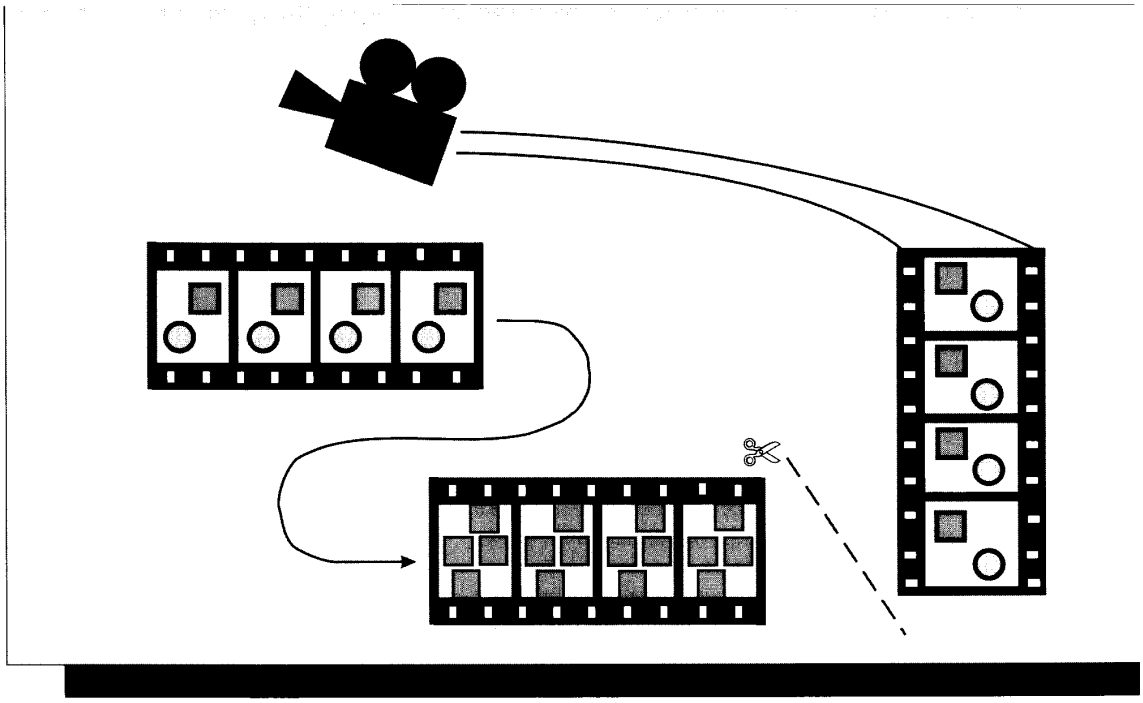

Cinematic Primitives for Multimedia

Glorianna Davenport, Thomas Aguierre Smith, and Natalio Pincever
MIT Media Laboratory

Shots, sounds, and sequences advance the storytelling in film. Stratification helps us use these elements in multimedia.

Throughout its history, cinema has challenged filmmakers to experiment with new storytelling modes. Interactive multimedia proposes that the participant viewer can affect selection and sequencing of cinematic story elements. Our challenge is to develop robust frameworks for representing story elements to the machine such that they can be retrieved in multiple contexts. We propose that content can be represented in layers, or *strata*. This model for layered information will allow programs to take advantage of the relation between cinematic sequences and the world they represent.

Interactive storytelling

Interactivity between participant users and machine representations is an integral element in today's multimedia computer environments. Thus, interactive multimedia is a user-directed form of storytelling. Whether we talk about hypermedia publications, digital video editing systems, real-time collaborative exchanges, or personalized access to large on-line video servers, we want to offer participants the most natural possible access to information. Computer programs coordinate the interaction by mediating between user input and segments of multimedia data. User navigation through multimedia systems and successful computer mediation of the data in such systems depends on an effective representation of the content.

