but which are being actively worked on include: object recognition (Nagasaka and Tanaka 1992), object tracking (Ueda and others 1991), and motion segmentation (Otsuji and others 1991; Zabih and others 1993). Research is also being conducted in automatic segmentation and tagging of audio data by means of parsing the audio track for pauses and voice intensities (Arons 1993), other audio cues including sounds made by the recording devices themselves (Pincever 1990), as well as specialized audio parsers for music, laughter, and other highly distinct acoustic phenomena (Hawley 1993). Advances in signal separation and speech recognition will also contribute 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 semantic and syntactic structure of video data in order to support annotation, browsing, retrieval, and resequencing of video according to its content. In the near term, it is computer-supported human annotation that will enable video to become a rich, structured data type.

1.1 Video Representation Today

In developing a structured representation of video content for use in the annotation, retrieval, and repurposing of video from large archives, it is important to understand the current state of video annotation in order to create specifications for how future annotation systems should be able to perform. To begin with, we can posit a hierarchy of the efficacy of annotations:

At least, Pat should be able to use Pat's annotations.

Slightly better, Chris should be able to use Pat's annotations.

Even better, Chris's computer should be able to use Pat's annotations.

At best, Chris's computer and Chris should be able to use Pat's and Pat's computer's annotations.

Today, annotations used by video editors will typically only satisfy the first desideratum (Pat should be able to use Pat's annotations) and only for a limited length of time. Annotations used by video archivists aspire to meet the second desideratum (Chris should be able to use Pat's annotations), yet these annotations often fail to do so if the context of annotation is too distant (in either time or space) from the context of use. Current computer-supported video annotation and retrieval systems use keyword-based representations of video and ostensibly meet the third desideratum (Chris's computer should be able to use Pat's annotations), but practically do not because of the inability of keyword representations to maintain a consistent and scaleable representation of the salient features of video content.

1.2 Why Keywords Are Not Enough

In the main, video has been archived and retrieved as if it were a non-temporal data type that could be adequately represented by "keywords." A good example of this approach can be seen in Apple Computer's *Visual Almanac* that 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 video retrieval and repurposing. The keyword approach is inadequate for representing video 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, similarity, or inference between descriptors. Looking for shots of “dogs” will not retrieve shots indexed as “German shepherds” and vice versa.
- Keywords do not describe *relations* between descriptors. A search using the keywords “man,” “dog,” and “bite” may retrieve “dog bites man” videos as well as “man bites dog” videos—the relations between the descriptors highly determine their salience and are not represented by keyword descriptors alone.
- Keywords do not *converge*. Since they are laden with linguistic associations and not a structured, designed language, keywords, as a representation mechanism for video content, suffer from the “vocabulary problem” (Furnas and others 1987). Different users use sufficiently different keywords to describe the same materials such that keyword annotation becomes idiosyncratic rather than consensual.
- Keywords do not *scale*. As the number of keywords grows, the possibility of matching a query to the annotation diminishes. As the size of the keyword vocabulary increases, the precision and recall of searches decrease.

Because of the deficiencies of keyword-based annotation and retrieval systems, current video archives cannot rely on computers to overcome the inherent barriers to sharability and durability in human memory. In fact, even with today’s “computerized” systems video archives rely on human memory as the crucial repository of the knowledge not contained in computational representations.

1.3 Towards a Global Media Archive

A video annotation language needs to create representations that are durable and sharable. The knowledge encoded in the annotation language needs to extend in time longer than one person’s memory or even a collective memory, and needs to extend in space across continents and cultures. Today, and increasingly, content providers have global reach. German news teams may shoot footage in Brazil for South Korean television that is then accessed by American documentary filmmakers, perhaps ten years later. We need a global media archiving system that can be added to and accessed by people who do not share a common language, and the knowledge of whose contents is not only housed in the memories of a few people working in the basements of news archives and film libraries.

The visual language we have designed may provide an annotation language with which we can create a truly global media resource. Unlike other visual languages that are used internationally (e.g., for traffic signage, operating instructions on machines, etc.), a visual language for video annotation can take advantage of the affordances of the computer medium. We have developed an iconic visual language for video annotation that is computationally writable and readable, and makes use of a structured, semantic, searchable, generative vocabulary of iconic primitives. It also uses color, shading, anti-aliasing, and animation in order to support the creation of durable and sharable representations of video content.

2 Representing Video

Current paradigms of video representation are drawn from practices which arose primarily out of “single-use” video applications. In single-use applications, footage is shot, annotated, and edited for a given movie, story, or film. Annotations 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, semantics, and non-uniformity with which these organizations annotate their archives limits the repurposability of their representations and their video content. The challenge is to create representations which support “multiuse” 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*. In order to create representations for reusable video, we need to understand the structure and function of what is being represented.

2.1 Streams vs. Clips

Video is a temporal medium that represents continuities and discontinuities of space, time, and action. The first task of a representation of video content is to provide a set of units into which the temporal streams of audio and video data can be parsed. In film theory, this task of parsing the streams of video and audio data into units is called *segmentation* (Bordwell and Thompson 1990). The task of representing the basic structures of video data is the task of creating useful segmentations of that data.

One might think that for the purposes of retrieval and repurposing a segmentation of video into frames, shots, sequences, and scenes would be sufficient. However necessary these traditional segmentations are for video representation they are insufficient for representing video content. First of all, each of these segmentations has certain inherent limitations as a content representation. Frames by themselves are too fine a segmentation and remove the temporal aspects of video content from a representation. Scenes are often too large of a segmentation to be useful for repurposing; by virtue of their completeness they render their parts less easily repurposable. Shots and sequences are a useful level of granularity, but in and of themselves these segmentations do not represent their contents. Finally, and most importantly, there are many aspects of video content which *continue across shot and scene boundaries* (e.g., music, dialogue, character, etc.) or *exist within shot boundaries* (e.g., action, camera motion, etc.).

Today, most systems for representing and manipulating video create a segmentation of video into *clips*. As will be explained below, representing video by segmenting it into clips is a representational strategy that does not support multiple reuse of the representations or of the data represented. The core task of representing video for repurposing is to create *a segmentation of the data out of which multiple segmentations can be generated*. As will be explained below, a *stream-based* representation of video content enables multiple segmentations of video to be generated (Davenport and others 1991).

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, sequence, 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 would be segmented into clips which would then be annotated by attaching descriptors. The clip is a fixed segmentation of the video stream that separates the video from its context of origin and encodes a particular chunking of the original data.

