Representing the Content of Video: Artifact or Process?

Scott L. Minneman[1] and **Stephen W. Smoliar**[2]

Abstract. An approach to semantics based on traditional paradigms of knowledge representation (e.g., developing reductionist models of video document genres), while not entirely off the mark, may be significantly misdirected. Understanding the semantics of video and multimedia must begin with understanding how video (and film) are "read" and "written." The purpose of this paper is to set an agenda for coming to an understanding of reading and writing multimedia and to address the representationalist implications of achieving that end. Further, we illustrate this research agenda, showing how we have applied concepts and methods from film theory (to the reading process) and interaction analysis (to the work of multimedia production), and what our preliminary findings might mean for computational support and knowledge representation in multimedia.

1 COMPUTATIONAL APPROACHES TO MEDIA

When dealing with documents, the most important problem that knowledge representation fails to confront is that the data structure nature of the document is actually secondary to the more fundamental nature of those *processes* which are required for that document to be "produced" and "consumed." (In the case of paper documents, these processes are more familiarly known as "writing" and "reading.") This is because the *social* implications of communication tend to be more important than the substantive nature of the message being communicated. Thus, if information technology is going to contribute to our understanding of how documents actually communicate, it will be not by refining structural models of the documents themselves but through modeling those processes of production and consumption. Such an inquiry is not well served by the intellectual foundations of computer science but must, instead, turn to principles laid down by disciplines such as semiotics and hermeneutics [1].

Semiotics and hermeneutics have traditionally been applied to text; and it still tends to be taken for granted that what we call "documents" generally consist of text printed on paper. However, this assumption changes as multimedia technology expands the scope of what can be put into a document. Unfortunately, when we see how information technology has tried to appropriate other media, such as video, we see the same faulty assumptions which have impeded our understanding of how text communicates. Current multimedia technology tries to approach video strictly in terms of its *structure*: how it is composed of *shots* and *transitions* between those shots [10]. Unfortunately, "parsing" a video or film into those structural primitives has little to do with how that video communicates, just as diagramming a sentence fails to tell us very much about how we actually *read* that sentence.

The next Section is an attempt to scope out a broader view of the complexity of the nature of "consuming" (i.e., "reading") video and film. It will quickly become apparent that the conventional paradigm of encoding a message and using knowledge representation to model the content of that message has only the slightest to do with the entire "big picture." Thus, in Section 3 we turn our attention instead to those processes of production and consumption which are far closer than the document itself to the nature of communication. Finally, we conclude by discussing how knowledge representation may provide support for those processes, even if the more subtle issues of message content continue to be elusive.

2 WHAT NEEDS TO BE REPRESENTED IN MOVING IMAGE MEDIA?

The major problem we face is that very few of our intuitions about the content of most video source material are particularly consistent with our intuitions about knowledge representation. Thus, while there are a variety of different ways in which the natural language content of, say, a newspaper story may be represented in some logical calculus and an equal variety of issues which have been explored regarding how that calculus may be used most effectively [13], there is far more to video than a textual narration which plays a role similar to the natural language found in newspapers. Most of the topics discussed in Dudley Andrew's *Concepts in Film Theory* [1] basically address how the content of video and film extend beyond what can be modeled by the state of the art of natural language processing. Andrew's enumeration is probably far from exhaustive, but it still constitutes a set of germane topics and pivotal questions which focus attention on current shortcomings of knowledge representation. Therefore, these topics will now be briefly reviewed.

2.1 Perception

On the surface, Umberto Eco's semiotic approach to the perception of film is similar to that of computer vision [1]:

[1] Xerox Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, California 94304, USA
[2] FX Palo Alto Laboratory, Inc., 3400 Hillview Avenue, Bldg. 4, Palo Alto, California 94304, USA

ECAI 96 Workshop. *KRIMS 96*
Edited by G.Vouros, W.Klas, T.Sellis, M.Vazirgiannis

> ...Eco goes ...to the pre-objective blotches of light, dark, and color which are the atoms of every image. Like the elementary particles of other semiotic systems (spoken language and music) these blotches are articulated via position and opposition to form fragments of recognizable semantic forms (triangles, vectors, and so forth) which are themselves articulated into iconic forms such as arms, legs, and trees.

However, in order to be computationally feasible, computer vision must assume an *a priori* model of what Eco calls semantic forms and iconic forms [7]. That model is the fundamental knowledge base which drives all vision processing. However, because semiotics assumes that such a model must incorporate not only the physics of the world but also the cultural context within which individuals interpret what they see in that world, it denies that such a model can be given *a priori* [1]. Rather, the model is *constructed* as part of the perceptual process or, as was much earlier suggested by David Hume [11], by a tight coupling between perception and cognition.

An approach based on current knowledge representation techniques thus falters because one simply cannot begin with a knowledge base of "iconic forms such as arms, legs, and trees." Every individual possesses a repertoire of such iconic forms, but they are a *personal repertoire* of *perceptual categories*. That repertoire is constructed as a result of perceptual experiences enhanced by values which are assumed through emotional associations with those experiences [7]. Those categories may be subsequently *labeled* as "arms, legs, and trees." However, that labeling process involves *linguistic*, as well as perceptual, experience; and linguistic experience may be seen as another form of perceptual categorization applied to the motor behavior of speech and the auditory behavior of listening [8].