0272-17-16/91/0700-0067$01.00

Today, after 12 years of developing laboratory prototypes, companies have begun to distribute commercial multimedia applications. ABC News Interactive in New York, WGBH in Boston, Voyager in Santa Monica, Synapse Technologies in Los Angeles, and Time-Warner New Media in Los Angeles number among current publishers of interactive video programs. Museums increasingly commission interactive media pieces, including the Exploratorium in San Francisco, the Baseball Hall of Fame in Cooperstown, New Jersey, the Computer Museum in Boston, and the Smithsonian in Washington D.C. While many of these applications constrain user navigation by adherence to the author's preprogrammed cues, others are more open ended and encourage browsing.

In browsing, users' personal goals and intentions may collide with the system's ability to foster coherent meaning; viewer comprehension tends to break down, therefore, when participants begin browsing. While this breakdown is often blamed on weaknesses in the navigational tools, it more frequently reflects the lack of a semantic representation that can support the limited look-ahead functions required to build meaningful interactions with the user.

Increasingly, we will need to create multimedia systems that maintain a conversational mode of interaction with users by generating and tracking story frameworks. In the case of online video servers as well as home-movie editing assistants, the machine must respond to the user by selecting the "best" shots, sounds, and text chunks, then orchestrating or sequencing them to emphasize a particular story. The story content should reflect the user's background and intent.

Factors that affect multimedia's conversational success include the quantity of video and audio available in a system (largely solved once low bandwidth digital video becomes ubiquitous), machine-readable representation of the program content, articulation of complex narrative models, and natural language understanding. Significant advances in all of these areas continue to increase the power of multimedia information and simulation systems.[1-3]

Developments in cinematic storytelling must join technological advances in multimedia if we are to produce interesting machine-mediated experiences. The history of cinema offers a sensual and provocative record of the evolution of an expressive form. A rapid overview of this history in light of technical innovation allows us to examine the traditional levels of granularity—that is, dividing a stream of information into meaningful chunks—that have been used for the moving image.

The problem might be quite obvious but bears restating: For a machine to select and sequence video footage, the content of footage must be represented in machine-readable form. Today, no machine-readable information, with the exception of SMPTE (Society of Motion Picture and Television Engineers) time code, accompanies motion picture images as generated. For the purposes of archiving and editing, logs are sometimes created after collecting the footage. Such logs are time consuming to construct, suffer from a range of information errors, and, most importantly, lack the rigorous methodology needed for computer representation. To realize our vision of conversational multimedia, we need to attach known descriptive attributes to specific shots and sounds as we record them.

Some useful terms

granularity
Granularity refers to the descriptive "coarseness" of a meaningful unit of multimedia information. For image processing, the descriptive granularity has to be fine enough to address specific objects in the frame. When editing a movie, the granularity must be coarser to encompass thoughts, actions, and intentions.

chunks/segments
By chunks we mean segments of arbitrary length, usually with some semantic significance. The coarseness of these chunks determines the granularity of the system.

log/logging
Logging is the process through which shots in a piece of footage are cataloged, usually by in- and out-points, length, and a short description. Such a catalog is called a shot log, or log.

interactive multimedia
We like to characterize interactive multimedia as a user-directed form of conversational storytelling. In multithreaded narratives, the user affects the "plot" of the story through choices along the story space.

shot
This is the smallest addressable unit having the finest level of descriptive granularity. It consists of one or more frames generated and recorded contiguously and representing a continuous action in time and space. The single frame is the smallest meaningful entity.

sequence
A collection of shots no longer perceived as a set of individual shots, forming a natural unit. This unity is the result of continuity in various planes: temporal, spatial, perceptual, and so forth.

An extensive glossary of terms for film and media can be found in James Monaco's *How to Read a Film: The Art, Technology, Language, History, and Theory of Film and Media*, Oxford University Press, New York, 1977.

