Media Streams: Representing Video for Retrieval and Repurposing

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

In order to enable the search and retrieval of video from large archives, we need a representation of video content. 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. Given the current state of the art in machine vision and image 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 technologies, and which is accessible to processing in various structured forms (ASCII, RTF, PostScript), video is still largely opaque. We are currently able to automatically analyze scene breaks, pauses in the audio, and camera pans and zooms, yet this information alone does not enable the creation of a sufficiently detailed representation of video content to support content-based retrieval and repurposing. In the near term, it is computer-supported human annotation that will enable video to become a rich, structured data type.

Over the past three years, members of the MIT Media Laboratory's Machine Understanding Group in the Learning and Common Sense Section (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 annotation and retrieval of video data. This system is called **Media Streams**. Media Streams is written in Macintosh Common Lisp and FRAMER, 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 three high resolution, accelerated 24-bit color displays and uses Apple's QuickTime digital video format. With Media Streams, users create stream-based, temporally-indexed, iconic annotations of video content which enable content-based retrieval of annotated video sequences.

The system has three main interface components: the Director's Workshop (Figure 1); Icon Palettes (Figure 2); and Media Time Lines (Figure 3). The process of annotating video in Media Streams using these components involves a few simple steps:

In the **Director's Workshop**, the user creates iconic descriptors by cascading down hierarchies of icons in order to select or compound iconic primitives

As the user creates iconic descriptors, they accumulate on one or more **Icon Palettes.** This process effectively groups related iconic descriptors. The user builds up Icon Palettes for various types of default scenes in which iconic descriptors are likely to co-occur; for example, an Icon Palette for "treaty signings" would contain icons for certain dignitaries, a treaty, journalists, the action of writing, a state room, etc.

By dragging iconic descriptors from Icon Palettes and dropping them onto a **Media Time Line,** the user annotates the temporal

