# STRUCTURED CONTENT MODELING FOR CINEMATIC INFORMATION

**BENJAMIN RUBIN**
**GLORIANNA DAVENPORT**

### Structured Content Modeling for Cinematic Information

For cinematic material to become useful as an online information resource, a structured content model must be developed. This model should enable both a viewer and an automated retrieval/presentation system to navigate and manipulate picture and sound information. Here we present the functional requirements for such a model, and then discuss a video editing and viewing environment, currently under development at the MIT Media Laboratory, which utilizes content information.

Imagine the following scenario: a viewer is watching a videotape of an interview. The subject of the interview makes a point which interests the viewer, who decides she wants to use that piece of video to illustrate a paper she is writing. The viewer wants to say "grab *this*," and would like to have the computer pull out the appropriate piece of video for her without her having to laboriously locate the exact start and end of the segment.

What kind of content structuring is necessary for this to happen? Before addressing this question, it is interesting to observe that there are few models in the retrieval and manipulation of text information which perform comparable functions. Text *is* content in the context of computer information systems; text searches of varying levels of sophistication are generally sufficient for locating both documents of interest and specific points of reference within documents.

Text documents use well-established conventions to assist navigation and retrieval of information: paragraphs, sections, chapters, headings, and even typography help to delineate "chunks" of content, and indices and tables of contents facilitate fast location of topics. In films, transitions are not always explicitly marked; in place of headings and typography, films use a complex set of visual and auditory conventions to indicate shifts in topic or focus. Yet when examined closely, most any film, regardless of content or genre, can be broken up into parts which, while not completely analogous to their textual counterparts, still correspond approximately to paragraphs, sections and chapters.

The task of a structured content model, then, is to represent the natural hierarchy of cinematic material. Given a single frame from a film, the model should provide several levels of context: the frame is from a shot showing a crowd of civil rights marchers; the shot is part of a scene which depicts a rally in Selma Alabama; the scene is from a section of the film which deals with George Wallace's opposition to school desegregation; the section is part of a documentary about school desegregation in the South, and the documentary is part of a series on the Civil Rights movement.

A start and end point, set of characters, locations, actions, issues, dates and events should be associated with every entity defined in the content model. In addition, the model should include a notion of hierarchical inheritance. Even though, taken out of context, the Selma rally scene in the example above does not explicitly depict George Wallace, a query seeking material to illustrate protest against George Wallace should find it, since it comes from a section of the film about him.

Thinking about this kind of content model immediately raises a host of difficult questions. Who creates the model? Are there not many possible ways of break-

ing a film up into parts? Can several content models exist side by side? Similarly, many theoretical issues are raised by the notion of representing visual and auditory media using written language. How can we insure that a query for "demonstrations" or "opposition" or "violence" would find the scene above if it had been described to the content model as a "protest rally?"

**Current Work at MIT**

In building our video viewing and editing environment at the MIT Media Lab, we are testing assumptions about some of these questions. We are centering the environment around a simple content model. Our first test data consists of a three-hour documentary about the redevelopment of the downtown waterfront in New Orleans. Produced by Film/Video Section professors Glorianna Davenport and Richard Leacock, the film provides a rich, in-depth look at the political, financial, ethical and personal issues surrounding the renovation of an historic brewery, as well as the financial collapse of the 1984 Louisiana World Exposition and the effects of this collapse on the city. Viewed from beginning to end, the film stands on its own as a close, personal portrait of the decisions which shaped urban change in New Orleans.

Professor Davenport is now in the process of building a content model for the film. The lowest-level unit of the model is called a *segment*, and consists of a contiguous group of video frames. Segments are arranged in hierarchies of arbitrary depth and are grouped into *scenes*, which are loosely defined as coherent groups of shots which share a common location, time, set of characters, and subject matter. Associated with each segment and hierarchical grouping of segments (some of which are marked as scenes) is information regarding characters, actions, events, locations, issues and dates.

In conjunction with the content model, we are building a combined viewing/editing environment based on Project Athena's visual workstation. The workstation environment will allow viewers to watch and control video, as well as to pull out, define, and save segments of their own. The environment also provides editing tools, so that segments can be re-combined to form new sequences. At any time during viewing or editing, the content model is available to provide background information and assistance. The user will, for example, be able to say "grab *this* segment" and based on information from the content model, the editing tool will be smart enough to present the user with its best guess as to what "this" is.

Among the interface issues we are exploring in the workstation is the use of visual mnemonic icons to represent segments and scenes. We have found that if the viewer is at all familiar with the material she is working with, then a visual mnemonic, usually a reduced-scale single frame from within the segment, functions much better than a textual signifier to identify the video segment and in certain cases can provide the best mnemonic for text data.

Organizational tools let the user graphically arrange and save unordered groups of segments in palettes for later use. A time-line representation allows the user to edit segments together, forming new sequences which can be played back seamlessly. The content model will eventually be capable of returning both palettes and sequences in response to content queries.

To supplement the content model, we are also examining ways of associating visual and spatial information with each segment. We belive that data concerning camera position and motion, angle of view, character positions and directions of motion can allow the machine to assist the editor in selecting segments which cut together smoothly. To test this hypothesis, Film/Video Section graduate student Ben Rubin, who is working on the design of the editing system, is producing a videodisc containing un-edited material from which a number of short, dramatic scenes can be created. Spatial data about camera and character positions was collected during the production; the data will be integrated within the content model.

Ultimately, through the existence of a content model, spatial information, editing rules, and true random access to video and audio, we hope to create a new environment for exploration, study and manipulation of complex video material which preserves at all times the sense of a coherent, seamless cinematic experience.