Cinema and storytelling style

What makes a movie complete? What distinguishes a scene from a sequence? How do sound and picture complement each other to produce a seamless structure? With each edited production, narrative or documentary, filmmakers gain new understanding about how the pieces of cinema—shots and sounds—go together and how their progression affects the audience. As technology evolved, filmmakers invented new aspects of cinematic language. In this process, they reshaped the relationship between specific cinematic elements, the final movie, and the audience.

The silent era

In the beginning, motion pictures were shown without sound. Eisenstein's theory of montage emerged during this time. In the montage approach to editing, the narrative results from a carefully structured linear ordering of shots. The meaning arises from the associations made in the viewer's mind while watching these visual juxtapositions.[4] Sets of shots compose a visual sequence that frequently suggests a psychological state or thought progression. This storytelling method relies on metaphor more than on continuity of action, on associative context rather than on directed context.

Filmmakers of this era who needed to develop the action frequently used text cards to further the plot. They shot all images with prime lenses, since zoom lenses were not yet a staple of the setup. The idea of visual, spatial, and temporal continuity developed slowly.

Movies with sound

As theaters became outfitted for projection, presenters began to accompany movies with live music, most frequently piano, selected to induce particular reactions from the audience. These live accompaniments were usually free-form, not specifically rehearsed or synchronized to the action. As soon as it became technically possible to lock sound to the moving image for projection, sound became the ubiquitous partner of the motion picture.

The advent of sound heralded the death of montage and the rise of movies that depended on visual continuity and specific dialogue to further the action. While the audience perceived this track as synchronous with the picture, most release tracks for narrative film were (and still are) created in dubbing studios and foley pits during the postproduction process.

As the tools for production and presentation improved, so did the language of the moving image. Directors began to use perspective as a vehicle for manipulating audience reaction to story elements. The camera became coconspirator with the creative vision of the cinematographer, who demanded increasing flexibility. New lenses were invented, and the pressure was on to make the camera more mobile, as well as to discover new, faster film stocks. These advances in technology led to the construction of much more sophisticated sequences. Movie directors became fascinated with realism and almost overnight invented conventions that allowed viewers to build cognitive maps of the story space. The development of higher film speeds had a profound effect on the evolution of new cinematic techniques.[5]

Throughout the evolution of cinema, directors, cinematographers, and editors sought to discover ways in which cinematic language could both camouflage and create a perceptual awareness of time and space. Whether match cut or jump cut, the relationship between the last frame of one shot and the first frame of the next shot generated new theories of continuity and ideas for sequence construction.

Descriptions of content must be structured at the shot level to maximize the potential for computer-aided browsing and sequence assembly.

Granularity of meaning

In motion picture parlance, syntax refers to how action is framed in individual shots and how those shots can be ordered into sequences. While the audience experiences linear movies as unified entities, filmmakers experience movies as the generation of individual shot and sound elements and the forging of relationships between these elements.[6] At each step in the production process, key professionals—directors, cinematographers, editors—contribute their knowledge and vision to the design of shots, sounds, and sequences.

In multimedia applications, especially those that draw on centralized databases of visual information, we will need to adjust levels of granularity as we incorporate the same piece of footage into different contexts. For instance, if we want to learn about recent events in Poland, we might choose to view sequences introducing us to the country, its language, and its culture. Later, we might want to build a country report using information about the country with recent political news. Clearly, we need descriptions that present us with longer or shorter elements according to the detail our query requires. Descriptions of content must be structured at the shot level to maximize the potential for computer-aided browsing and sequence assembly.

A shot consists of one or more frames generated and recorded contiguously, representing a continuous action in time and space. In most instances, the beginning and end frames of a recording medium are used as an address for shots. We can compute the duration of a shot from these two numbers and a designated playback speed. Obviously, the original shot generated in the camera can be redefined many times during editing.