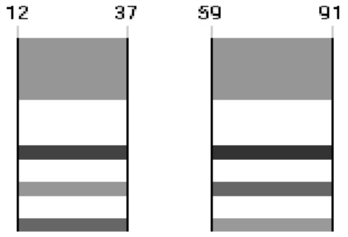


Figure 1. Two “clips” with Three Descriptors Each

In a stream-based representation, the stream of frames is left intact and is annotated by multi-layered annotations with precise time indexes (beginning and ending points in the video stream). Annotations could be made within any of the various categories for video representation discussed below (e.g., characters, character actions, objects, spatial location, camera motions, dialogue, etc.) or contain any data the user may wish.

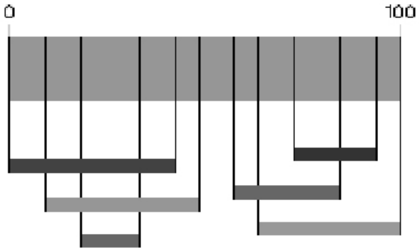


Figure 2. Stream of 100 Frames of Video with 6 Annotations Resulting in *66* Possible Segmentations of the Stream

Stream-based representation makes annotation pay off—the richer the annotation, the more numerous the possible segmentations of the video stream. Stream-based annotations generate new segmentations by virtue of their unions, intersections, overlaps, etc. Clips change from being fixed segmentations of the video stream, to being the results of retrieval queries into the network of stream-based annotations of the video stream. In short, in addressing the challenges of representing video for large archives *what we need are representations which make clips, not representations of clips.*

2.2 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. Video has a radically different semantic and syntactic structure than text, and attempts to represent video and index it in ways similar to text will suffer serious problems. For video, it is essential to clearly distinguish between its sequence-dependent and sequence-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 supersede or contravene their existing semantics. This is evidenced by a basic property of the medium that 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 motion pictures' basic semantic and syntactic functionality: the creation of meaningful sequences through the *montage* of visual and auditory representations of discontinuous times and discontiguous spaces. Eisenstein described this property as *montage* (Eisenstein 1947).

The early experimental evidence for the 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 (Isenhour 1975; Kuleshov 1973; Kuleshov 1974). The classic example of the "Kuleshov Effect" was evidenced by the following experiment (Pudovkin 1949). 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**
>
> **the same face of the actor**
> **— cut — a woman lying dead in a coffin**
>
> **the same face of the actor**
> **— cut — a child playing with a toy bear**

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. It is the Kuleshov Effect that makes the construction of cinematic sequences possible at all and that enables us to reuse existing footage to make new sequences.

Kuleshov's experiments began the work of cataloging the effects and principles which underlie all montage and are especially important for a representation of video that seeks to repurpose content and retrieve sequences by composing segments from various videos. 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 through resegmentation and resequencing. The challenge for video representation is to provide a framework for determining, representing, and relating those aspects of video content whose semantics are invariant and sequence-independent and those aspects whose semantics are variable and sequence-dependent.

What film theory teaches us is that a representation of video content cannot rely on existing representational strategies for other media or for the physical world. Video is itself a representational system with its own ontological properties and its own constraints on the construction and maintenance of representations of spaces, objects, characters, and actions 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.

2.3 Categories for Video Representation

A central question in our research is the development of a minimal set of categories for representing video content. One of the principal features that makes video unique is that it is a temporal medium. Any language for annotating the content of video must have a way of talking about temporal events—the actions of humans and objects in space over time. Therefore, we also need a way of talking about the characters and objects involved in actions as well as their mise-en-scene, that is, the spatial location, temporal location, and weather/lighting conditions in which these actions take place. The objects and characters involved in actions in particular settings also have significant positions in space relative to one another (beneath, above, inside, outside, etc.).

These categories—*actions*, *characters*, *objects*, *relative positions*, *locations*, *times*, and *weather*—would be nearly sufficient for talking about actions in the world, but video is a *recording* of actions in the world by a camera, and any representation of video content must address further specific properties. First, we need ways of talking about *cinematographic properties*, the movement and framing of the camera recording events in the world. We also need to describe the properties of the *recording medium* itself (film or video, color or black & white, graininess, etc.) Furthermore, in video, viewers see events depicted on screens, and therefore, in addition to relative positions in space, screen objects have a *screen position* in the two-dimensional grid of the frame and in the various layered vertical planes of the screen depth. Finally, video recordings of events can be manipulated as objects and rearranged. We create transitions in video in ways not possible in the physical world. Therefore, *cinematic transition*s must also be represented in an annotation language for video content. In working with video archivists from Monitor Television, we found that in their daily practice (in addition to the above mentioned intersubjective categories) video producers would ask for footage according to highly subjective *thoughts* about the video content which relate to the quality of the frame composition, color, and level of activity.

These categories need not be *sufficient* for media annotation (the range of potential things one can say is unbounded), but we believe they are *necessary* categories for media annotation in order to support retrieval and reuse of particular segments of video data from an annotated stream.

These minimal annotation categories attempt to represent information about media content that can function as a substrate:

- on top of which other annotations may be layered
- out of which new annotations may be inferred
- within which the differences between consensual and idiosyncratic annotations may be articulated

In a minimal representation of video content, the primary level of representation is of the semantically invariant, sequence-independent aspects of video. The semantically variable, sequence-dependent aspects of video content are represented in terms of this primary level of representation. Therefore, the representational system is optimized to represent that which one sees and hears in a video shot, rather than what one infers from the syntactic context of a video shot (Bordwell 1985). The process of representation is highly decontextualizing in order that these representations can support retrieval and repurposing of video content.

3 Media Streams: An Overview

Over the past four years, a small group of researchers in the MIT Media Laboratory's Machine Understanding Group (myself with the assistance of Brian Williams and Golan Levin under the direction of Prof. Kenneth Haase) has built *Media Streams,* a prototype for the representation, retrieval, and repurposing of video and audio data (Davis 1993a; Davis 1993b; Davis 1994a; Davis 1994b; Davis and others 1994; Sack and Davis 1994; Davis 1995).

Media Streams is written in two languages: the outstanding rapid prototyping environment of Macintosh Common Lisp (Apple Computer 1993a) with its CLOS (Common Lisp Object System) interface to the Macintosh ToolBox, and FRAMER (Haase 1994; Haase and Sack 1993), a persistent framework for media annotation and description that supports cross-platform knowledge representation and database functionality. Media Streams has its own Lisp interface to Apple's QuickTime digital video system software (Apple Computer 1993b). Media Streams has been developed on an Apple Macintosh Quadra 940 with two high resolution color displays.

Media Streams enables users to preprocess, annotate, browse, retrieve, and repurpose digital video and audio content with an iconic visual language designed for video representation. Its main functions are outlined in the following subsections.

