Socially-Aware Multimedia Authoring: Past, Present, and Future

DICK C. A. BULTERMAN, Centrum Wiskunde & Informatica and VU University Amsterdam
PABLO CESAR, Centrum Wiskunde & Informatica
RODRIGO LAIOLA GUIMARÃES, Centrum Wiskunde & Informatica and VU University Amsterdam

Creating compelling multimedia productions is a nontrivial task. This is as true for creating professional content as it is for nonprofessional editors. During the past 20 years, authoring networked content has been a part of the research agenda of the multimedia community. Unfortunately, authoring has been seen as an initial enterprise that occurs before 'real' content processing takes place. This limits the options open to authors and to viewers of rich multimedia content for creating and receiving focused, highly personal media presentations. This article reflects on the history of multimedia authoring. We focus on the particular task of supporting *socially-aware multimedia*, in which the relationships within particular social groups among authors and viewers can be exploited to create highly personal media experiences. We provide an overview of the requirements and characteristics of socially-aware multimedia authoring within the context of exploiting community content. We continue with a short historical perspective on authoring support for these types of situations. We then present an overview of a current system for supporting socially-aware multimedia authoring within the community content. We conclude with a discussion of the issues that we feel can provide a fruitful basis for future multimedia authoring support. We argue that providing support for socially-aware multimedia authoring can have a profound impact on the nature and architecture of the entire multimedia information processing pipeline.

Categories and Subject Descriptors: H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Evaluation / methodology*; H.1.2 [**Models and Principles**]: User/Machine Systems—*Human information processing*

General Terms: Human Factors

Additional Key Words and Phrases: Authoring systems, socially-aware multimedia, personalization

ACM Reference Format:

Bulterman, D. C. A., Cesar, P., and Guimarães, R. L. 2013. Socially-aware multimedia authoring: Past, present, and future. ACM Trans. Multimedia Comput. Commun. Appl. 9, 1s, Article 35 (October 2013), 23 pages.
DOI: http://dx.doi.org/10.1145/2491893

1. INTRODUCTION

In this issue, we as a community celebrate the fact that multimedia has been a significant area of research for over 20 years. Multimedia data has two general characteristics that separate it from

This work was supported in part by the EU projects TA2: Together Anywhere, Together Anytime (FP7-ICT-214793) and Vconect: (FP7-ICT-287760).

D. Bulterman is now affiliated with FX Laboratory, Palo Alto, CA.
Corresponding author's address: D. C. A. Bulterman, FX Palo Alto Laboratory, 3174 Porter Drive, Palo Alto, CA 94304; email: dick.bulterman@fxpal.com.

DOI: http://dx.doi.org/10.1145/2491893

other types of computing content: it is inherently intended for human consumption and it is inherently temporal. In this context, it is interesting to note that most multimedia research has held neither the end-user nor the notion of time as central themes. Although the end-user has played a central role in viewing multimedia information, enabling significant interaction with media has not played a seminal role in most aspects of multimedia research. We as a community have spent considerable effort on analyzing content, but much less on analyzing the user of that content.

It is interesting to reflect on the issues that have been treated by our community during the past two decades. The first edition of ACM Multimedia in 1993 helped define this agenda for multimedia research. Topics covered at that year's conference included the following.

—Disk scheduling algorithms
—Multimedia file systems
—Transport systems for media distribution
—Media content codes and formats
—Video conferencing and collaboration tools
—(Multi)media synchronization specification languages
—Multimedia authoring systems and interfaces

(The conference also introduced the then radical idea that a telephone could be used as a multimedia terminal device [Schmandt, 1993].)

This list of topics remained relatively static throughout the first ten years of ACM Multimedia, with the exception that content analysis rather than content coding became an important research issue in our community. In all of these areas, media technology played a fundamental role, while the user of multimedia content typically was treated as a passive edge device that simply initiated or consumed content.

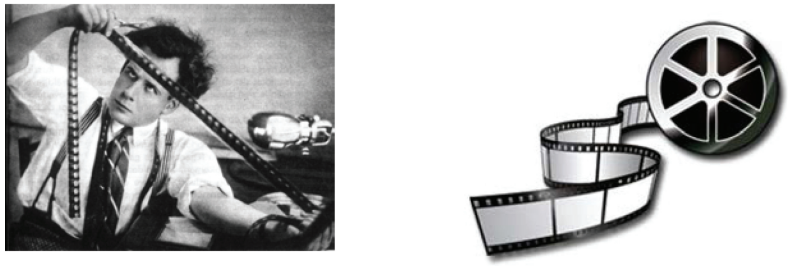
One of the reasons that the nature of the end-user has played only a marginal role is that the focus of much of our research has been to increase the efficiency of media processing. In this sense, the list of research topics that have received the most significant attention can be better described as follows.