Frequently a passage of sound is recorded along with the shot (either on the same or on a separate medium). We refer to this as *synchronous sound*. Synchronous sound consists of semantic dialogue and ambient sound. If this synchronous sound track is not physically bound to the picture (as in film or digital modes, which use separate picture and sound servers), it must carry some identifying tag to guarantee that it will maintain its synchronous relationship with the picture despite adjustments during editing. If you want to be able to relocate the original shot or synchronous sound passage from an edited program, a trace identifier must be included as part of the descriptive data set.

We urge you to contemplate the role that the environment serves in its several guises. The environment serves as the glue we can use to generate a description for all aspects of the shot.

We can describe a shot in four general ways, each linked to the others:

- Perspective: Whose point of view does the image represent?
- Camera and sound recorder: What is the camera or microphone doing relative to the scene? Or how is the camera or microphone mediating the environment around it?
- Content: What is happening in the environment as it is captured in the image and sound?
- Context: What additional meaning is represented in the shot based on its adjacency to other shots or to related world knowledge?

Figure 1 displays a map of the shot. This map indicates how much information we, as editor or viewer, must understand to make sense of the image before us. In this global view, we extend each component of the shot—perspective, camera, content, and context—to show the extraction of this aspect. We urge you to contemplate the role that the environment serves in its several guises. The environment serves as the glue we can use to generate a description for all aspects of the shot. It is a cross-reference point for our structured model. By rotating a three-dimensional layout of the model, we should be able to superimpose the several instances of environment and move between the levels of data corresponding to each of the aspects.

Collectors of content: Camera

While Figure 1 lays a framework for understanding the complexity of the shot, it does not clearly reflect its temporal nature. In reality, all of these elements are subject to change in time—as the shot is generated. This makes the idea of describing the image appear as a gargantuan task. However, if we turn to Ricky Leacock's description of a movie as "the filmmaker's perception of what takes place in the presence of a camera when the camera is turned on," we are reminded that what is captured on film involves the relationship between what happens in a given environment and the physical presence of a camera as it records over time in that environment. We can exploit the temporal nature of this process to track changes in state of both the camera and the images that the camera records over time and to retrieve this description with variable granularity.

This argument considers the camera a physical object, situated in space and time. The camera records images through a single lens to a discrete image plane. When we begin a shot, we know the date and exact time. We know the camera location, lens, and direction of view (more or less exactly, depending on our level of instrumentation). Moreover, the camera person knows the general intention of the filming and can identify, with some precision, the significant people and objects in proximity to the camera. We refer to this surround as the ambient reality of the camera (see Figure 2).

The physical act of the camera motion, of light passing through the lens to the image plane, and of a specific number of frames being generated over time, imply that the ambient reality established in whatever detail at the head of a shot will, if properly encoded, be constant throughout the shot except as changes of state occur. Changes of state happen either to the recording equipment or to the objects inhabiting the particular environment in which we are recording. Therefore, we can represent information about the ambient reality to the computer over time in as much detail as we wish.

Stratification defines a model of layered information we use to describe what is happening in a shot.[7] We can attach a range of attributes including frame delimiters, who, what, when, where, why, and so forth over the temporal span of the shot. We can attach more generalized descriptions to the scene level or program, and these will be inherited by the shot. Likewise, the program as a whole can inherit information specific to a shot.

To keep better track of state changes, we are currently constructing a graphical logging device that will allow us to annotate the layered information related to video footage with greater ease. The variable granularity will allow users or programs to query the database and to identify shots and sequences in a temporally appropriate scale. The model we suggest links inheritance to granularity.