3.1 Media Streams Functionality

3.1.1. Preprocessing

Media Streams makes use of existing and reliable signal-processing techniques for automatically creating meaningful segmentations and visualizations of digital video and audio data. When a QuickTime movie is first loaded into the system Media Streams creates shot-breaks for the video and pause-breaks for the audio. The system also automatically creates multiple representations of the video and audio data's structure at different temporal and spatial resolutions which are used in visualizing and navigating the data (for video: thumbnails and a videogram; for audio: waveforms and pause-break bars).

3.1.2. Annotation

In Media Streams, annotators use an iconic visual language to create stream-based annotations of video content. Media Streams utilizes a hierarchically structured semantic space of iconic primitives which are combined to form compound descriptors which are then used to create multi-layered, temporally indexed annotations of video content. These iconic primitives are grouped into the descriptive categories designed for video representation and are structured to deal with the special semantic and syntactic properties of video data discussed above. These categories include: space, time, weather, characters, objects, character actions, object actions, relative position, screen position, recording medium, cinematography, shot transitions, and subjective thoughts about the material.

In Media Streams, the annotation language is designed to support the annotation of the consensual aspects of video content—what one sees and hears, rather than what one infers from context—in order to facilitate the convergence of iconic annotations and the repurposability of the content described by these annotations. Media Streams does not aim to support all types of annotations, but only those physically-based descriptions whose semantics supports repurposing. Other types of annotations may be layered on top of and use those created in Media Streams, but the goal here is for finding the most minimalist way of saying the most salient things about the content so as to support content-based retrieval for repurposing.

Media Streams' annotations do not describe video clips, but are themselves temporal extents describing content within a video stream. As *stream-based* annotations they support multiple layers of overlapping descriptions which, unlike clip-based annotations, enable video to be dynamically resegmented at query time.

The system also supports the reuse of other people's descriptive effort through the ability to retrieve and group related iconic descriptors into palettes.

3.1.3. Browsing

Browsing in Media Streams makes use of the representations of video content which are automatically generated as well as annotations created by human users. For example, users can use a jump button (identical to the "track advance" button on consumer CD players) in order to jump by content to the next logical change in the video stream be it the next shot break or the next new character in a shot.

3.1.4. Retrieval

Media Streams supports the retrieval of annotated video segments and sequences in two ways: *by description* or *by example*. Query by description is the use of the annotation language as a query language in order to describe footage that one wants to find. Query by example is using already annotated footage itself as a query. Unlike most conventional video retrieval systems, Media Streams supports query of annotated video according to its *temporal* and *semantic* structure.

3.1.5. Repurposing

Media Streams is designed to support the repurposing of video content in all of its functions and components. The functionality that most clearly shows this is the way in which Media Streams redefines retrieval in terms of composition. A query for a video sequence will not only search the annotated video streams for a matching sequence but will *compose* a sequence out of parts from various videos in order to satisfy the query. We refer to this retrieval strategy as *retrieval-by-composition*. In answering user queries, Media Streams can repurpose the content in its own archive in order to *make* video sequences as a way of satisfying requests to *find* them.

3.2 Media Streams System Components

Media Streams attempts to address two fundamental interface issues in video annotation and retrieval: creating and searching the space of descriptors to be used in annotation and retrieval; and visualizing, annotating, browsing, and retrieving video shots and sequences. Consequently, the system has two main interface components: the Icon Space (Figure 3) and the Media Time Line (Figure 5).

3.2.1. Icon Space

The **Icon Space** is the interface for the selection and compounding of the iconic descriptors in Media Streams (Figure 3). To date Media Streams has over 3500 iconic primitives. In the **Icon Workshop** portion of the Icon Space (the upper half) these iconic primitives can be compounded to form compound icons. Through compounding, the base set of primitives can produce millions of unique expressions.

In the **Icon Palette** portion of the Icon Space (the lower half of Figure 3), 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. In Figure 3, the query for the union of female character icons and full body character actions has retrieved compound icons of various females doing various actions. Importantly, the Icon Palette enables 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 a similar piece of video.

3.2.1.1. Icon Workshop Organization

What enables the user to navigate and make use of our large number of primitives is the way the Icon Workshop organizes icons into cascading hierarchies. We refer to the iconic primitives in the Icon Workshop as *cascading icons*. The Icon Workshop has two significant forms of organization for managing navigational and descriptive complexity:

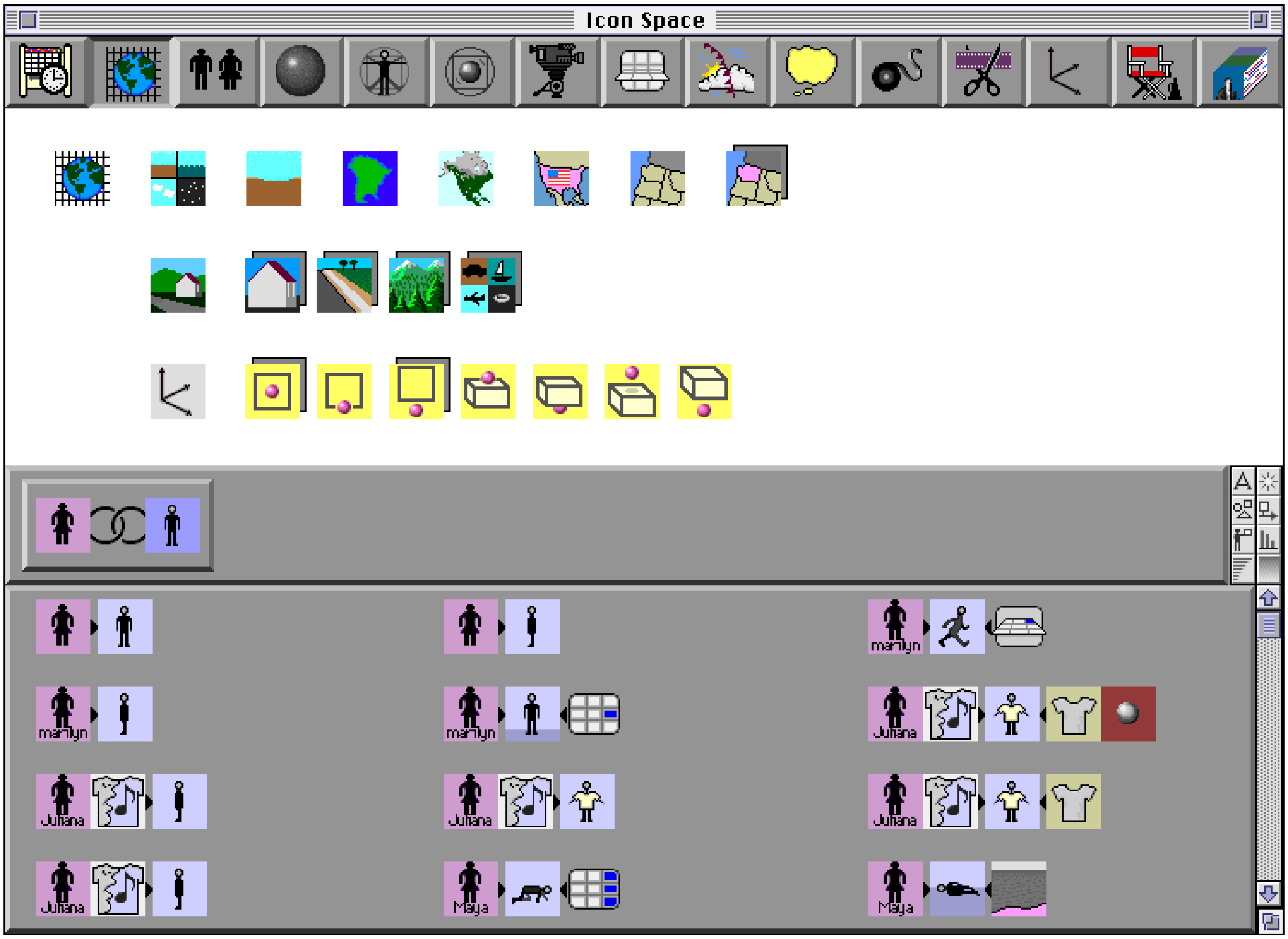


Figure 3: The Icon Space

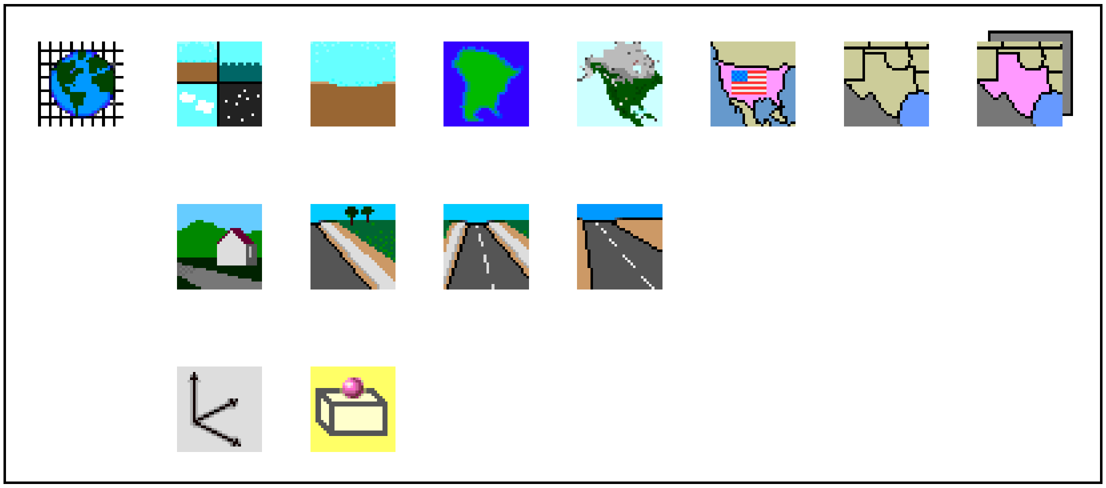
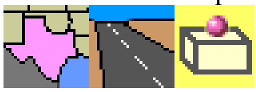


Figure 4. An Icon Path to *On Top of a Street in Texas*

- *Cascading Hierarchy with Increasing Specificity of Primitives on Subordinate Levels*

Cascading icons are organized in hierarchies from levels of generality to increasing levels of specificity. Similarly to cascading menus on the Macintosh, when a user cascades down an icon hierarchy by clicking on a cascading icon, its subordinate icons are displayed to the right of the cascading icon. These subordinate icons are arranged *horizontally* and represent an increased level of specificity. Some of the icon hierarchies cascade to as many as 7 or 8 levels deep, yet, similarly to the semantic hierarchies of the CYC Project (Lenat and Guha 1990), the design of the categories themselves and their first two or three levels is the hardest and most important representational task.

- *Compounding of Hierarchically Organized Primitives Across Multiple Axes of Description*

In many icon hierarchies in the Icon Workshop, there exists an additional form of organization. When subordinate icons are arranged *vertically*, they represent independent axes of description whose icon hierarchies can be cascaded through separately and whose respective subordinate icons can be compounded together across these axes to form compound iconic descriptors. This form of organization enables a relatively small set of primitives to be compounded into a very large and rich set of descriptors. To illustrate these forms of organization in our iconic language we can look at how the compound icon for "the scene is located on top of a street in Texas," , was created. Figure 4 shows the cascading icon hierarchy for *spatial location* extended out to the icons for *Texas*, *street*, and *on top of*, which the user compounded to create the icon for "the scene is located on top of a street in Texas."

The user clicked on the *spatial location* icon, which cascaded to show its *vertically* arranged subordinate icons *geographical space, functional space*, and *topological space*. Each of these cascading icons has further *horizontally* arranged subordinate icons each of which may go several levels deep. For example, the icons in the path from *geographical space* to *Texas* each represents a distinct level of progressive specification (geographical space --> land --> continent --> North America --> United States of America --> Southern Mid-Western States --> Texas). As indicated by the gray square behind the *Texas* icon, it too has further levels of specificity below it which can be displayed by clicking on the icon. In the Icon Workshop, at all but the terminal levels in the hierarchy, there exist many icons which themselves have further levels of specification. At any level in the hierarchy, icons can be compounded across the vertical organization to create compound icons. In addition to clicking, cascading icons can be accessed by dropping an existing compound icon onto the Icon Workshop that opens the icon hierarchies up to the terminals of the components of the dropped icon. The structure of the Icon Workshop enables a vast space of icons and their possible combinations to be easily navigated by the user.

It is also important to note that in the icon hierarchy of the Icon Workshop, the same iconic primitives can often be reached by multiple paths. The system knows the paths users take to get to these primitives; this enriches the representation of the compounds which are constructed out of these primitives. Having multiple paths allows different categorization schemes to coexist in the Icon Workshop. These multiple paths are also important in retrieval because they can guide generalization and specialization of search criteria by functioning as a semantic net of hierarchically organized classes, subclasses, and instances. This is especially useful in the organization of object icons, in which, for example, the icon for *blow-dryer* may be reached under *hand-held device*, *heat-producing device*, or *personal device.*

3.2.2. Media Time Line

The **Media Time Line** (Figure 5) is the core browser and viewer of Media Streams. It enables users to visualize video at multiple timescales simultaneously, to read and write multi-layered iconic annotations, and provides one consistent interface for annotating, browsing, and retrieving video and audio data.

