I 162 Video Streamer – Eddie Elliott I 162

Motion Image Processing: Striking Possibilities

by Andrew W. Davis

Multimedia research done by advanced groups at the MIT Media Lab points toward real-time manipulation and analysis of moving imagery

Multimedia applications are taxing the way we think about computer resources and challenging the existing structures for data storage, retrieval, editing, and manipulation. Models most of us use for filing away our drawings, still images, and text files break down when the scheme is applied to multimedia clips. Multimedia adds the dimension of time to imagery—the time axis maps to a physical property called *motion*. Processing requirements are compounded: image, sound, and text files must be linked in the database, both by association and by time, yet available individually for editing. In addition, the form and content of the multimedia data should be typed to allow reasonable retrieval from multimedia databases of the future. The environment to support such applications is being researched at the Media Laboratory of the Massachusetts Institute of Technology. Work is proceeding today on a variety of fronts; the implications for "The Future of Imaging" are profound.

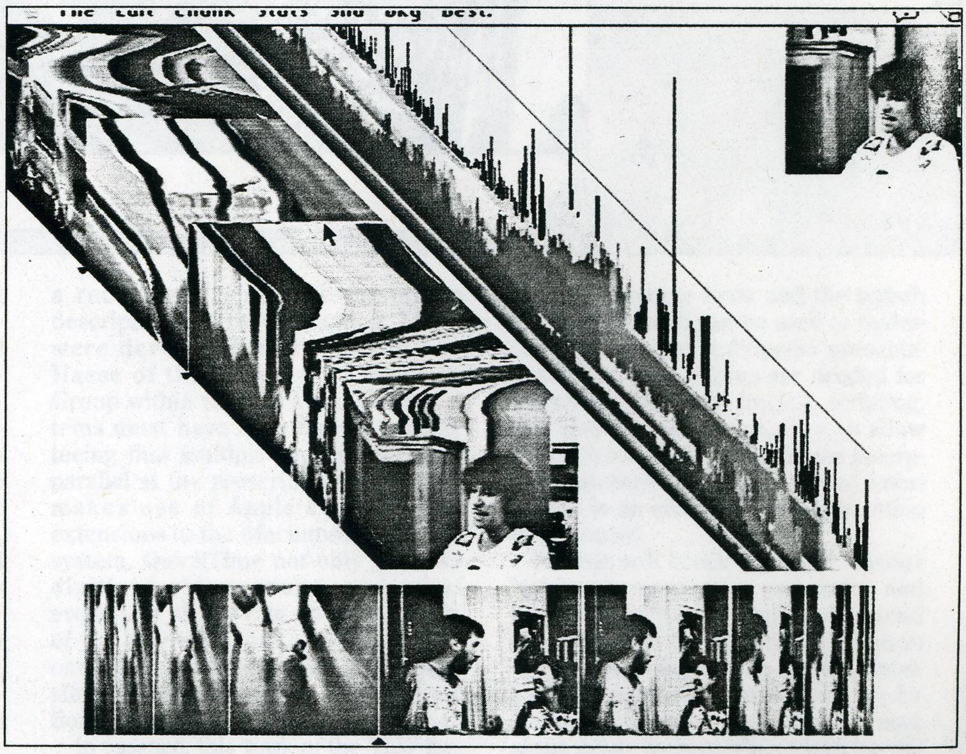
Motion video databases explain themselves

What if you could take a video clip, break it down into its individual components, and treat each one as a separate case? With advanced motion image processing tools you would be able to reuse resources from a multimedia case library while not losing track of the inter-relation between video, audio, and other files. This is the challenge being addressed by Ron MacNeil in the Visible Language Workshop (VLW) headed by Professor Muriel Cooper.

MacNeil's research is developing tools to offer the user information about the *form and content* of video data. The focus is on the *creator* of multimedia solutions and using rhythmic content to characterize audio and video motion. The VLW team is developing a set of motion image processing tools called TYRO as part of a knowledge-based multimedia presentation creation facility. These tools could play an important role in the near future when fast networks will make access to large databases more feasible and when multimedia libraries become as common as today's text and graphics databases.