Collectors of content: Microphone

As with the camera, the microphone or sound recorder mediates the environment in which it is situated. Ambient sound offers us clues about the physical and geographic aspects of place. When parsed, the original synchronous sound track for

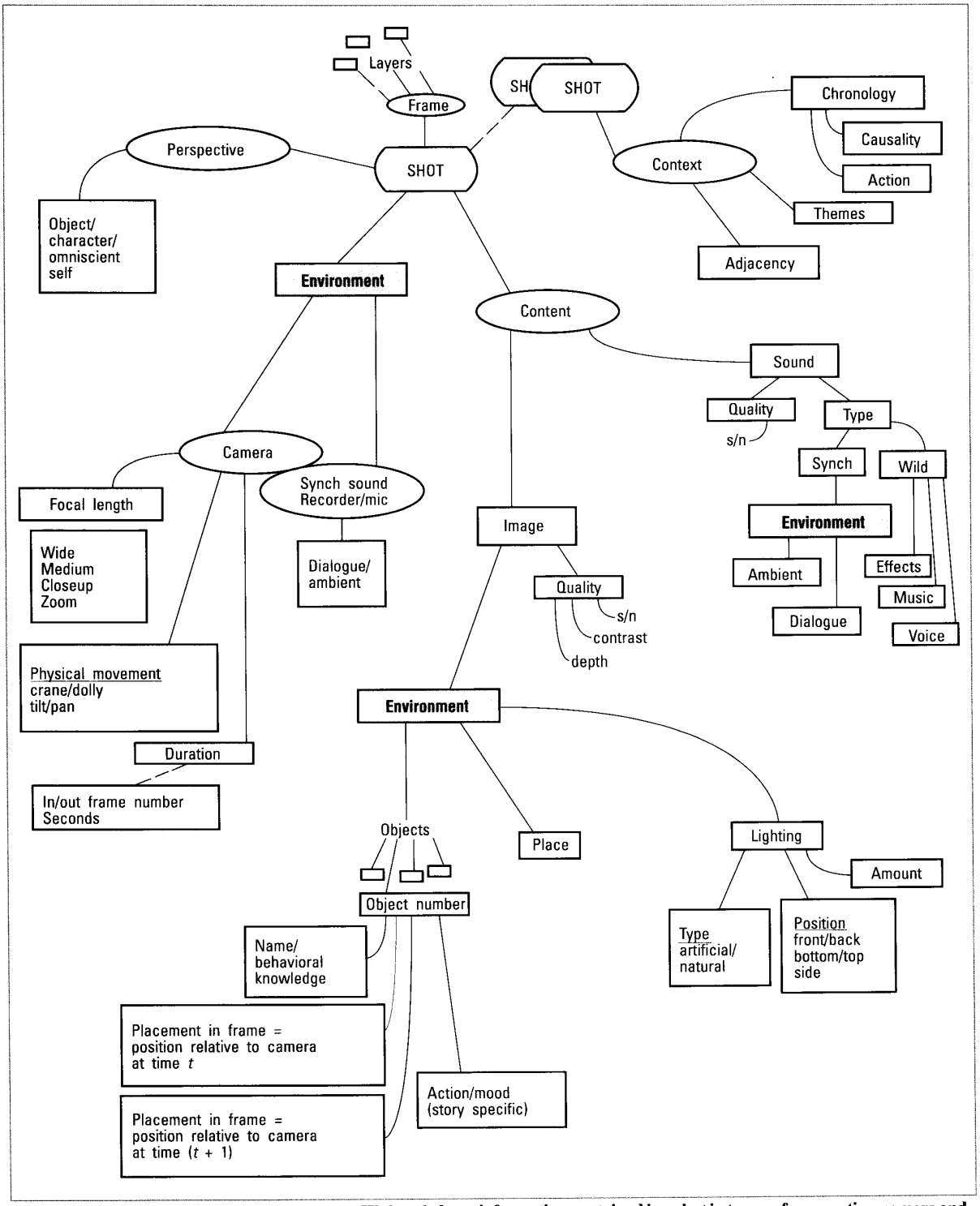

Figure 1. The shot: toward a structured model. We break down information contained in a shot in terms of perspective, camera and sound, content, and context. In each case our understanding of the relation of the recorded information to the environment helps us articulate meaning.

any picture can help us segment the video/audio stream into shots or help us determine significant action points. Notwithstanding, sound poses some difficult challenges to our representation methodology.