2.2 Representation

In film theory representation involves the relation of the entire ensemble of the stimuli of the viewing experience (those "pre-objective blotches of light, dark, and color") "to the world it produces through our imagination" [1]. This is because, as Andrew puts it, "No matter what appears on the screen, audiences will instinctively shape it into a representation of something familiar to them" [1]. Representation is thus a mapping of the semantic and iconic forms of the film experience to the semantic and iconic forms of other experiences, perceived in either the real world or the imagination.

In the context of knowledge representation, such a mapping may be viewed as an analogy [2]. However, the knowledge representation approach to analogy assumes that both the domain and the range of the mapping are rigid structures. As we have seen, such rigidity is not compatible with semantic and iconic forms, whether they are part of our perceptual experience or part of the imagination to which perceived forms are mapped. Like the real world, the imagined world is also based on perceptual categories [8]; so the mapping of representation must have spaces of perceptual categories for both its domain and its range.

2.3 Signification

Of course the mapping of representation can no more be assumed to be given *a priori* than can the semantic and iconic forms which are being mapped. Rather, this mapping arises from an *interpretative* process engaged by the "reader" of the film or video experience. That process reflects an *active role* on the part of the reader (that "process of consumption") through which the stimuli acquire not only associations with the range of the representation mapping but also *significations* which embody how those stimuli are engaged by "the role of the reader" [6].

Signification is the primary focus of semiotics as it is currently pursued [5]. However, semiotics has been, for the most part, concerned with *static* constructs, such as text. Because interpretation is a *process*, the role of time becomes more important when the medium is *not* static. The interpretation of film and video must be performed under the constraint that interpreting occupies time, just as the "playing out" of the film or video occupies time. Thus, the producer of a film exercises a certain degree of control over how the consumer experiences time; but the consumer's experience must incorporate interpretation, rather than simply sensation or perception. Once again, these are relationships which clearly strain the current expressive capabilities of knowledge representation.

2.4 Narrative Structure

What is it that guides the consumer through a process as complex as interpretation? Usually, it is the fact that what is being consumed is a *story*; and every consumer comes to a story with certain assumptions about how stories are structured and told. This is as true if the story is told by a film as it is when the story is told by text, either printed or delivered by a narrator. In both cases, however, the ways in which stories are *structured* have a lot to do with how they are understood. Citing the hermeneutic studies of Paul Ricoeur, Andrew observed that "every text is comprehensible only because of a system (grammar) that gives us access to it and inevitably limits what the text can say" [1]. Because so much work has gone into characterizing such systems, it should be no surprise that some of those results should have found their way into approaches to knowledge representation [3]. Unfortunately, these approaches tend to view a story as being structured for its own sake. The role of the reader is trivialized to the task of either querying the knowledge base which represents a story or simply reproducing it. The interpretative process which lies at the heart of signification remains ignored. However, unless we understand how a system of structuring relates to that process of interpretation, it is hard to establish the validity of the structure itself.

2.5 Adaptation

Another issue which is important in understanding the content of a film is that film is often used to *adapt* material which had been presented through some other medium. We all know how film and video have been used to present their own versions of "classic" or familiar novels. However, the adaptational role of film may be generalized beyond its most obvious applications [1]:

> Every representational film *adapts* a prior conception. Indeed the very term "representation" suggests the existence of a model. Adaptation delimits representation by insisting on the cultural status of the model, on its existence in the mode of the text or the already textualized. In the case of those texts explicitly termed "adaptations," the cultural model which the cinema represents is already treasured as a representation in another sign system.

From this point of view Andrew [1] argues that there are three "modes of relation" between a film and the source material being adapted: borrowing, intersection, and fidelity of transformation. These modes will now be briefly reviewed.

In the case of borrowing, "the artist employs, more or less extensively, the material, idea, or form of an earlier, generally successful text. ... Here the main concern is the generality of the original, its potential for wide and varied appeal; in short, its existence as a continuing form or archetype in culture." The opposite of borrowing is intersection, where "the uniqueness of the original text is preserved to such an extent that it is intentionally left unassimilated in adaptation. The cinema, as a separate mechanism, records its confrontation with an ultimately intransigent text" [1]. Such intersection arises, for example, when characters in a contemporary setting speak to each other in Shakespeare's language, as opposed to *West Side Story*, which borrows the idea of *Romeo and Juliet* but does not bring along any of Shakespeare's text. In both these cases, however, there will be the issue of whether or not the adaptation is being *faithful* to the source, be it borrowed or intersected. Unfortunately, the assessment of fidelity often rests upon the interpretation of the reader, since it is that interpretative process which determines whether or not the elements of the film experience "fit." Thus, watching Richard III implement his machinations from behind a computer workstation may "work" in one adaptation and fall completely apart in another; but the efficacity of the technique may have more to do with what the consumer is relating to when confronted with the image than with whether that image of Richard with contemporary office equipment is a "faithful" one.