—The efficient coding of media content.
—The efficient capture and storage of coded content.
—The efficient searching and selection of coded content.
—The efficient distribution and delivery of coded content.

The need for efficient systems has influenced the nature of tools, languages, and interfaces associated with storing and accessing multimedia content. It has also influenced the manner in which end-users are presented with multimedia data. For the sake of efficiency in storage and transport, most media is delivered as a pre-composed composite object that allows only limited customization at the receiver's end. With the possible exception of the dynamic selection of content encoding based of infrastructure quality, the dominant model for content has been that of a virtual reel of movie film: content is created, encoded, and packaged as an immutable object that needs to be shared in a reliable, cost-effective and predictable manner.[1]

The reel-of-film model (Figure 1) has provided many benefits, but it has also resulted in some significant limitations. Chief among these is the manner in which users can meaningfully interact within a

[1]In this article, we speak of 'viewing' content as if all multimedia were visual in nature. We do this for convenience only. Nearly all aspects of processing encoded media apply to any form for temporally-based multimedia content.



(a) (b)

Fig. 1. A reel of film (or audio tape) has been the dominant model for multimedia content during the past 20 years. From (a) an authoring perspective, the main challenge has been to make the creation of this reel a manageable process. From (b) a coding, storage, selection and transfer perspective, the result has been a closed object abstraction.

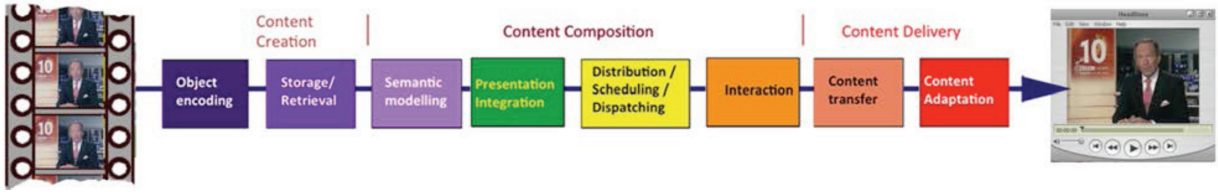


Fig. 2. A typical processing pipeline for multimedia content.

media object. The user is allowed to select composite objects, to view them, and to terminate that viewing. Depending on the rendering environment (and the intermediate content storage model used), the user is sometimes allowed to temporally fast-forward or rewind content during viewing. Put together, this viewing control has not changed much from the VCR-like model of the 1970s. While content-based navigation has become the norm in HTML-like text documents, user-level hyper-navigation in video and audio is missing in all current media formats and players.

The physical film model for multimedia has implications for how content has been processed in media pipelines. Consider the abstraction presented in Figure 2. Here we see a raw fragment of media that is presumably captured by anything from a high-end professional camera to a handheld smartphone. In this pipeline, the content is encoded (usually using closed codecs). The content is then stored, either locally or in a content retrieval network. There will be some metadata created for that object—ranging from a file name or some camera-/location-related metadata, to entire semantic descriptions of the object—allowing it to be retrieved at some later time. Once this object is fetched, there may be some integration with associated objects (such as a branding icon or a set of end-user comments), the object will be dispatched via a guaranteed or best-effort network, all under the control of some high-level start/stop/pause interaction at the client. The content portions ultimately selected will be transferred and decoded, and then displayed by the client.

If we take a step back from this process, we see that the 'authoring' activity occurs at the very beginning of the processing pipeline. In the preceding abstraction, the actual authoring is related to the composition of the scene (in this case, the newsreader, the logo, and the content projected behind his back) all are composited before the media object is encoded. At an even higher level, the semantic structure of the content—the number of news stories in the compilation, their positioning within the 40 minute temporal scope of the object, and the structure and duration of each individual story within

the object—all are predetermined by the editorial staff in charge of the production. We refer to the process of selecting, organizing, prioritizing, and compositing of source object into the final presentation as the *temporal binding of content*.

