DRO
Deakin University's Research Repository
DEAKIN UNIVERSITY AUSTRALIA Worldly

This is the published version:

Adams, Brett and Venkatesh, Svetha 2004, Authoring multimedia authoring tools, IEEE multimedia, vol. 11, no. 3, pp. 1-6

Available from Deakin Research Online:

http://hdl.handle.net/10536/DRO/DU:30044306

©2004 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

Copyright : 2004, IEEE

dro.deakin.edu.au
Deakin University CRICOS Provider Code: 00113B

Frank Nack
CWI, Netherlands

Authoring Multimedia Authoring Tools

Brett Adams and Svetha Venkatesh
Curtin University of Technology

It's hard to find people with no video-capture experience these days. In fact, capturing devices, while continually becoming smaller and easier to use, have increased in capacity. They're also more connectable and interoperable, and their propensity to show up where we least expect them is surprising. Perhaps the average household toaster will soon come equipped with a video camera—"No, honey, that's the multimedia card slot. The toast goes over there…."

Yet, despite all these advances, the video-capture experience is still frustrating. So, what's the problem? Precisely, it's difficult to determine what to capture and how, and how to handle the ensuing process required to transform the raw captured footage into a presentable multimedia artifact.

Frustrated users

Consider a typical home videographer in a holiday scenario. Holidays often combine two incentives to bring out the camera: new sights and sounds, and memorable moments. Depending on how trigger-happy our videographer is, he or she might start the film rolling and in no time gather quite a bit of footage. In the worst case, no further processing occurs, and the captured footage assumes the title of movie. This could mean hours of footage peppered with wandering, panning, and zooming camera work. The end result is often boring, and the viewer falls asleep.

Our home videographer might be willing to process the captured footage with standard commercial software. But this is often frustrating because the available tools are essentially low-level film manipulations under the hood, requiring much patience and some editing knowledge. The most useful outcome is often just a shortened version of the original footage.

Not surprisingly, this postproduction process is difficult for the amateur videographer. People expect to emulate what they've experienced, and they're saturated with highly produced media—finished products whose authors are learned in the use of the film medium. But Michelangelo painted and sculpted well because he practiced to distraction. Churchill orated famously, despite a speech impediment, because he rehearsed. We can't expect our amateur videographer to automatically author effective multimedia.

Defining the domain

Researchers are only too eager to try to solve such an interesting problem. The "Multimedia Authoring Approaches" sidebar shows a few approaches to multimedia authoring. A quick glance reveals a heterogeneity of context that is representative of the field of computer-assisted multimedia authoring and serves to highlight an important requirement for researchers: They must precisely define the intended domain of the authoring technology in question. Here, "domain" means the complex of assumed context (target user, physical limitations, target audiences, and so on) and rules (conventions), which collectively constitute the air in which the solution lives and becomes efficient in achieving its stated goals. Defining this domain doesn't require explicitly enumerating and instantiating every aspect, nor specifying them to a fine point. Genres, abstractions, and catch-alls exist precisely because they help specify ranges of items that creators of authoring tools can more easily handle.

But why do we need to address these side issues? They're kind of hard to determine anyway. Why can't we just concentrate on simple concretes, such as the data type we're dealing with and its intrinsic properties? Raw video capture and manipulation, object tracking, cataloging and matching, and so on would surely provide a firm foundation, wouldn't it? Such a data-type-centric approach might seem appealing, but it doesn't

continued on p. 4

1070-986X/04/$20.00 © 2004 IEEE

Published by the IEEE Computer Society

Multimedia Authoring Approaches

Traditionally, the process of creating a finished video presentation includes three main phases: preproduction, production, and postproduction—or, in abstract terms, planning, execution, and polishing. Here, we present some multimedia authoring approaches from the literature, grouped by the process phase they focus on. We prefer not to emphasize any particular commercial provider, but we encourage you to search for audio or video authoring tools for each relevant phase. You'll be amazed at what you find.

Planning

- R. Baecker et al., "A Multimedia System for Authoring Motion Pictures," *Proc. ACM Multimedia 1996*, ACM Press, 1996, pp. 31-42.
- B. Bailey, J. Konstan, and J. Carlis, "DEMAIS: Designing Multimedia Applications with Interactive Storyboards," *Proc. 9th ACM Int'l Conf. Multimedia*, ACM Press, 2001, pp. 241-250.

Execution

- L.-W. He, M. Cohen, and D. Salesin, "The Virtual Cinematographer: A Paradigm for Automatic Real-Time Camera Control and Directing," *Proc. Computer Graphics Int'l*, IEEE Press, 1996, pp. 217-224.
- B. Tomlinson, B. Blumberg, and D. Nain, "Expressive Autonomous Cinematography for Interactive Virtual Environments," *Proc. 4th Int'l Conf. Autonomous Agents*, 2000, pp. 317-324.
- B. Barry and G. Davenport, "Documenting Life: Videography and Common Sense," *Proc. Int'l Conf. Multimedia and Expo* (ICME 03), vol. 2, IEEE Press, 2003, pp. 197-200.