For instance, in interviews sound is usually the principal source of meaning. To represent the content of the interview, which is frequently captured as a single shot, the text of the interview needs to be broken down into conceptual units. The length of these units can vary enormously. While text transcripts of interviews and of all dialogue offer certain value to multimedia manipulations, neither speech recognition nor natural language understanding are sufficiently developed to help in constructing a representation from real-world dialogues. Inevitably, we will need to encode a summarized conceptual representation of dialogue tracks as part of our stratified representation of movie content.

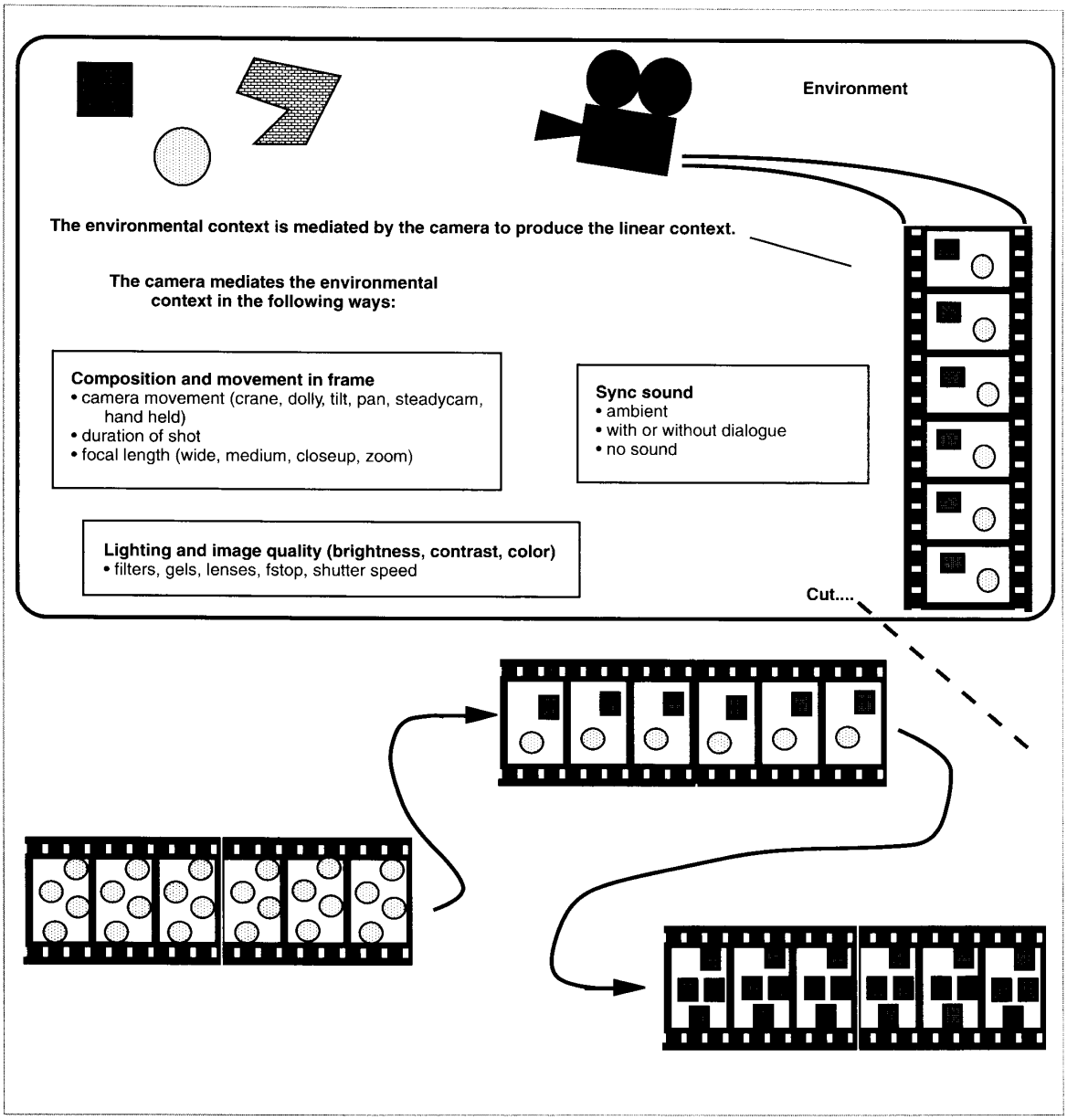

Figure 2. The situated camera. In editing, you can create context (chronology, causality, action) and develop themes by ordering image units into sequences; employing optical effects such as fades, dissolves, and wipes; using sound such as ambient, live, voice, and special effects; and adding text and break points.

The data camera

Some readers might already know that the cost of data entry for this stratified description will quickly become overwhelming if we rely only on after-the-fact human annotation. Because our representation method relies on the idea of the situated camera, we are currently considering how to build a camera to collect some of our descriptive information during the shooting process. We call this camera a *data camera*. Such a camera will record multiple tracks of data, including

- time code (SMPTE is already generated in professional electronic recording systems);
- camera position (by incorporating a global positioning system, as well as dynamic tracking of focal length and focus setting recorders); and
- voice annotation of who-what-why information.

Such data tracks can easily interface to a stratified logging system. Information collected during shooting is more reliable than after-the-fact annotation and will equally serve both modes of linear editing and interactive multimedia browsing applications.

The sequence

While the shot and associated audio are primitive elements, the sequence represents a second order of cinematic structure. Sequences are created by placing one shot next to another. While each shot describes a part of an action, the sequence allows us to understand a whole action, event, or thought. Whether comprised of one or more shots, a sequence is distinguished by the fact that it is no longer perceived as a set of individual shots. The transition between shots may be further blurred by the existence of a mix track. We believe that both the machine and the individual editor can use our descriptive methodology to build meaningful, context-rich sequences.