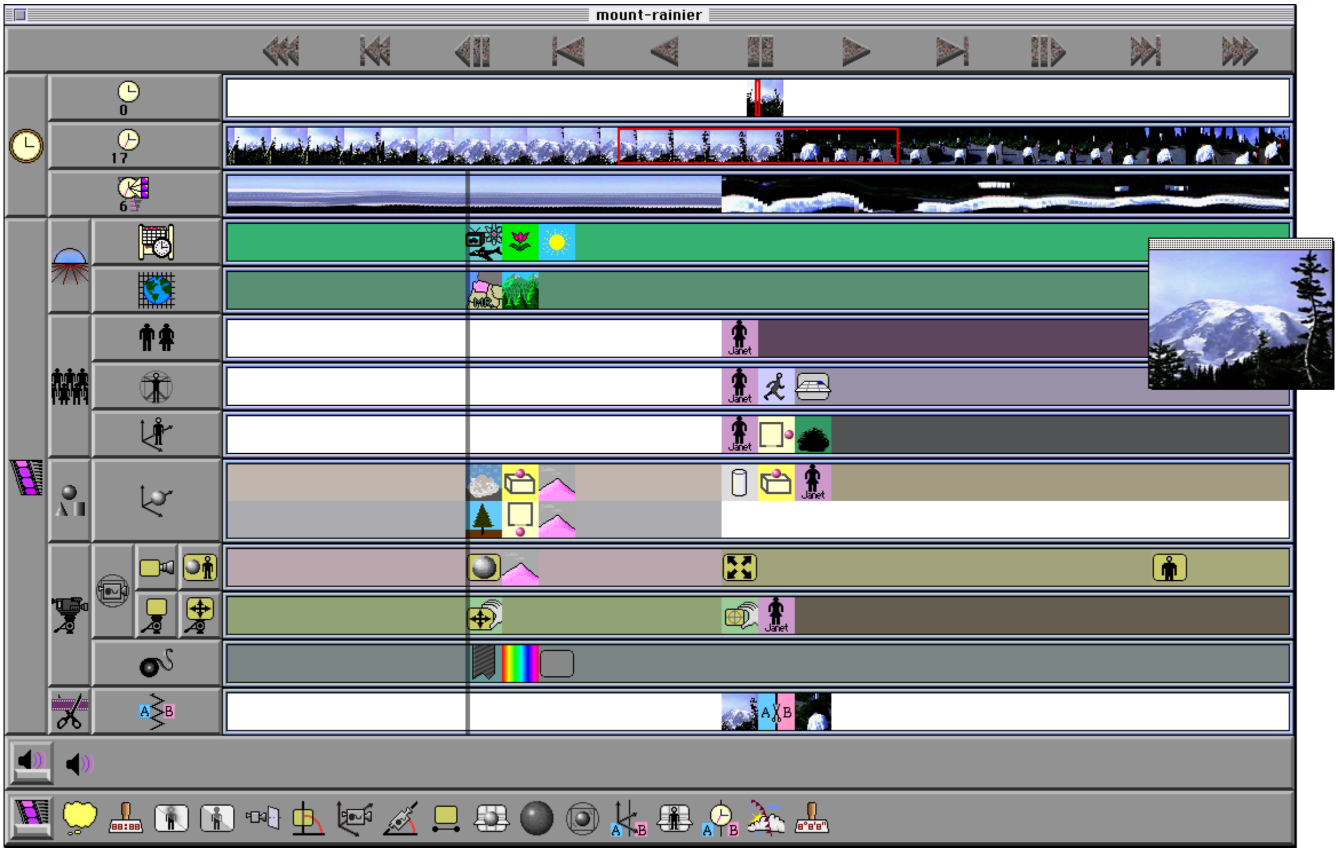


Figure 5: The Media Time Line

The Media Time Line separates annotation into various streams. These annotation streams reproduce and refine the icon categories of the Icon Space. They enable the icons used in annotation to be *viewed in context*. This design principle, like the hierarchical organization and compounding of the Icon Space, enables the large number of Media Streams' compound icons to maintain their intelligibility.

3.2.2.1. Visualizing Video Structure

Since video is a temporal medium, the first challenge for representing and annotating its content is to visualize its content and structure. In the Media Time Line we represent video at multiple timescales simultaneously by trading off temporal and spatial resolution in order to visualize both the content and the dynamics of the video data. We create a sequence of thumbnails of the video stream by subsampling the video stream at one frame per second. For longer movies, we sample at one frame per minute as well. The spatial resolution of each thumbnail enables the user to visually inspect its contents. However, the temporal resolution is not as informative because the sequence is being subsampled at one frame per second.

In order to overcome the lack of temporal resolution, we extend a technique pioneered by Ron MacNeil of the Visible Language Workshop at the MIT Media Laboratory (MacNeil 1991) and used in the work of Mills and his colleagues at Apple Computer's Advanced Technology Group (Mills and others 1992). We create a videogram. A videogram is made by grabbing a center strip from every video frame and concatenating them together. Underneath the subsampled thumbnail frames of video in the Media Time Line, the videogram represents the fine temporal resolution of the dynamics of the video with a reduced spatial resolution. However, because camera operators often strive to leave significant information within the center of the frame, a salient trace of spatial resolution is preserved.

In a videogram, a still image has an unusual salience: if a camera pans across a scene and then a center strip is taken from each video frame, a still will be recreated which is coherently deformed by the pace and direction of the camera motion and/or the pace and direction of any moving objects within the frame. Our contribution is that by presenting two different, but coordinated views of video data—the thumbnails, with good spatial resolution and poor temporal resolution, and the videogram, with poor spatial resolution but good temporal resolution—the system enables the viewer to use both representations in tandem in order to visualize the structure of the video information. In the Media Time Line, the videogram is sampled at a rate of one 4 pixel-wide strip every 1/30 second while the corresponding thumbnails appear above (outlined in a bounding box) sampled at a rate of one 32 pixel-wide thumbnail every second. This idea of playing *spatial* and *temporal* resolutions off one another is also utilized in Laura Teodosio's work on "salient stills" (Teodosio 1992) and holds promise as a general guideline for creating new