From a point of view of knowledge representation, the problem here is that adaptation involves more than the content of a single film. It involves the content of some source as well, along with the relationship between the two presentations. Needless to say, if that content cannot be adequately modeled for a single presentation, relating two such presentations will be even more infeasible.

2.6 Valuation

Thus far we have examined the film experience as if it were entirely perceptual and cognitive. However, it is clear that the film experience is also emotional. Furthermore, not only is it the case that emotional response tends to lie at the heart of whether or not we decide to have a film experience at all (or remember that experience and relate it to others); but also recent studies of the brain have revealed that cognitive processing is more tightly coupled to emotional processing than had been previously supposed [4]. Thus, it is probably fair to say that a consumer is not necessarily fully capable of perception, representation, or interpretation without also being able to attach emotional value to what is being perceived, represented, and interpreted [7].

This is clearly problematic, since the technology of knowledge representation is predicated upon the Cartesian assumption that any description of a part of the world can be separated from our emotional response to that particular part. Returning to the terminology of Edelman [7], we have perceptual categories not only because we are "wired" to have them but also because we are wired to *want* to have them. Perceptual experience is far too rich to allow the processing of every conceivable interpretation in terms of semantic and iconic forms. Such processing has to be regulated; and regulation is the overall objective of emotional processing [4], be it the regulation of autonomous bodily functions or the regulation of the very way in which we establish how the world is made up of objects and events.

2.7 Identification

However, not only is the psychology of emotion critical to the perception and interpretation of film content; but also the entire psychology of personality is part of the picture [1]:

> Questions about the connections cinema maintains with reality and with art ... were subsumed under the consideration of cinema's rapport with the spectator. A new faculty, the unconscious, instantly became a necessary part of any overarching film theory, and a new discourse, psychoanalysis, was called upon to explain what before had been of little consequence, the fact and the force of desire.

Content is thus grounded not only on semantic and iconic forms, not only on emotions which regulate how those forms come to be perceived and interpreted, but also on the very *drives* upon which every personality is constructed. Furthermore, those drives rest as much on the narrative content of a film as they do on its image and audio content [1]:

> All stories proceed by a flow that aims to fill a lack. The storytelling ritual is a universal phenomenon because of the constitutional emptiness in experience, what before psychoanalysis was called "the human condition." Stories satisfy our need to sense the filling of a lack and the achievement of stasis. How that lack is identified (as a maiden, a father, a treasure, an integrated view of the past, or whatever) is of less moment here than how it is managed in narrative. For stories, while seeking a timeless goal ("happily ever after"), are in fact defined by the opposite of stasis, by flow, change, and interaction. In this context we might add incidentally that the flow of film in its basic perceptual mode necessarily prepares the way for its narrative dimension, explaining perhaps the primacy of narrative over other forms of film.

Once again we see the essential tension between the dynamic nature of film and video content and the static nature of the knowledge representation technology.

2.8 Figuration

In his penultimate chapter Andrew recognizes that the content of any film or video is not just semantic but also *rhetorical*

[1]. Semiotic theory has thus been obliged to progress from the semantic analysis of words, phrases, and sentences, through the structural analysis of narrative discourse, and through the psychological analysis of emotion until it arrives at the need for a rhetorical analysis of presentation by which the producer may invoke powers of persuasion upon the consumer [1]:

> The more serious student will see in this shift the recognition ... that film is ordered not as a natural language but at best as a set of practices and strategies that are in some way "ready-to-hand" but hardly form a system in any strong sense of the term. This aspect of *bricolage* at the heart of the medium suggests that meaning in film comes largely by way of conventions which began as [rhetorical] figures [of speech]. A dissolve denotes the passage of time today only because for years it figured that passage palpably through the physical intertwining of adjacent but distinct scenes.

Here, ironically, we may see a path by which knowledge representation may enter the picture. If we think of film not strictly in terms of its content but in terms of that "set of practices and strategies" which give rise to a "content-bearing artifact," then we may be better off searching for knowledge-based tools to support those practices and strategies. The paradigms of work, both individual and cooperative, may be more conducive to technological support than the paradigms of the objects being made by the work.

2.9 Other Factors

Andrew's analysis, which we've summarized in the previous sections, can be faulted for its concentration on issues that are fundamentally individual, even psychoanalytic, while it is clear that film and video come from and are viewed within a complex social sphere whose influences cannot responsibly be neglected. These social factors further erode the plausibility of traditional knowledge representation making much headway in the understanding of these media. Thus, before we pursue film-making from the perspective of work paradigms (and the support of those paradigms through knowledge representation), we would like to discuss briefly three key social factors.

2.9.1 Genre

Andrew's discussion of genre is subsumed under his approach to valuation [1], but genre is important in its own right as that social factor which determines how sets of narrative conventions (ranging from plot through characterization to cutting style) become recognizable types of moving image material (e.g., documentary, romance, pornography) [17]. Genres are used in both production and consumption; multimedia authors use genre to acknowledge the extent to which the process of interpretation must play in the consumption of the products of their efforts.[3] The consumers of film, too, are expected to bring a level of sophistication to their viewing; so the author need not spell out what should be obvious (e.g., that narration is a guide to what's being seen, not the voice of somebody who's just outside the framed view). There is a complex interplay between the producers, the consumers, and the editors of multimedia content that allows for genre to operate, while supporting the fluid development of new genres.