TYRO is a visual programming environment that uses a case-based reasoning approach to capturing and reusing knowledge about the design of multimedia presentations. TYRO consists of four individual tools or data elements derived from the input audio/video clip (Figure 1). The thumbnail displays every 16th frame of a sequence; this is a traditional means of visualizing video, but it offers little information about rhythmic content. The videograph is a composite made by taking a 4-pixel slice from the center or diagonal of each frame. The videograph allows the user to see and annotate the *visual rhythm* of the video. The waveform displays the audio track and indicates how the breaks and pauses "synch" to the rhythm of the video. The transcription track organizes transcribed text, color coded by speaker, against pauses in the sound track.

TYRO is intended to address the needs of multimedia authors who will require design assistance and who will have access to complex data bases. The long range goal of the



All photos by Alan Blount

Figure 2. Video Streamer "electronic flip book" presents video tools in a fresh way; parsing will aid in finding frames and sequences automatically.

Andrew Davis, well-known in the electronic imaging world, has been imaging marketing specialist for several major East Coast computer and display firms; he is currently an independent marketing consultant specializing in scientific imaging on personal computers. A frequent contributor to this magazine, he can be reached at: (508) 460-6355.

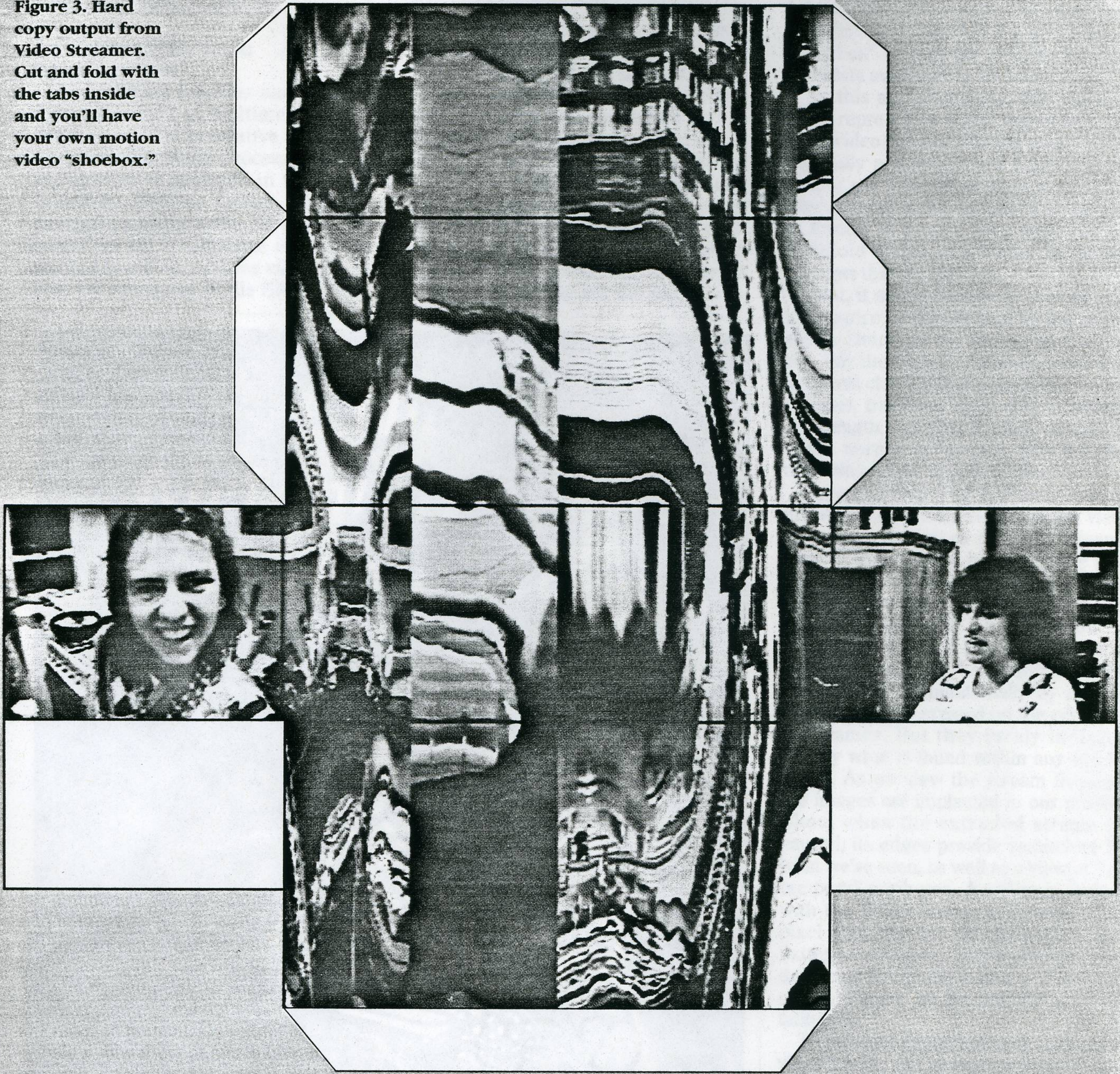


Figure 3. Hard copy output from Video Streamer. Cut and fold with the tabs inside and you'll have your own motion video "shoebox."

VLW team is to develop a tool that will automatically generate multimedia presentations, in part so that the *databases themselves can generate customized videos to explain their contents.* TYRO therefore uses an artificial intelligence model of motion image processing and uses a case approach to problem solving: by studying successful models of video, by building up a rhythmic database of what models work and what models fail, TYRO will be able to find new ways to help "authors" blend video and audio resources to meet future needs.

TYRO is built in Apple's Macintosh Common Lisp and uses an experimental representation language as well as a recursive framework for media description and representation. These were developed by Professor Ken Haase of the Music and Cognition Group within the lab. Multimedia systems must have a facility for guaranteeing that multiple events happen in parallel at the prescribed time. TYRO makes use of Apple's QuickTime extensions to the Macintosh operating system. QuickTime not only can play digitized video, sound, and other event files at any rate up to the limits of the hardware, but also can coerce one rate to another, which makes possible both unidirectional and bi-directional constraints.