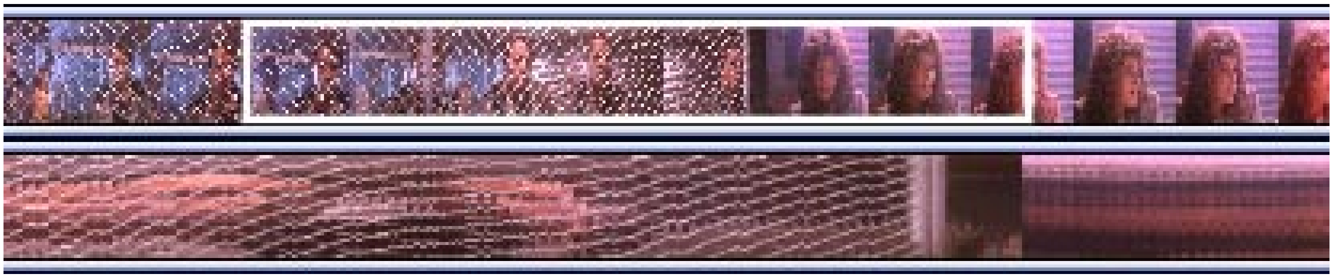


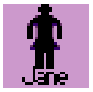
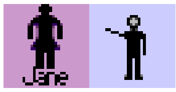
Figure 6: Media Time Line Detail — Video Thumbnails and Videogram

visualizations of video data. An example of this spatial/temporal tradeoff can be seen in Figure 6 in which the movement of Arnold through the frame is differently visible in the first thumbnails within the bounding box and in the left hand side of the videogram. In the videogram, the two swaths of extended face with the space between them correspond to Arnold's moving in the video from left to right through the center of the frame, his pausing, and then moving again.

With little practice, users can learn to quickly scan the dynamics and structure of video content from this dual temporal/spatial representation. Shot breaks are clearly visible as are camera pans, zooms, tracking, and the difference between handheld and tripod recorded video footage. Finally, the deformation of the still image in the videogram provides a coded signature of camera and/or object motion as in the example above.

3.3 Annotating Video in Media Streams

The process of annotating video in Media Streams using these components involves a few simple steps. In the Icon Space, the user can retrieve related iconic descriptors to form a customized icon palette or create iconic descriptors by cascading down hierarchies of icons in order to select or compound iconic descriptors. By dragging iconic descriptors from the Icon Space and dropping them onto a Media Time Line, the user annotates the temporal media represented in the Media Time Line. Once dropped onto a Media Time Line, an iconic description extends from its insertion point in the video stream to either a shot break or the end of the video stream. A vertical select bar specifies the current position in a movie and displays the icons that are valid at that point in time. The user can specify the end point of an annotation by dragging its icon off the select bar and can adjust the starting and ending points of an annotation by dragging the annotation's edges. A description is built up by dropping down icons for the various categories of video representation. The granularity and specificity of the annotation are user determined. By annotating various aspects of the video and audio streams (time, space, characters, characters' actions, camera motions, etc.), the user constructs a multi-layered, temporally indexed representation of video content.

In addition to dropping individual icons onto the Media Time Line, the user can construct compound icon sentences on the Media Time Line, which, when completed, are then available for use in the Icon Space and may themselves be used as descriptors. For example, the user initially builds up the compound icon sentence for "Jane waves" by successively dropping the icons  and onto the Media Time Line. The user then has the "glommed" icon  in the Icon Space to use in later annotation.

In addition to annotating video content, users can transcribe dialogue, and use the categories for video representation to describe events in the audio stream. In annotating the presence or absence of audio events, our representation makes use of the fact that in listening to audio, one thinks about the source that produced the audio. This concept correlates to Christian Metz's notion of "aural objects" (Metz 1980). Icons for different objects and characters are compounded with the icon for the action of producing the heard sound in order to annotate audio events.

In Media Streams, the interface for annotation is the same as the interface for retrieval: annotation is the process of describing footage one has; storyboarding is the process of describing footage one wants to make; query formulation is the process of describing footage one wants to find. In Media Streams, one interface is used for annotation and retrieval-by-composition.

4 Why Icons?

The most obvious and unique feature of Media Streams' user interface is its iconic visual language for video annotation and retrieval (Davis 1995; Davis 1993a; Davis 1993b). The representation and retrieval structures in Media Streams could be manipulated by many types of human-computer interface; however, the choice of an iconic visual language for this task is not an arbitrary or unimportant one. It represents a solution for the design of practical video annotation systems today as well as a statement about the future of systems for media manipulation. By decreasing the tedium and increasing the reusability of annotation effort, Media Streams' iconic visual language may solve many of the current problems of the stock footage industry whose antiquated technology and practices are inadequate to the task of on-time and accurate retrieval of video data (Greenway and Mouchawar 1994). A uniform and widespread iconic visual language for video annotation and retrieval will enable the creation of a global media archive in which video can be stored and reused.

Media Streams' iconic visual language also points toward the development of new forms of visual literacy which will become predominant in the coming age of computational media. We are currently in a crucial phase of a second "Gutenberg shift" (McLuhan 1962) 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 resequence video according to representations of its content. These representations of visual media will themselves be visual. An iconic visual language for video annotation and retrieval will support new forms of video writing (repurposing of video content) within a widespread practice of asynchronous many-to-many daily video communication.

There have been prior, pre-computational efforts to create iconic languages to facilitate global communication (Bliss 1978; Neurath 1981) and provide international standard symbols for specific domains (Dreyfuss 1972). We developed Media Streams' iconic visual language in response to trying to meet the needs of annotating video content in large archives. It seeks to enable:

- Accurate and readable time-indexed representation of actions, expressions, and spatial relations
- Gestalt visualization of the dense, multi-layered structure of video content
- Quick recognition and browsing of content annotations
- Designed visual similarities between instances or subclasses of a class (visual resonances in the iconic language)
- Articulation of the boundaries between consensual and idiosyncratic annotations (icons can have attached textual annotations and can thus function as the explicit consensual tokens of various idiosyncratic textual descriptions)
- Global international use of annotations
- Usable by illiterate and preliterate people

Media Streams' iconic language encompasses icons which denote both things and actions and thus embodies a distinction analogous to Chang's (Chang 1986) distinction between object icons and process icons. The difference here is that the objects and processes denoted by the Media Streams' icons are not computational ones, but aspects of the video content which they represent.

The iconic language gains expressive power and range from the compounding of primitives and has set grammars of combination for various categories of icons. In Korfhage's sense Media Streams is an iconic language as opposed to being merely an iconography (Korfhage and Korfhage 1986). Similar to other syntaxes for iconic sentences (Chang and others 1992; Tanimoto and Runyan 1986), icon sentences for actions have the form of subject-action-object or subject-action-direction, while those for relative positions have the form of subject-relative position-object. Icon sentences for cinematographic properties are of the form camera-movement-object (as in "the camera-is tracking-Steve" or "the camera-zooms in on-Sally").