The social contract of genre interpretation is particularly important in settings involving interactive multimedia where it isn't necessarily the case that a complete linear narrative will be played out. Individuals, always largely responsible for personal sensemaking in the consumption of moving media productions, become further burdened with a process of integrating the work into a cohesive whole, assuming that is something they desire, in which case there's often an obvious path through a multimedia work that will provide them with what they seek. Genres are also both culturally and historically embedded: productions from other times feel very different from those of today (see 1950s lifestyle training films for a good example), work from other cultures can be very obtuse.[4] The question remains, though, of how one can play to genres that vary among one's audience members: What does the local stories section of a TV news broadcast in China look like?

Genre recognition, then, is a slippery slope. Human viewers simultaneously see the pattern and the variations; the variations are not simply noise, but also the past and/or the future and/or comedic relief and/or interjections. Granted, many genres are very formulaic and historically stable (e.g., the daytime soap opera), to the point that the formulas may even be captured in a knowledge base. However, while that knowledge base may some day allow a system to provide a weekly summary (and make its designer very wealthy), it will not distinguish a soap opera from a parody of a soap opera.

2.9.2 Audience

The extent to which it's possible to draw on the resources of peers and collective consciousness in the production and consumption of moving image material is a very challenging dimension of understanding what we're up to here. Unfortunately, this aspect is largely unknown during the early production stages of most pieces—that a film may be a hit is seldom available as a resource to draw upon (although sequels can play with the social conventions that may have sprung up around a particular work). Multimedia productions for more limited audiences have more latitude in this area; it is possible, for example, in the human interface community, to parody the Apple Knowledge Navigator interface, counting on the community to fill in unfamiliar members with the resources needed for successful interpretation.

Computers don't play well in this arena, not being full-fledged members of the filmic audience (although, in popular media, textual analysis of film criticism could help). Furthermore, as the foundation for hypermedia technology, computers have introduced a new factor in reading and writing—random access—which significantly alters the extent to which the common audience experience can be drawn upon as a resource. Thus, the producers of *Myst* have no control over whether or not every consumer ever gets off the island, making it virtually impossible for them to draw upon factors such

[3] Even the producer who chooses to break genre bounds must do so in such a way as to make that break from convention be comprehensible to the audience(s).

[4] However, the globalization of many media services is heralding the emergence of a (Western) media Procrustean bed.

as sequence or context in establishing a relationship with their audience.

2.9.3 Stardom

Not only do the viewers serve as resources in filmic material, so do the subjects. Producers and consumers alike draw upon, play with, contrast, compare, and parody previous roles that they've played or previous times that particular source material has been produced. Someone may become typecast, always playing the villain or playing one scene in a style that recalls a previous role. Stardom, along with genre, factors heavily in consumers' choices in what to view. This pattern perpetuates itself by exposure; the marketing strategy is obvious—we are shown promotional material that reflects the films or television programs we are currently making time to see. A trailer for a touching romance will likely attract few new patrons if it's shown to a theatre full of testosterone-pumped action-adventure fans.

3 IMPACTS OF THESE THEORETIC FACTORS

Much of the work of film theory has concentrated on how these theoretic elements show up in the constructed artifacts (e.g., what constitutes the genre "western" over the decades). The most salient point for the knowledge representation community is the depth and breadth of the gulf between what can be computationally represented of (and automatically derived from) filmic material and what we, as viewers, immediately bring to bear in comprehending a work of moving image media. We get the gestalt from the get-go, not by coming up from primitives. In an attempt to get a better appreciation of how these concerns show up across the wide range of settings involved in multimedia production and consumption, we've turned our sights to another dimension: how these theoretic factors arise in the authoring process.

To get a handle on these matters, we have employed a variety of observational methods to examine the day-to-day work practice of multimedia producers in a varied corporate setting. The bulk of our observations are drawn from open-ended interviews and video analysis, drawn from the tradition of work-practice studies [16]. While this work must be considered very preliminary, many interesting points are coming to light.

The editor/producer we concentrated most of our efforts on, hereafter referred to as Chris, is employed in the "Creative Services" group at a medium-sized industrial research facility. One member of a small, self-managed team, she is typically involved in a number of simultaneous projects ranging from day-to-day presentation support in the facility's auditorium to CD authoring/mastering to editing material for a researcher's conference presentation to authoring Web pages for other groups within the corporation. Chris was approached as the subject of this initial study, in part, because of this wide-ranging set of responsibilities and skills; as is typical and desired when using interaction analysis, we didn't have pre-formulated notions of exactly what we were looking for or where we might find material relevant to our broad topic.

3.1 Open-Ended Interview

We began this exercise by preparing a list of questions which we wanted to raise in an interview with Chris. We encouraged her to answer these in specific terms, i.e. with reference to one or two particular projects, ongoing or recent. Many of the questions were designed with the topics reviewed in Section 2 in mind:

Perception: What are the objects (artifacts and/or events) depicted [in that project]?

Representation: What points are you trying to make? How do you get your clients to work within a viable filmic structure?