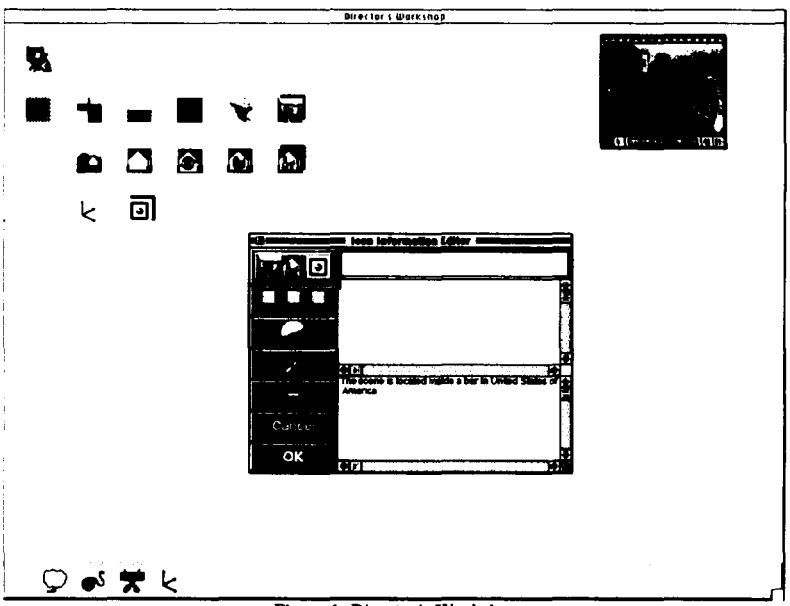

Figure 1: Director's Workshop

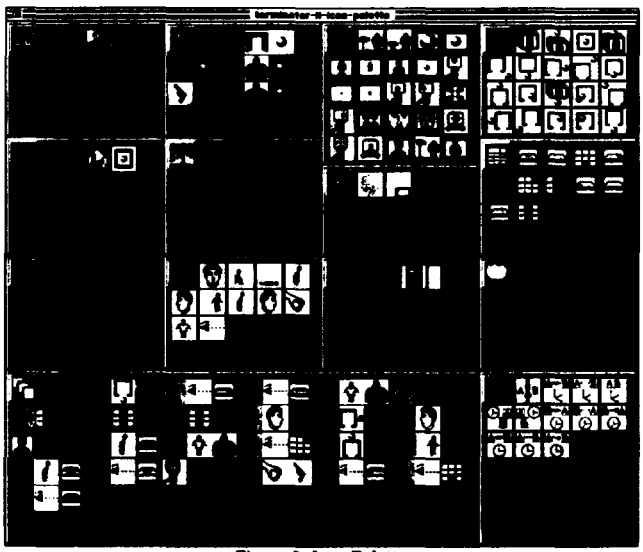

Figure 2: Icon Palette

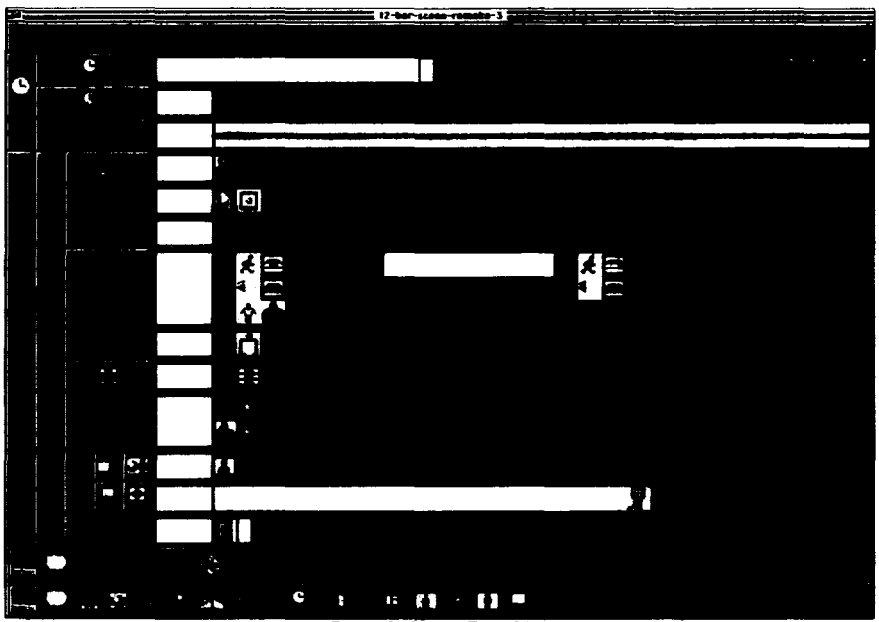

Figure 3 : Media Time Line

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 scene break or the end of the video stream. The user then ends the iconic description at the point in the video stream at which it no longer applies.

In addition to dropping individual icons onto the Media Time Line, the user can construct compound icon sentences by dropping certain "glommable" icons onto the Media Time Line, which, when completed, are then added to the relevant Icon Palette and may themselves be used as primitives. By annotating various aspects of the video stream (time, space, characters, character actions, camera motions, etc.), the user constructs a multi-layered, temporally indexed representation of video content.

The Media Time Line 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 annotation, browsing, query, and editing of video and audio data.

With Media Streams, users can create shareable representations of media content which enable the construction of large archives of reusable temporal media. Without tools like Media Streams, a thousand hours of video content will be less useful than one. With tools like Media Streams, we move closer to the day when all the digital video in the world can become an accessible and reusable resource for human and computational imagination.

REFERENCES

Davis, Marc. "Media Streams: An Iconic Visual Language for Video Annotation." Telektronikk 4.93 (1993): 59-71.

Davis, Marc. "*Knowledge Representation for Video.*" Forthcoming in: Proceedings of the 1994 National Conference on Artificial Intelligence in Seattle, Washington, 1994.