4.1 Extensibility of the Icon Language

Currently, we have two ways of extending the iconic visual language of Media Streams beyond the composition of iconic primitives. Icons and the components of compound icons can be titled. This enables the user to attain a level of specificity of representation while still making use of the generality and abstraction of icons. For example, if I were to annotate the video of an automobile with the descriptor "XJ7," this description may be very opaque. If, however, I title a car icon XJ7, in addition to the computer learning that XJ7 is a type of car, a human reading this annotation can simply and quickly see the visual similarity between an "XJ7" car icon and icons for other types of automobiles.

Users can also create new icons for character and object actions by means of an animated icon editor. This editor allows users to define new icons as subsets or mixtures of existing animated icons. This is very useful because a wide range of possible human motions can be described as subsets or mixtures of existing animated icons.

Applying the results of work on automatic icon incorporation would also be a fruitful path of exploration (Fuji and Korfhage 1991). Already in our icon language, there are many iconic descriptors which we designed using the principle of incorporation (by which individual iconic elements are combined to form new icons). Creating tools to allow users to automatically extend the language in this way is a logical extension of our work in this area.

5 Representation Example: A Language for Human Action

The central problem of a descriptive language for 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. There exists significant work in the formalization of temporal events in order to support inferencing about their interrelationships (Allen 1985) and to facilitate the compression and retrieval of image sequences by indexing temporal and spatial changes (Arndt and Chang 1989; Del Bimbo and others 1992). Our work creates a representation of cinematic action that these and other techniques could be usefully applied to. For even if we had robust machine vision, temporal and spatial logics would still require a *representation* of the video content because such a representation would determine the units these formalizations would operate on for indexing, compression, retrieval, and inferencing.

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. 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. Lenat and Guha (Lenat and Guha 1990) point 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.). 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).

Our approach tries to address these issues in multiple ways with special attention paid to the problems of representing human action as it appears in video. It is important to note in this regard—and this holds true for all aspects of representing the content of video—that unlike the project of traditional knowledge representation that seeks to represent the world, our project is *to represent a representation of the world.* This distinction has significant consequences for the representation of human action in video. As described above, in video, 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 annotating 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.

In order to address the representational challenges of action in video we do not explicitly annotate actions according to their particular semantics in a given video stream (a shot of two people shaking hands is not annotated as "greeting" or alternately as "agreeing"), but rather according to the motion of objects and people in space. We annotate using physically-based description in order to support the reuse of annotated video in different contexts—be they cinematic or cultural ones. In our representation, we index examples of sequence-dependent semantic differences in order to represent contextual synonymy or lack thereof.

In Media Streams' user interface for action representation, our icons for action are *animated* and thus take advantage of the affordances of iconography in the computer medium as opposed to those of traditional graphic arts (Baecker and others 1991). Furthermore, we represent actions for characters and objects separately because of the unique actions afforded by the human form. We *horizontally* subdivide character actions into full body actions, head actions, arm actions, and leg actions. Under each of these categories of human action (and their own subdivisions) action is represented in two ways: *abstract* physical motions and *conventionalized* physical motions.

Media Streams' *abstract* action representation provides a hierarchical decomposition of the possible motions of the human body according to articulations and rotations of joints. Since Media Streams enables multi-layered annotation, any pattern of human motion can be described with precision by layering temporally indexed descriptions of the motion of various human body parts.

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.

One may ask "where is the representation of emotion in all of this?" If we remember the insights of the Kuleshov Effect, the answer becomes clear. Imagine a shot of a man *smiling*. Is it a "happy" shot? One might think so. But what if I edit this shot in a sequence so as to reveal that a gun is pointed at the head of the *smiling* man? Is he still "happy"? Perhaps the emotion is now better described as "fearful" or "pleading"? In both cases though the man is still *smiling*. Emotion is not a property of a shot that necessarily survives resequencing. Therefore Media Streams represents the underlying physiognomy of emotion by offering a typology of facial gestures, rather than emotions themselves which are the result of the semantics of video sequences, not of the semantics of video shots.

6 Conclusions and Future Work

In studying the semantic and syntactic properties of video we have developed both a representation and an interface which enable content-based annotation, retrieval, and repurposing. In the summer of 1994, Media Streams was subjected to an 8 person, 3-day user test that yielded promising results: we found that the system is learnable; that users reuse each other's annotation effort; and that, unlike keyword-based systems, different users' descriptions of the same footage are semantically convergent (Davis 1995). It is our hope that this technology will contribute to the creation of a stream-based, reusable, global archive of digital video. We believe that designing video representations for reusable content articulates an important challenge and opportunity for visual languages in the 1990's (Glinert and others 1991) and is the key to the development of large-scale multimedia applications in the coming decades (Davis and others 1994). Our next step is to use our system to create a large archive of annotated digital video in order to explore mechanisms for computational video storytelling. It was this goal that originally inspired the creation of Media Streams out of the necessity of having a representation of video content.

Acknowledgments

The research discussed above was conducted at the MIT Media Laboratory. 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 Wendy Buffett for enduring my dissertation.

References

Allen, James F. "Maintaining Knowledge about Temporal Intervals." In: Readings In Knowledge Representation, ed. Ronald J. Brachman and Hector J. Levesque. 510-521. San Mateo, California: Morgan Kaufmann Publishers, Inc., 1985.

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

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

Apple Multimedia Lab. The Visual Almanac. San Francisco: Apple Computer, 1989.

Arndt, Timothy and Shi-Kuo Chang. "Image Sequence Compression by Iconic Indexing." In: Proceedings of 1989 IEEE Workshop on Visual Languages in Rome, Italy, IEEE Computer Society Press, 177-182, 1989.

Arons, Barry. "Interactively Skimming Recorded Speech." Ph.D. Thesis, Massachusetts Institute of Technology, 1993.

Baecker, Ronald, Ian Small, and Richard Mander. "Bringing Icons to Life." In: Proceedings of ACM CHI'91 Conference on Human Factors in Computing Systems,1-6, 1991.

Bliss, Charles Kasiel. Semantography-Blissymbolics. 3rd ed., Sydney, N.S.W., Australia: Semantography-Blissymbolics Publications, 1978.

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

Bordwell, David and Kristin Thompson. Film Art: An Introduction. 3rd ed., New York: McGraw-Hill Publishing Company, 1990.

Chang, Shi-Kuo. "Visual Languages and Iconic Languages." In: Visual Languages, ed. Shi-Kuo Chang, Tadao Ichikawa, and Panos A. Ligomenides. 1-7. New York: Plenum Press, 1986.

Chang, Shi-Kuo, G. Costagliola, S. Orefice, G. Polese, and B. R. Baker. "A Methodology for Iconic Language Design with Application to Augmentative Commmunication." In: Proceedings of 1992 IEEE Workshop on Visual Languages in Seattle, Washington, IEEE Computer Society Press, 110-116, 1992.