Signification: How do you work with getting viewers to interpret your productions in the way that you or your clients desire?

Narrative Structure: How much effort does it take to get a workable structure [in these productions]? What is the tie-in between that structure and the points you are making? Do things cluster in any useful way?

Adaptation: How often is the work you do a translation of a document from another medium? How does such translation occur?

Valuation: What were the successful things you did in editing [a particular video]? What will the viewer(s) think was successful? How radically do things change during actual editing? What serves as a catalyst for these changes?

Identification: How do you get into the perspective of an "ordinary" viewer? How does this influence what you do? What personalities are evident in the videos you produce? Do you expect viewers to identify with people or settings in the work?

Because this interview was open-ended, we did not explicitly cover all of the questions that we had prepared in advance, nor did we try, in any systematic way, to touch upon all of the topics addressed in Section 2. Furthermore, we found that the conversation drifted into other areas, including the topics discussed in Section 2.9. Nevertheless, the interview highlighted some of the severe mismatches between artifact-oriented understandings of moving-image media and Chris' everyday work. For her, multimedia production took place in a complex organizational realm populated by clients, subjects, budgets, scripts, schedules, storyboards, equipment, and myriad other influential factors. The shape of particular pieces arose as much from a social process of coming to grips with the dimensions of the possible as from an artistic vision.

> There's many, many times when I'm sitting in the edit suite with somebody and I'm saying, well, if I cut it this way it implies this, and if I cut this way it implies this, which is closer to the truth? ...and then, OK, do you *want* to imply the truth or not?

What we see here is a "ground-level" confrontation with the difficulties which arise when the theory of signification has to be put into practice. Furthermore, that practice must deal with the *process* nature of signification, specifically raised in Section 2.3, and how the dynamics of that process is explicitly related to the dynamics of the film (which arises from a particular choice of editing cuts). However, because those cuts often become key delimiters of *events* in the video, they

also serve to define that narrative structure which guides an individual viewer's process of signification.

Chris' pragmatic approach should not suggest that a deep knowledge of the points we've introduced from film theory don't apply to everyday production settings; rather it suggests how they must be considered in a broader-ranging social context:

> The preliminary stage was, he came to us and said "I don't want to write a white paper, so I want to make a videotape of it that will do that," so we knew that it had to be—have enough technical detail to suffice as a tech report but still be watchable by those of a somewhat more generic audience.

This is a case where adaptation is the issue of highest priority; but what does it mean to adapt the idea of a white paper into a video? Clearly, this is a case where fidelity of transformation must be honored above all else; but, within that context, what is the most appropriate approach to adaptation. Is borrowing more appropriate, or is intersection preferable?

This matter is further complicated by issues of genre. The "white paper document" is a well-defined genre in most research contexts; and it is typefied by dense technical content. Video tends to be most powerful in delivering a point of view, rather than exhaustively accounting for all relevant factual issues. Consequently, the white paper's density of facts and apparent objectivity is unlikely to be effectively communicated in the video domain; there *is* no direct mapping (borrowing or intersection) of the white paper genre from text to video. As we shall see, Chris often had to deal with her clients wanting to communicate large amounts of technical content, leading to the apparent paradox that adding content does not always increase the power of communication.

In these settings, Chris must bring to bear her knowledge of viable filmic strategies to the communicative goals of her customers. These parties sometimes have only the vaguest notions of what it is that they're seeking (e.g., "something attention-grabbing for our trade show booth"). Often, her customers fail to recognize the particular strengths of the visual media:

> On [a recent tape] we convinced them that their usual wide shot wasn't right; we inserted a simple close-up where things happened ... it made everything move faster, made it click better, it seemed to improve the amount of ... connection made between what you saw on the screen and what was being said on the audio track.

And the weaker aspects and pitfalls of the media:

> The biggest trade-offs that we're always making is content—the amount of content that people want to put in—a point that people frequently miss is: more content does *not* equal more communication. If you put too much content in, your viewer will probably absorb less of it than if you had put the right amount in.

Both of these examples boil down to questions of valuation, as well as the underlying personality issues of identification. At the same time making things "click" is ultimately a matter of rhetoric: They probably "click" because particular figures[5] have been appropriately engaged. In other words whether or not a given video communicates successfully ultimately involves a tight interaction among considerations of valuation, identification, and figuration.

3.2 Observation of Practice

We also observed Chris in the edit suite and watched how the elements we've been discussing arose and didn't, and were supported and weren't, in her moment-to-moment work. The edit suite at the research center is reasonably well-equipped, with a state-of-the-art digital non-linear editing system. These systems have limited on-line capacity, so one aspect of working with them is managing segments and parallel projects so that the necessary source material is available. During one session when we were observing Chris at work, she was digitizing footage from a previous project shoot back onto the editing system's disk banks for editing into another piece. Although portions of this material had been used in a previous production (and thus would have been available on the backup tapes used to move projects off of spinning disks), the helical-scan backup drives are very slow and she felt it would be more efficient simply to go back to the source tapes and digitize it again. This practice, however, raised a concern:

> I'm going through here *really* quickly and setting ins and outs assuming that the last take of each of these shots is the best one, but that's not always the case—we're usually working out problems during the first few, but any one of the last ones could be the best take.