In essence, this work of the VLW has two facets: development of motion image processing tools, and the search for models which can be used to understand and create multimedia presentations. Tools and editors are needed for searching, browsing, linking, scripting, and visualization capabilities to allow research into new design issues emerging from real-time, multi-layered information in an electronic communication environment.

Models will break down the barrier between programmers and users, and permit users to customize and extend software. The combination of motion image processing and an artificial intelligence approach to programming by example will enable the computer user of the future to generalize a multimedia program from a database of acceptable

All photos by Alan Blount

examples and to filter, prioritize and design information on-the-fly.

An electronic flip book: brand new view

While TYRO provides one unusual method to look at multimedia clips, another equally innovative approach toward motion image processing is the foundation of research in the Lab's Interactive Cinema Group. Using an electronic analogy to the old paper "flip books" in which individual frames are stacked in a deck, *extruded video*, developed by researcher Eddie Elliott under the supervision of Professor Davenport, flows video along a time axis to give video a spatial context useful for subsequent processing.

According to Elliott, the Interactive Cinema Group is exploring ways to present a moving picture stream and to organize multiple video streams in order to assist the *viewing* experience in addition to the production. In its current incarnation, the Video Streamer accepts video frames from a frame grabber and stacks them on the Macintosh screen in such a way that the current image is on the top and the stack pushes images diagonally off to the upper left over time (Figure 2). The result is a streamlined view of a video clip, a "shoebox" shaped file of video images, with data available on all six surfaces.

To this author at least, this block of video represented a new way of presenting a video stream. Moving pictures are normally presented in a rapid succession of still frames that produce the illusion of motion. We usually view strings of frames as though they were still on celluloid strips; the Video Streamer changes this perspective.

First, it stacks frames so that we can view more of the stream in the same space. Lining up the frames so that we see only the edges, we become aware of temporal characteristics such as camera motions, transition types and locations, shot length, and shot rhythm, perceived in the normal full-screen view but are not visible in the film-strip view.