Davenport, Glorianna, Thomas G. Aguierre-Smith, and Natalio Pincever. "Cinematic Primitives for Multimedia." IEEE Computer Graphics and Applications 11.4 (July 1991): 67-75.

Davis, Marc. "Media Streams: An Iconic Visual Language for Video Annotation." Telektronikk 4.93 (1993a): 59-71. (Also available at: http://www.w3.org/People/howcome/p/telektronikk-4-93/Davis_M.html)

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

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

Davis, Marc. "Media Streams: Representing Video for Retrieval and Repurposing." Ph.D. Thesis, Massachusetts Institute of Technology, 1995.

Davis, Marc, Catherine Baudin, Smadar Kedar, and Daniel M. Russell. "No Multimedia Without Representation." In: Proceedings of Second ACM International Conference on Multimedia in San Francisco, ACM Press, 181-182, 1994.

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

Dreyfuss, Henry. Symbol Sourcebook: An Authoritative Guide to International Graphic Symbols. New York: McGraw-Hill, 1972.

Eisenstein, Sergei M. The Film Sense. Translated by Jay Leyda. San Diego: Harcourt Brace Jovanovich, Publishers, 1947.

Fuji, Hideo and Robert R. Korfhage. "Features and a Model for Icon Morphological Transformation." In: Proceedings of 1991 IEEE Workshop on Visual Languages in Kobe, Japan, IEEE Computer Society Press, 240-245, 1991.

Furnas, G.W., T.K. Landauer, L.M. Gomez, and S.T. Dumais. "The Vocabulary Problem in Human-System Communication." Communications of the ACM 30 (11 1987): 964-971.

Glinert, Ephraim, Meera M. Blattner, and Chrisyopher J. Frerking. "Visual Tool and Languages: Directions for the 90's." In: Proceedings of 1991 IEEE Workshop on Visual Languages in Kobe, Japan, IEEE Computer Society Press, 89-95, 1991.

Greenway, Tom and Ronaldo Mouchawar. "A Visual Language for the Management of Digital Film and Video (Motion Imagery) Archives." In: Proceedings of 136th Technical Conference and World Media Expo of the Society of Motion Picture and Television Engineers (SMPTE) in Los Angeles, California, Society of Motion Picture and Television Engineers, Inc., 1-15, 1994.

Haase, Ken. "FRAMER: A Persistent Portable Representation Library." In: Proceedings of European Conference on Artificial Intelligence in Amsterdam, The Netherlands, 1994.

Haase, Ken and Warren Sack. "FRAMER Manual." Internal Document. Cambridge, Massachusetts: MIT Media Laboratory, 1993.

Hawley, Michael. "Structure out of Sound." Ph.D. Thesis, Massachusetts Insitute of Technology, 1993.

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

Korfhage, Robert R. and Margaret A. Korfhage. "Criteria for Iconic Languages." In: Visual Languages, ed. Shi-Kuo Chang, Tadao Ichikawa, and Panos A. Ligomenides. 207-231. New York: Plenum Press, 1986.

Kuleshov, Lev. "The Origins of Montage." In: Cinema in Revolution, ed. Luda Schnitzer, Jean Schnitzer, and Marcel Martin. 67-76. New York: Da Capo Press, 1973.

Kuleshov, Lev. Kuleshov on Film: Writings by Lev Kuleshov. Translated by Ronald Levaco. Berkeley: University of California Press, 1974.

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

MacNeil, Ron. "Generating Multimedia Presentations Automatically Using TYRO: the Constraint, Case-Based Designer's Apprentice." In: Proceedings of 1991 IEEE Workshop on Visual Languages in Kobe, Japan, IEEE Computer Society Press, 74-79, 1991.

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

Metz, Christian. "Aural Objects." Yale French Studies 60 (1980): 24-32.

Mills, Michael, Jonathan Cohen, and Yin Yin Wong. "A Magnifier Tool for Video Data." In: Proceedings of CHI'92 in Monterey, California, 93-98, 1992.

Nagasaka, Akio and Yuzuru Tanaka. "Automatic Video Indexing and Full-Video Search for Object Appearances." In: IFIP Transactions, Visual Database Systems II, ed. E. Knuth and L. M. Wegner. Elsevier Publishers, 1992.

Neurath, Otto. International Picture Language. New York: State Mutual Book and Periodical Service, 1981.

Otsuji, Kiyotaka, Yoshinobu Tonomura, and Yuji Ohba. "Video Browsing Using Brightness Data." SPIE Visual Communications and Image Processing '91: Image Processing SPIE 1606 (1991): 980-989.

Pincever, Natalio C. "If You Could See What I Hear: Editing Assistance Through Cinematic Parsing." M.S. Thesis, Massachusetts Institute of Technology, 1990.

Pudovkin, Vsevolod Illarionovitch. Film Technique and Film Acting. Translated by Ivor Montagu. New York: Bonanza Books, 1949.

Sack, Warren and Marc Davis. "IDIC: Assembling Video Sequences from Story Plans and Content Annotations." In: Proceedings of IEEE International Conference on Multimedia Computing and Systems in Boston, Massachusetts, IEEE Computer Society Press, 30-36, 1994.

Tanimoto, Steven L. and Marcia S. Runyan. "PLAY: An Iconic Programming System for Children." In: Visual Languages, ed. Shi Kuo Chang, Tadao Ichikawa, and Panos A. Ligomenides. 191-205. New York: Plenum Press, 1986.

Teodosio, Laura. "Salient Stills." M.S. Thesis, Massachusetts Institute of Technology Media Laboratory, 1992.

Tonomura, Yoshinobu, Akihito Akutsu, Kiyotaka Otsuji, and Toru Sadakata. "VideoMAP and VideoSpaceIcon: Tools for Anatomizing Content." In: Proceedings of INTERCHI'93 Conference on Human Factors in Computing Systems in Amsterdam, The Netherlands, ACM Press, 131-136, 1993.

Ueda, Hirotada, Takafumi Miyatake, Shigeo Sumino, and Akio Nagasaka. "Automatic Structure Visualization for Video Editing." In: Proceedings of INTERCHI'93 Conference on Human Factors in Computing Systems in Amsterdam, The Netherlands, ACM Press, 137-141, 1993.

Ueda, Hirotada, Takafumi Miyatake, and Satoshi Yoshizawa. "IMPACT: An Interactive Natural-Motion-Picture Dedicated Multimedia Authoring System." In: Proceedings of CHI '91 in New Orleans, Louisiana, ACM Press, 343-350, 1991.

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

Zhang, HongJiang, Atreyi Kankanhalli, and Stephen William Smoliar. "Automatic Partitioning of Full-Motion Video." Multimedia Systems 1 (1993): 10-28.