Also, going back to footage that was anything other than freshly shot had other potential problems:

> There's some concern with reusing this stuff that I'll have to check on later. Part of what this piece is trying to show is that [the customer] has state-of-the-art manufacturing capabilities, but if [we] accidentally show some piece of outmoded equipment to a really knowledgeable viewer, it'll undermine that message. This is really aimed at a less technically astute viewer, but [the customer] is worried about it, nonetheless.

This particular exercise revealed that the making of a technical video document can introduce issues of adaptation which are usually too subtle for most commercial film-making and are concerned more with the dispositions of the audience than with the mechanics of adaptation. Chris' concern here is that, while the video may achieve its desired effect for its intended audience, particular members of that audience may read too much into the presentation and take away an entirely different, and possibly damaging, message. The lesson here is that borrowing is often very valuable in trying to get across a point, providing one borrows from just the right source!

The technologies of digital video production, currently found on the cutting edge of computational systems, seem to perpetually lack the polish and integration of more mature technologies. It may well be that the challenges of using these systems and the accomplishment of having collectively

[5] In Andrew's discussion of figuration, he goes to considerable length to tease out how the figures of speech which enhance oratory can be generalized to similar "figures" in the film domain. Correlating a speaker's voice with an image of what he is describing is an example of such a "figure of film."

bent them to the will of a diverse team is one of the important take-aways in a production effort. The articulation of the possibilities with examples is a clearly recognized strength of established production groups: Having an impressive "demo reel" and the ability to produce quickly something evocative to give the client a sense of how the final product will appear are powerful assets in attracting clients. On the other hand, one of the greatest liabilities is often the absence of any assistance in managing an excessive number of resources which may be required for a particular project. Chris' system, for instance, lacked many basic niceties that would have helped with segment management, confounding, for instance, issues of mnemonic naming and uniqueness.

> I tend to give each project a number that we use at the start of all of the filenames so that they'll bunch together when I do a listing. I know, it's really ad hoc, but it works well enough to keep us out of major trouble—the source tapes, on the other hand, are stored next door in a roughly historical order . . . if you're looking at a finished piece and want to get back all the way back to the source tape, the path is really convoluted.

Is there any value in support at the "parsing" level? From the perspective of multimedia technology, the greatest success in parsing has come at the level of segmenting source material [10]. Segmentation is clearly a concern during production; but, with many kinds of footage (e.g., a talking head with illustrative b-roll), setting viable in and out points (i.e., *semantic* segment boundaries) and working on continuity between shots, again at a semantic level, are more important issues.

> I remember working with this stuff—remember how in the best wide shot the [objects] were moving left to right? Well, when we wanted to cut in for a close-up, all we had were these shots going right to left, too jarring, so we took the right to left stuff and reversed it to synthesize footage going the right way.

Was this a problem of poor planning? No, this is just the reality of the naturalistic origins of the source material, the cost and time of shooting, of not knowing exactly which shots would be used where. Chris indicated that they don't always know the structure of a piece as they're going in, that producing a piece is often a process of exploration. The raw video is an artifact, a souvenir of that process, that structures the telling of the story that's been uncovered. Thus, before that raw video can be put to use, it must be analyzed at the lowest level of perception for what objects it presents and how it presents them. The film-maker must then move on to the problem of representation, making sure that those objects are mapped to the proper experiences, even if such a mapping can only be achived if the source objects are first transformed in a way that will reinforce the validity of the experience.

Even in cases where a piece has been carefully scripted, the production remains a process of improvization among the various players. The customer is getting clearer about what is necessary; the editor is learning about which portion of that message can be conveyed with video, which must be narrated, which are animations, and which might best be communicated by accompanying printed material; the quality of the shot footage constrains what can be shown or provides irresistable opportunities for particular segues; trial audiences might be getting the wrong message, or bored, or insulted.

> In [a recent video], one particular talking head just wasn't cutting it, so it got swapped out at the last minute for an animation . . . the same basic information came though, but the question was exactly how to communicate it.

This is precisely what is meant by *bricolage* in the quotation cited in Section 2.8. Different figures can serve the same communicative function. Often, it is just a matter of resolving which figure works best ("clicks").

It is important to conclude this section by reiterating that these observations must be read as preliminary. We are only beginning to scratch the surface of what could be learned about the social and technological setting of multimedia production. These issues will have to be investigated in greater depth before truly useful niches for information technology can be fully characterized.

4 AN ALTERNATIVE ROLE FOR KNOWLEDGE REPRESENTATION: PROCESS SUPPORT

In Section 2 we tried to demonstrate that traditional knowledge representation was simply not up to the task of dealing with many key issues of content which had been raised in film theory. In Section 3 we then showed how the practice of making videos often involves juggling several of these issues at once, even at times when they might appear self-contradictory. All this could easily lead one to believe that the entire knowledge representation arena is doomed and fundamentally flawed: Many researchers were following in the academic footsteps of computational linguistics, mistaking the formalisms they had imposed on the domain for the domain itself, and declaring successes germane only in these artificial worlds. As we spent more time, however, threads of potential emerged from our observational work with practitioners.