Second, since the video stream is inherently dynamic, the extruded view also presents itself as a flowing stream. Note in the example how the edge patterns in the "shoebox" extruded view give clues to the characteristics of the stream's contents. The duration of shots as well as the types of transition between them can be identified both visually and computationally by analyzing the patterns created on the edges of the frames. But they rarely indicate exactly what is found within any single frame. As we view the stream flowing, the images are implanted in our minds. Then, when the extruded stream is paused, its edges provide reminders to what we've seen, as well as a view of the temporal attributes. A utility included with the Video Streamer arranges the pixel data into the six rectangles of a flattened shoebox. When printed, the data can be formed into a physical model (Figure 3): life imitates art imitating life.

To facilitate coordination of audio and video tracks, and to work with a file format that is rapidly becoming a standard for multimedia, the Interactive Cinema Group, like the VLW, is using Apple's QuickTime. The Video Streamer software allows the user to easily browse backwards and forwards through time. The cursor can be used to select any frame in the stack as the starting point. Audio tracks are keyed to the video frames, so that when the video is played backwards, the audio is played backwards at the correct rate while the words are spoken "forwards."

Video parsing: Find that change

While the computer is accepting QuickTime video inputs, it is also calculating a 64 bin histogram for each of the R, G, and B values of the pixels along the top and left edges (total = 220 pixels), shown in the example.

The Video Streamer does this for each frame, subtracts the histograms for each adjacent frame, then compares these differentials from frame to frame to get a second differential characteristic. This is evaluated against a threshold for detecting "probably cut" frames. High values are tagged on the 3-D video streamer display. This assists users in automatically parsing a video stream into distinct shots.

The existing video parser was intentionally kept simple in order to minimize the CPU loading and to allow the Macintosh to perform tagging in real time, along with the other QuickTime processing. A more elaborate video stream parser could conceivably watch for signal characteristics that signify other types of transitions besides cuts, such as dissolves, fades, wipes, and keys. It might also be tuned to recognize camera motions such as pans and zooms.

A more sophisticated parser that incorporated a variety of characteristics (such as spectral analysis) would need to employ some weighting algorithm to arrive at a difference value or a collection of values. If compression algorithms are used—certain in future commercial implementations of this technology— then the compression software may provide additional useful parameters for the parsing algorithm.

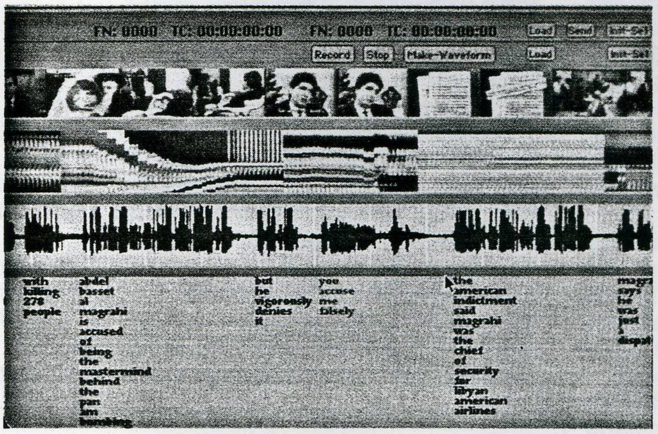


Figure 1. TYRO visual programming environment: toward multimedia databases that explain their own content with automatically-generated video.

Parsing is likely a good use of standard statistical analysis, artificial intelligence, and/or neural network computing. Further clues to the video stream may be available by analyzing the accompanying audio stream as well.

Volumetric motion video content analysis

While Eddie Elliott's research is focused on the information available on the surface of the Video Streamer, Professors Edward Adelson and Alex Pentland in the Vision and Modeling Group are looking at motion image processing within a volumetric framework (*see this issue's cover*). Continuous frames of a video sequence are stacked up like a deck of cards to form a 3-D volume with X,Y, and T (time) axes. Different perspectives of the spatial-temporal volume reveal motion patterns not otherwise perceptible in any single 2-D video frame.