Polishing

- A. Girgensohn et al., "A Semi-Automatic Approach to Home Video Editing," *Proc. 13th Ann. ACM Symp. User Interface Software and Technology*, ACM Press, 2000, pp. 81-89.
- J. Casares et al., "Simplifying Video Editing Using Metadata," *Proc. Symp. Designing Interactive Systems* (DIS 02), ACM Press, 2002, pp. 157-166.

Holistic: From planning to polishing, and optionally to repolishing

- M. Davis, "Editing out Video Editing," *IEEE MultiMedia*, vol. 10, no. 2, Apr.–June 2003, pp. 54-64.
- F. Nack, "From Ontology-Based Semiosis to Computational Intelligence: The Future of Media Computing," *Media Computing Computational Media Aesthetics*, C. Dorai and S. Venkatesh, eds., Kluwer Academic, 2002, pp. 159-196.

continued from p. 1

help define the chief problem of technology: What is the nature of the gap we're trying to bridge?

Defining the gap

Consider the following example. In the editing room of a professional feature film, the editor often needs to locate a piece of footage to fill a need. The problem is retrieval, and the gap is information on the desired clip's whereabouts. In the editing room of one of the *Lord of the Rings* movies, there were times when editors had to chisel hours of footage down to seconds.

Contrast that situation with a home videographer trying to assemble a vacation movie from assembled clips. Here, the main challenge is basically to determine how much of what should go where. The problem is which sequence makes an interesting progression, and the corresponding gap is an understanding of film grammar and narrative principles (or some other theory of discourse structure). This situation is analogous to finding a citation and knowing the principles of essay writing, respectively.

We aren't saying every home videographer wants to produce something like a Hollywood feature. But people do have certain, at least implicit, expectations. They want, for example, a sense of continuity; a cohesive theme; and well-composed, cleanly edited material. When their production ends up being little more than an unmediated record of events, they're naturally disappointed. Sometimes the user can't name the source of this dissatisfaction, let alone remedy it, indicating that many people aren't even aware of what they can do with the medium.

Closing the gap

Once we've better defined the gap our technology is trying to bridge, we're in a position to propose solutions. For our amateur videographer, the solution is perhaps either education or a lowering of expectations. Such procrustean solutions, however, amount to surrender, and users are generally ill-inclined to surgery of any sort, be it film school or subliminal messages of pessimism.

However, we could provide the video equivalent of a thesaurus and spellchecker: squiggly lines

under the offending text. We could even offer grammar checking, providing additional supports at a syntactic level. This would help a little, but it still wouldn't be enough. The problem is twofold. First, we must formulate the cogent argument or message we wish to display—for example, this scene goes before that one and supports this idea. This is called the *discourse*. Second, we must express this discourse in a surface manifestation (such as video) using the particular medium's powers to heighten the discourse's effectiveness.

Formulating content structure

Getting beneath the surface of the problem, what can we assume our amateur videographer knows? People are good at content: What is it? This footage contains two people skiing. What semantic associations does the footage carry? The older man and the younger man are father and son; skiing can be fun yet painful. This commonsense analysis is just the sort of knowledge that's difficult to automatically extract from media, model, and ply with computation.

However, if we liken the selection and ordering of footage to the building of an argument, we note that the user's inherent understanding of the isolated pieces (fathers and sons) isn't sufficient to help him or her form a reasoned chain of logic. The goal might be an enjoyable home movie, but the user doesn't know how to get there. Those very same pieces of knowledge (sons, fathers, skiing, and pain) don't say anything about how to combine them after plucking them from their semantic webs to serve the argument.

Enumerating possibilities from a given set of rules and constraints, however, is something we've had success at expressing algorithmically. For example, simple software wizards can help give the needed structure some flexibility, letting the user include discourse elements relevant to his or her presentation within the provided structure.

Modeling discourse structures is a necessary first step. The particular, relevant theory of discourse and corresponding models will vary. If the goal is an entertaining home movie, some type of narrative model is appropriate—perhaps something simple that allows coarse structures such as crises, climaxes, and resolutions. Or, a more complex model might be appropriate—for example, Dramatica, which views a story as an argument and has a host of structures and roles that a user could potentially instantiate (http://www.dramatica.com/). Rhetorical structure theory might apply to domains in which the veracity of the resulting generated presentation is more important.[1] Having structured a discourse in abstract terms for a presentation, we must now instantiate those discourse terminals with living, breathing content; for amateur video, this means footage of concretes that match the abstract contract.

Expressing form

The next problem to address is how our home videographer can capture the footage. A given video-clip capture veritably bristles with parameters. Light, composition, camera and object motion, camera mounting, *z*-axis blocking, duration, and focal length (to name just a few) all greatly impact the captured footage, its content, and its aesthetic potential. We can't expect our amateur to have domain-specific knowledge such as how to map discourse elements to well-formed, effective surface manifestations.