4.1 Segment Management

As was observed in Section 3, keeping track of an unmanageable quantity of resource units is one of the most difficult parts of the production process. A variety of approaches have been taken to the indexing and retrieval of such units within the paradigms of multimedia databases [9]; but it is unclear that a video editor having to work in a real-time "seat of the pants" mode is going to be able to afford the time either to set up such a database or to hack away at figuring out the right way to query it. We believe that one plausible approach to what is required, instead, is a sufficiently flexible technique for outlining one's various plans and concerns before (and during) jumping into the task. However, it is important to recognize that such outlining has to take place at several different levels (called "activity spaces" in the SEPIA system [15]). Thus, one will have to organize one's materials in different (probably simultaneous) ways at different stages in the production effort; and such a bookkeeping task may be better managed by a machine than by any human user (particularly one concerned with more pressing problems of creation).

4.2 Activity Capture

Activity Capture refers to a suite of technologies developed at Xerox PARC that might be applied to the problems highlighted in Section 3 in a variety of ways [14]. The tools provide the means to initiate digital multimedia recordings, a variety of ways to index those recordings, and ways to retrieve the indexed material in other settings. These technologies might be applied directly, during the capture of source material, and/or at a meta level, to document the various processes of producing a multimedia piece.

In the direct application the tools would be employed during the gathering of source materials in order to find particular material at a later time. Automatic and manual means could be employed to gather index marks for shot boundaries, annotations could be made on particular segments, and ancillary information could be gathered about a collection of shots that would aid in later retrieval. The primitive application programmer interfaces of many of the devices and packages involved might make integrating them into our architecture a challenge, but the downstream pay-offs would likely be significant.

Another level where Activity Capture technologies could provide valuable help is in giving the various players in the social mileu of multimedia production a means to access recordings of their own process. We have noted that, as work progresses on these projects, a deeper appreciation of others' concerns is acquired; days after a meeting with a customer, a producer might have the background needed to understand the finer points of what was discussed. Having an indexed recording of that meeting might provide a way to streamline this process of coming to a shared understanding of a project. This is not a panacea, however, as we've also noted the extent to which these goals naturally shift and how these processes depend on contact with the other parties in the effort.

4.3 Automatic Segmentation

While we've been quite harsh towards any hopes for going from low-level (shots and transitions) representations of moving media material to an understanding of content, we see considerable value in tools that would provide this and other breakdowns of multimedia material to producers and editors. Our goal for such techniques is not that of using these "primitives" to provide higher level interpretations of the material but to give editors better tools to relieve them of some of the drudgery of their work.

Video analysis is somewhat different from most computational vision research in that many of the techniques do not seek to model the world, but aim only to describe the recorded material. While challenging, this is a more modest goal than full-blown computer vision; and major strides have been made towards it. The simplest are probably those that do cut detection based on a search for significant frame differences. Other transition types (wipes, fades) can be more challenging; but successful algorithms have been demonstrated. Various systems for scene analysis have been developed [9] that do such crude, but useful, things as cutting the frame into rectangular segments and perform dominant color matching to identify image classes. Object tracking, wherein the trajectories and transformations of, for instance, collections of edge-detected pixels are detected and identified, is probably the most ambitious work of this sort [12]; but additional work must be carried out to establish its relevance to actual work practices.

Speaker Identification takes another tack at moving image stream segmentation [18]. Hidden Markov Models are trained up for each of the speakers in a scene, and those models are used to identify who's speaking when. (Note that this is *not* speech recognition; no attempt is made to determine what is being uttered.) This technique is particularly effective in long shots with distinguishable speakers (or characteristic noises). This might well be very effective for tracking down memorable moments in a shoot (e.g., the good shot came right after Chris gave them lighting direction).

While each of these methods, properly applied, is promising in its own right, real excitement should come from their combination into a suite of tools that allow production teams to focus on less mundane issues.

5 DISCUSSION

While the categories of analysis from film theory give us a leg up on understanding important aspects of multimedia, qualities of these media also invalidate some of the central assumptions operating in the filmic arena. The role of the consumer, random access, and exploration fundamentally change what can be counted on during (and in the wake of) seeing a multimedia piece.

5.1 Codification

Genres have been slow to develop in new multimedia products. Many factors play into this observation, but one that we believe is pivotal is that the consumer of multimedia titles is an unpredictable beast. Unlike the moviegoer, whose attention is fairly undivided and whose faculties will all be brought to bear upon the big screen for 112 minutes, a CD buyer, in marked contrast, cannot be counted on for a sustained interaction (the producer needs a hook), intense concentration (they may be doing other things at the same time), or controlled facilities (watching something on a tiny screen with tinny speakers isn't very absorbing). On the other hand, these media can be liberating, if used well; our editor, Chris, again:

> I can use all the very best material because I don't *have* to make it flow from A to B in a seamless flow ... that part of it was just this huge light bulb going "Wow, this is great."

The development of new channels for distribution of these media will influence the development of new genres, but producers and consumers have the opportunity to participate in this process. Inclusion of aspects like a passive mode to generate interest (e.g., show me bits of this CD[6] while I'm exercising) as well as support for recognizable ways for full engagement might be of great service to producers and consumers alike.