The effect is similar to stacking a series of 2-D CAT scans to form a volumetric model, and then taking arbitrary planar cuts through the volume. *X-Y-T analysis of video has numerous applications for motion understanding and object identification.* A slice along an X-T plane shows patterns indicating object motions, the rhythm of certain events, and the intersection of frame contents.

It is also possible to use other principals developed in the world of voxel processing, such as applying trans-

(continued on page 56)

CARLOTTO

(continued from page 28)

tunately, the objects must usually be identified and extracted manually. Since IU systems capable of automatically identifying objects, relationships, and events do not yet exist, more opportunistic approaches are required.

The following are several representations and techniques we currently have under study at TASC for image retrieval:

- *Global Energy Distribution*: Images are represented by their energy distribution as a function of frequency using pyramids. Uses "query by example" to find similar images.
- *Local Feature Matching*: Images are represented by vectors of local features (intensity, color, texture, position, etc.). Matching algorithm attempts to find for each feature in one image the most similar feature in the other image using a constrained search algorithm. Uses "query by example" to find similar images.
- *Intermediate-Level Indexing*: Images are represented by vectors that describe the properties of regions (e.g., color, size, and shape). Retrieves images that contain regions with specified properties.
- *Shape Matching*: Uses generalized Hough transform to represent edge information in images. Matching algorithm finds those images in database that contain similar shapes (matching algorithm can be made insensitive to differences in rotation and scale).

The first two methods show promise in retrieving images that look the same. The third technique appears promising for retrieving maps, color, and multispectral imagery. The last technique is a variant of template matching.

Mtool is a mensuration tool developed by TASC which provides a graphically-oriented interface to a powerful set of photogrammetric algorithms. These algorithms use the dimensions of known 3-D objects in the scene to convert the 2-D positions of other objects into 3-D models. Mtool provides access to both imagery and the elements of the models. It supports the extraction of image features, the display and overlay of models and images, and estimation of modeling errors. Here we show Mtool being used to model several satellite dishes using two uncontrolled ground photographs.

Markets, potential

As is the case with all of the Image Understanding development lines, many potential markets for content-based image database technology exist in areas such as defense, biomedical imaging, digital photography, digital media database development, the arts, and others.

Since IU systems capable of identifying arbitrary objects in general imagery do not yet exist, the greatest potential for success is in the development of solutions tailored to particular domains or image characteristics. Near-term success will be on methods that extend existing DBMS technology to images (e.g., to retrieve images by attached freetext), operate within a limited domain (e.g., multispectral imagery), or have limited capabilities (e.g., to retrieve images that contain specific shapes).

Even the selected areas covered in this article suggest that IU is and will continue to be a highly interdisciplinary field. Emphasis will continue to be on the development of specific capabilities for near-term applications, but will continue to be guided by the vision of developing intelligent systems that can really see. ■

DAVIS

(continued from page 25

parency effects to give viewers a totally new way of seeing video and viewing motion. When time is mapped to volume and volume is made transparent by the computer, video information takes on new meaning.

Motion image processing/analysis

By its very nature, multimedia, which contains time sequenced frames of audio and video information, will expand the amount of data available manifold. The challenges of dealing with multimedia resources are significant. Research today is exploring methods to access and manipulate multimedia databases. The focus at the Media Lab is on developing models to improve understanding and then to use these models on ever more powerful computers which can deal with large databases, complex modeling schemes, and provide real time performance. Programs such as TYRO provide multimedia browsers and editors and are now being used by designers to build case libraries which represent high level chunks of design knowledge.

TYRO and the Video Streamer use motion image processing to address various aspects of multimedia viewing, creation, and analysis. *It is likely that any multimedia-based activity in the future will involve a mix of these requirements.* For example, editing involves both reviewing raw material and manipulating it to create more refined materials. Such situations are encountered while viewing hypermedia documents, while logging footage prior to editing or for entry into video databases, and during film analysis.

