MARC DAVIS PUBLICATIONS

School of Information Management and Systems
University of California at Berkeley

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

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 se-

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.

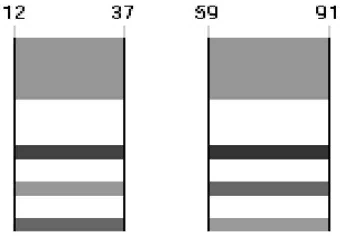

Figure 1. Two “clips” with Three Descriptors Each

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.

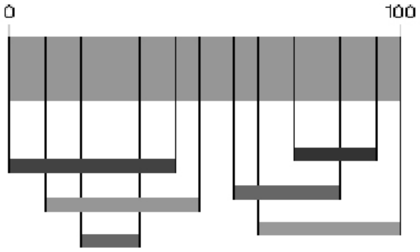

Figure 2. Stream of 100 Frames of Video with 6 Annotations Resulting in *66* Possible Segmentations of the Stream

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

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS

Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS

Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS

Sync your system to PACER to automate legal marketing.

WHAT WILL YOU BUILD? | **sales@docketalarm.com** | **1-866-77-FASTCASE**

fastcase®
Smarter legal research.