In traditional cinema, the editor tries to create a seamless experience. As just mentioned, sound can be used to blur distinctions at the edges of shots, sequences, and scenes. As a final step in cinematic postproduction, after the editor has fine-tuned or "locked picture," a sound editor is frequently employed to "clean up the track," embellish the feeling of being there by inserting sound effects, and create audio transitions. In addition to sync sound, a sound editor can add a track providing explanatory commentary or music to set a mood, to make information more explicit, or to heighten the perception of an action.

The utility of sound effects on reel can clearly be extended into multimedia systems. Sound libraries, such as the "Sound Ideas" CD collection, are currently being marketed. As we develop models for sequence construction, we need to be able to link to sound libraries in order to build seamless sequences.

One of the goals of "conversational" interactive multimedia systems is to create coherent linear sequences on the fly, according to user preferences. As sequences are built, the context inherent at the shot level is harnessed and made explicit. In both documentary and narrative movies, context is developed through manipulation of chronology, introduction of causality, or simply structuring a completed action. Interactive multimedia differs from traditional cinema in that participant viewers may need to choose between multiple paths, may need to review their course, and may wish to actively restructure a sequence. Nevertheless, both the viewers' experience and associated learning proceeds in time and can be enhanced by the generation of seamless narratives. This requires dynamic look-ahead and scalable descriptions.

Stratification is a model for annotation that approximates the way in which the editor builds an understanding of what happens in individual shots.

Using representative strata

While movie viewing is sequential, movie production is inherently nonlinear. To construct coherent sequences, the editor must understand what took place in the environment mediated by the camera. As part of this process the editor builds a cognitive map of the environment and keeps track of who, what, when, where, and why over the temporal duration of a shot.

Stratification is a model for annotation that approximates the way in which the editor builds an understanding of what happens in individual shots. Much of this information could be collected quite easily by a data camera at the time of shooting. Therefore, stratified annotation appears to have several advantages for multimedia environments. First, it can comfortably support granularity. Second, other programs can easily parse the descriptions. Third, the original annotation can be tracked through multiple representations.