Whereas current commercial products let users create catalogs of still images, QuickTime movies, and sounds, using thumbnails of graphics and keywords, the research at the Media Lab is using rhythm and content to not only catalog resources, but to aid the user in reusing those resources. The emerging tools will tie together macro views of the video stream, micro views of the contents, and abstract views of the audio/video signal. Motion image processing will also provide keys to archival retrieval programs, content-based editors, and other systems which can take advantage of knowing the content of recorded video. And the tools developed along the way inevitably have implications for the way we'll view and manipulate motion imagery as well. ■

F R A M E S

Video Streamer - Eddie Elliott

April 1992
Number 7

A Monthly Publication for Sponsors of The Media Laboratory

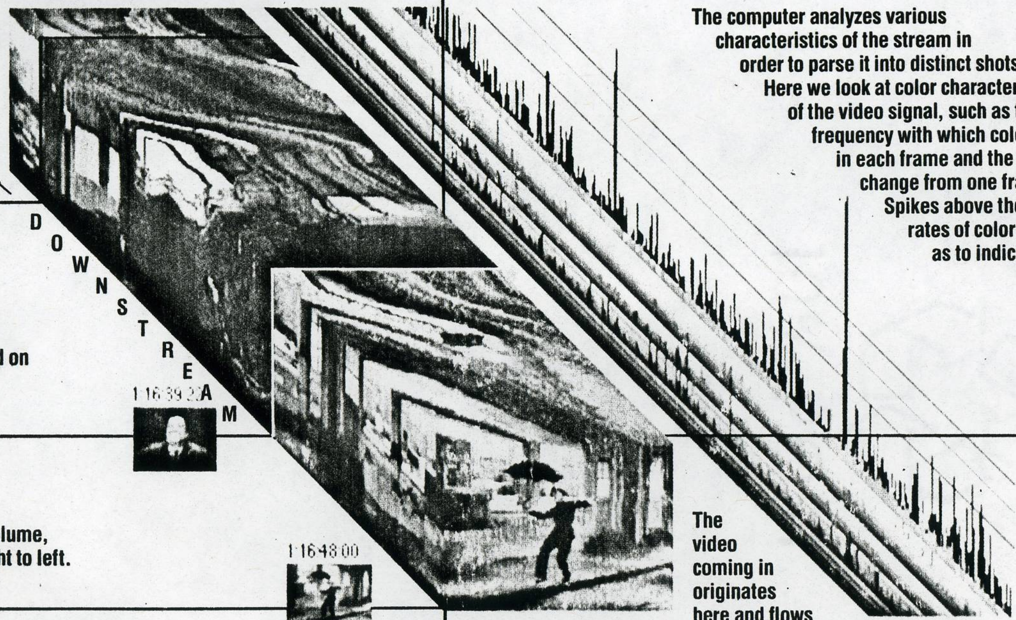
Massachusetts Institute of Technology

AND SHOWING THE

Extruded video, developed by Interactive Cinema Research Assistant Eddie Elliott under the supervision of Professor Glorianna Davenport, flows video along a time axis to give it spatial context.

The computer analyzes various characteristics of the stream in order to parse it into distinct shots. Here we look at color characteristics of the video signal, such as the frequency with which colors occur in each frame and the rate of color change from one frame to the next. Spikes above the red line indicate rates of color change so significant as to indicate likely shot boundaries.

1:16:27:10



The duration of shots as well as the types of transitions between them can be easily identified both visually and computationally by analyzing the patterns created on the edges of frames.

1:16:39:23

As this shot tracks Gene Kelly down the street, features of the background smear across the top of the video volume, indicating a camera pan from right to left.

1:16:48:00

The video coming in originates here and flows downstream in time toward the upper left.

FLOWING THE SHOW

The digital video signal is a bit stream that can be manipulated mathematically. We are developing various tools to help us **feel video**, and to **teach the computer** about important relationships within the video stream.

We can form video into shapes and slice it — literally — up, down, and sideways to better understand its content and the cinematographic style of its creator.

Our expanding **digital video toolkit** will make possible the creation of new media forms, including video collages and real-time interactive multimedia experiences in which the computer automatically offers up information in response to a user's needs.

We can use the cursor to scroll around within the captured video stream. The large, center frame shown by this utility indicates where the cursor is. Previous frames are to the left, while future frames lie to the right. This particular view shows the last frame of a close-up of Gene Kelly just before the cut to a wide tracking shot.