[6] These might be carefully scripted paths through the work.

5.2 Technological Changes in Editing Suites

What changes are being wrought in these disciplines, and how do these play into and/or drive the observations of this paper? In Section 3.2, we highlighted a moment where unavailable footage was *derived* from existing materials. Artificial pans, zooms, matte inserts, slow-mos, and stills can be created from raw footage of various kinds. This changes the role of segmentation steps in production and requires even more intimate visual familiarity with qualities of each shot (e.g., people can only sometimes be left-right reversed, writing almost never, backwards human motion is very distinguishable); it may trickle down into new kinds of shooting, as well.

6 CONCLUSIONS

We have observed a big mismatch between the realities of multimedia production and the heroic image of the lone editor toiling to bring an artistic vision into fruition. Seldom is our editor frustrated by the need to find a generic reddish sunrise over water with boats (and we assert that one will seldom find an individual with on such a quest). The issues that arise in real practice revolve around having facile resources for organizing and keeping track of the moment-to-moment concerns of real work. (Which of these clips have I watched? Which are still candidates for b-roll in this unfinished section? How much time can I responsibly spend sorting out this awkward transition?)

Also contrary to popular belief, the multimedia artifact is *not* the primary product. The real "game" here is *communication*. This begins with the set of understandings that get forged among the interested parties in the production effort. The "game" then proceeds downstream to the viewing process; but that is a story for another paper, where we must explore the extent to which the overall process of production actually reveals itself to the viewer.

ACKNOWLEDGEMENTS

We would like to thank Jennifer Ernst (and the rest of the Creative Services Group) for many valuable insights into the nature of the video editing process. Rich Gold, Victoria Bellotti, and Steve Harrison have each contributed with helpful readings of and/or conversations about the material contained herein.

REFERENCES

[1] D. Andrew, *Concepts in Film Theory*, Oxford University Press, New York, NY, 1984.

[2] A. Barr and E. A. Feigenbaum, eds., *The Handbook of Artificial Intelligence*, volume 1, chapter 3, 141–222, William Kaufmann, Los Altos, CA, 1981.

[3] A. Barr and E. A. Feigenbaum, eds., *The Handbook of Artificial Intelligence*, volume 1, chapter 4, 223–321, William Kaufmann, Los Altos, CA, 1981.

[4] A. Damasio, *Descartes' Error: Emotion, Reason, and the Human Brain*, Grosset/Putnam, New York, NY, 1994.

[5] U. Eco, *A Theory of Semiotics*, Advances in Semiotics, Indiana University Press, Bloomington, IN, 1976.

[6] U. Eco, *Six Walks in the Fictional Woods*, The Charles Eliot Norton Lectures, Harvard University Press, Cambridge, MA, 1994.

[7] G. M. Edelman, *Neural Darwinism: The Theory of Neuronal Group Selection*, Basic Books, New York, NY, 1987.

[8] G. M. Edelman, *The Remembered Present: A Biological Theory of Consciousness*, Basic Books, New York, NY, 1989.

[9] B. Furht, S. W. Smoliar, and H.-J. Zhang, *Video and Image Processing in Multimedia Systems*, Multimedia Systems and Applications, Kluwer Academic Publishers, Boston, MA, 1995.

[10] A. Hampapur, R. Jain, and T. Weymouth, 'Production model based digital video segmentation', *Multimedia Tools and Applications*, **1**(1), 9–46, (March 1995).

[11] D. Hume, *A Treatise of Human Nature*, Dent, London, England, 1911. Introduction by A. D. Lindsay.

[12] D. P. Huttenlocher, J. J. Noe, and W. J. Rucklidge, 'Tracking non-rigid objects in complex scenes', in *Proceedings: International Conference on Computer Vision*, pp. 93–101, Berlin, GERMANY, (May 1993).

[13] J. McCarthy, 'An example for natural language understanding and the AI problems it raises', in *Formalizing Common Sense: Papers by John McCarthy*, ed., V. Lifschitz, 70–76, Ablex Publishing Corporation, Norwood, NJ, (1990).

[14] S. Minneman et al., 'A confederation of tools for capturing and accessing collaborative activity', in *Proceedings: ACM Multimedia '95*, pp. 523–534, San Francisco, CA, (November 1995). ACM, ACM Press.

[15] N. A. Streitz et al., 'SEPIA: A cooperative hypermedia authoring environment', in *Proceedings: 4th ACM Conf. on Hypertext (ECHT'92)*, Milan, ITALY, (November-December 1992). ACM. Available at http://www.darmstadt.gmd.de/-publish/ocean/publications/SEPIApaper-home.html.

[16] L. A. Suchman, *Plans and Situated Actions: The Problem of Human-Machine Communication*, Cambridge University Press, Cambridge, ENGLAND, 1987.

[17] G. Turner, *Film as Social Practice*, Routledge, London, ENGLAND, 1988.

[18] L. Wilcox, D. Kimber, and F. Chen, 'Audio indexing using speaker identification', in *Automatic Systems for the Identification and Inspection of Humans*, pp. 149–157. SPIE, (July 1994). Volume 2277.