In addition, stratification can help support computer-generated links between similar pieces of footage. Editors use linking to associate sequences when creating linear narratives. Stratified descriptions of content will therefore expedite on-the-fly creation of machine-generated, multithreaded parallel narratives that users can adjust through linking. Such narratives are vital for effective multimedia environments. ❑

Acknowledgments

Ideas in this article evolved over the past six years. They result both from building multimedia systems and from producing multimedia publications. While we have given a few directly relevant references, many other students and research affiliates of the Film/Video Group (recently renamed the Interactive Cinema Group) contributed to our exploration and understanding of cinematic primitives. A short list follows: Dan Applebaum, Hal Birkeland, Brian Bradley, Hans Peter Brøndmo, Amy Bruckman, Ben Rubin, and Martha Swetzoff. Rele-

vant research was conducted under grants from MIT's Project Athena and in collaboration with research affiliates from the Asahi Broadcasting Corporation, particularly Keishi Kandori, Kenzo Furukawa, Yoshio Obata, Toyokazu Yoshida, Kuniyoshi Chihara, and Hiroshi Ikeda. Thanks also to Ricky Leacock for ideas communicated in many conversations and lectures between 1978 and 1988.

References

1. S. Morgaine, *MIDS: A System for Describing Image Content for Multimedia Design*, MSVS thesis, MIT, 1989.
2. S.M. Stevens, "Intelligent Interactive Video Simulation of a Code Inspection," *CACM*, Vol. 32, No. 7, July 1989, pp. 832-843.
3. W. MacKay and G. Davenport, "Virtual Video Editing in Interactive Multimedia Applications," *CACM*, Vol. 32, No. 7, July 1989.
4. S.M. Eisenstein, *Film Form*, Harcourt Brace and Co., New York, 1949.
5. G. Toland, "How I Broke the Rules in Citizen Kane," in *Focus on Citizen Kane*, R. Gottesman, ed., Prentice-Hall, New Jersey, 1971, pp. 73-77.
6. K. Reisz and G. Millar, *The Technique of Film Editing*, Focal Press, Boston, 1978.
7. T.G. Aguierre Smith, "Stratification: Toward a Computer Representation of the Moving Image," a working paper from the Interactive Cinema Group, MIT Media Lab, 1991.

Glorianna Davenport is an assistant professor of media technology, first holder of the Asahi Broadcasting Corporation Career Development Chair, and Director of the Interactive Cinema Group, a research group at MIT's Media Lab. Formerly an independent documentary producer, Davenport has been involved in the development of computational multimedia since 1982. Her work includes interactive programs such as New Orleans Interactive and the Elastic Charles, as well as random access video viewing and editing tools.

Davenport received the Gyorgy Kepes Fellowship Prize for excellence in the arts at MIT in the fall of 1990. She holds an MA from Hunter College and a BA from Mt. Holyoke College.

Thomas G. Aguierre Smith is currently a masters degree candidate in the Interactive Cinema Group at the MIT Media Lab. His research interests include multiuser networked video databases, problems related to on-line video archiving, video indexing systems, and how video can be used as a research tool. His current research involves developing a computer representation of moving image content called "Stratification."

Smith received a BA degree with a concentration in anthropology from the University of California at Berkeley.

Natalio Pincever received his MS in media arts and sciences at the MIT Media Lab in 1991. His current research focuses on cinematic parsing of audio tracks for the Interactive Cinema Group. His research interests include digital audio processing, desktop audio, advanced human interfaces, multimedia communication, and hypermedia.

Pincever holds a BS degree in electrical engineering form the University of Florida. He is a member of the IEEE Computer Society.

The authors can be reached at the Interactive Cinema Group, E15-432, MIT Media Laboratory, 20 Ames Street, Cambridge, MA 02139.