Film theory has done much work for us in defining appropriate parameterizations for given discourse-related goals.[2] For example,

- for emphasis, shot duration patterning and audio volume;
- for emotional state, color atmosphere;
- for spatial location, focal depth and audio clues; and
- for temporal location, shot transition type and textual descriptions.

So, it's possible to encode mappings relating the presence of discourse structures to cinematic parameters. However, an important caveat is that the degree of freedom for the requested manifestation parameterization—for example, a shot of your son next to his friend—is not unlimited if the user's context is the kind of impromptu setting common for amateur videographers. At times, it's possible to relieve such complications through reuse of existing media.

Media reuse

Reuse of media artifacts is a natural benefit of maturing media technology, mainly its metadata aspects. Examples include extending in time or purpose the initial use to new audiences, multiple versions of the same presentation, and so on. Reuse springs from a desire to harness the initial effort required to create media artifacts and to make it abound to repeated or repurposed

uses, as with industrial manufacture: The same bolt, if cast to a standard, can secure a washing machine or a 747, so why can't we apply the same principle here?

To answer this question, we must identify the interfaces of a piece of footage. This in turn depends on the domain context. In a genre hierarchy stretching from a moving photo album to a simple thematic narrative, to the more complex traditional narrative, it's difficult to take footage from a lower genre and reuse it in a genre above it. The metadata to place and manipulate footage in the easier genre is insufficient to the task in the target genre.

Imagine you filmed a narrative version of your son's birthday party, including all the fun and difficulties leading up to the climactic opening of presents. Then, you later decide to do a montage of major family events during the past year. It would be simple to select shots from the already existing narrative and insert them into your review. The only constraint would be that the footage must in some sense be a highlight. In the following year, you decide to create a birthday movie for your daughter. Somehow, you end up missing the footage of the presents! In a moment of desperation, you decide to sneak in some footage from your son's movie. But this could cause some problems: If certain children appear in one or two shots and are never seen again, eyebrows will elevate. Or perhaps you've found some footage suitably devoid of people, but you've overlooked the fact that you changed the curtains last Christmas. These are examples of *unmet continuity constraints*. Another potential problem is that maybe in your son's video you have presents framed in extreme close-up, and this runs against the grain of the style of movie you're employing for your daughter. This would be an example of *manifestation constraints* in answer to aesthetic goals.

In short, authoring presentations of any sort require a level of understanding about the content or meaning of the media presented. Reuse of existing material is no exception. Add to this the fact that the genre in use can partly predicate that meaning, and we come to the moral of the story: We must define the source and target genres. Coupled with this information, our metadata, however we obtain it, provides computational assistance for creating our presentation.

Example multimedia authoring system

Accordingly, we've designed a multimedia authoring system to fulfill these requirements. A typical home movie creation flows through a storyboard, capture, and edit life cycle. The process starts with the selection of a narrative template—an abstraction of an occasion, such as a wedding or a birthday party—viewed in narrative terms (climaxes and so on). The user either obtains this template from an existing library (generalized, specialized, or cannibalized) or constructs it from scratch for the intended event. The user specifies a creative purpose in terms of recognizable genres (for example, action or documentary), which in turn map to affective goals. Our multimedia authoring system then automatically transforms the narrative template, coupled with these goals, into a storyboard comprising scenes and attached shot directives with the help of aesthetic structuralizers. This storyboard is the basis for an interactive capture process that results in footage that the system can automatically edit into a film or alter for affective impact.[3]

The end

If we are to create effective multimedia authoring tools, we must avail ourselves of the various disciplines that we normally throw in the "someone-else's-problem" basket. Discourse theory, domain distinctives such as media aesthetics, human–computer interface issues, and multimedia data description standards all have a part to play. However, none of these fields stands still, so we need to continually query them for new insights that might impact our multimedia authoring endeavor. **MM**

References

1. C.A. Lindley et al., *The Application of Rhetorical Structure Theory to Interactive News Program Generation from Digital Archives*, tech. report INS-R0101, CWI, Amsterdam, 2001.
2. H. Zettl, *Sight, Sound, Motion: Applied Media Aesthetics*, 3rd ed., Wadsworth Publishing, 1999.
3. B. Adams and S. Venkatesh, "Weaving Stories in Digital Media: When Spielberg Makes Home Movies," *Proc. ACM Multimedia Conf.*, ACM Press, 2003, pp. 207-210.

Readers may contact Brett Adams or Svetha Venkatesh at Dept. of Computing, Curtin University of Technology, GPO Box U1987, Perth, Western Australia 6845; adamsb@cs.curtin.edu.au or s.venkatesh@exchange.curtin.edu.au.

Contact Media Impact editor Frank Nack at CWI, Kruislaan 413, PO Box 94079, 1090 GB Amsterdam, the Netherlands; frank.nack@cwi.nl.