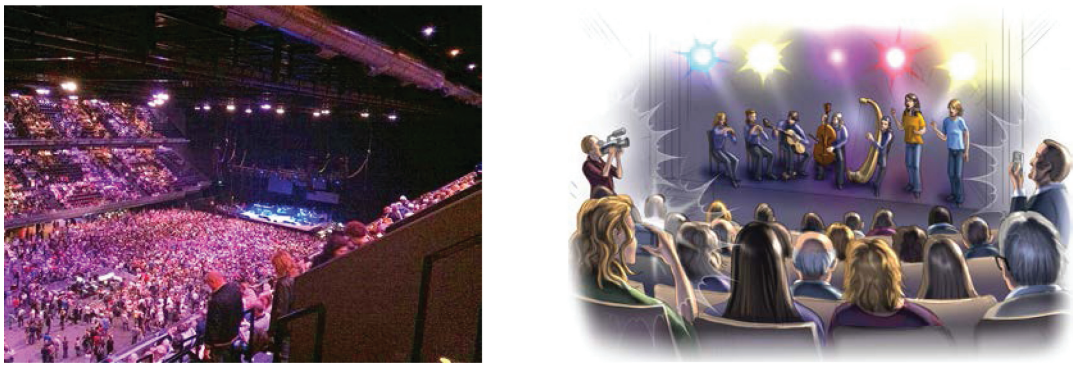
Most multimedia authoring systems promote *early binding* of content. As in the preceding illustration, the choices are made once for a (potentially) broad audience. Early binding is usually appropriate when creating generalist presentations. Examples are not only news broadcasts, but also the typical broadcast-yourself style of YouTube videos. Early binding is useful for the efficient coding, storage, transmission, and delivery of media content: there are no surprises in terms of what happens next, since the order and duration of the media object content is determined by the encoding of the virtual movie reel. In contrast, *late binding* of content allows an end-viewer to have more control over which segments are presented, their relative placement in a presentation, and perhaps even the duration of each segment. Note that the notion of a segment in this context is scalable: in the news example, a segment could be a story, a portion of that story, or even a shot (or frame) within that story. At a high level of abstraction, typical recommender systems can be seen as a form of late binding of content: they allow top-level selection to made based on the preferences of the viewer instead of a broadcast scheduler. When taken to an extreme, late binding can allow for totally customizable presentations, with each fragment determined based on some form of preference processing.

In this article, we survey past approaches to multimedia authoring, with a special focus on support for situations in which both the original presentation creator and the presentation viewer play a role in determining presentation content. We call this *socially-aware* multimedia authoring. In this class of multimedia, we investigate means of late binding of content based on the social relationships between information providers and information consumers. We focus on community authoring applications, where content is contributed from many sources and distributed within a relatively closed circle of viewers who have varying degrees of affinity with the content produced. We consider general issues provided by the paradigm and review results obtained from a series of initial experiments with a prototype socially-aware environment. In these experiments, the value of a pull model of content packaging is contrasted with a more conventional push model of pre-created content. We also discuss a number of broader research issues that we feel can flow from a more viewer-centric model of media production.

In the following sections, we provide an overview of the requirements and characteristics of socially-aware multimedia authoring within the context of exploiting community content. We continue with a short historical perspective on authoring support for these types of situations. We then present an overview of a current system for supporting socially-aware multimedia authoring within the community content. We conclude with a discussion of the issues that we feel can provide a fruitful basis for future multimedia authoring support. Our conjecture is that support for late binding of content, together with a process of understanding, managing and adapting to (dynamic) user relations and preferences, will require new content authoring tools that consider personalization as a fundamental requirement rather than as a luxury add-on, as is currently the case. As a consequence, content creation will need to consider context migration and personal protection (both in terms of content privacy and information integrity).

2. COMMUNITY CONTENT AND SOCIALLY-AWARE MULTIMEDIA

Much of the media landscape has been, and continues to be, dominated by commercially produced content. Whether image, video, audio, or (to a lesser extent) text, users today have become accustomed to experiencing highly polished media messages. Does this mean that user-generated content is unimportant? No. For many users, media objects with a high degree of personal value consist of fragments that come from personal archives: photos of family and friends, videos of small children, audio fragments that capture the sounds of people who have played an important role in one's life. Although there may



(a) (b)

Fig. 3. Two extremes of media capture at a concert: (a) media for the masses, and (b) media only a mother could love.

always be anomalies, it is clear that a short video showing a child's first violin solo will not attract the same audience as, say, a slickly-produced commercial music video. This does not make the violin fragment less valuable within a select social group.

In order to better understand the role of highly-personal user content, we consider two aspects: content capture/creation and end-user viewing control. Historically, capturing personal media was a special activity that occurred at major life events (birthdays, holidays, weddings, funerals, etc.). With the ubiquity of capture devices, there is ample evidence that this capture context is shifting: most of the media currently recorded concerns less structured and more transient daily events: a shared meal, a ride on the bus, a visit with friends.

To help structure our discussion, we consider a medium-complexity media capture session: that of a parent who is collecting media at a concert. Figure 3 shows two extremes of capture/sharing possibilities. In Figure 3(a), we see a major public event: a concert held at an 80,000 seat stadium. In Figure 3(b), we see a concert at a local high school. We consider the implications of both settings in turn.

In the content-in-the-large setting of Figure 3(a), our parent may record parts of a concert and share them with friends to highlight his/her personal enthusiasm from being at the event. The actual content recorded with be incidental to the higher-level message of "I was here!" Of course, if a few thousand of the participants produced such self-labeling content and all the content were gathered on a central server, and if that content were temporally aligned, it is also possible to build an aggregate 'concert' story out of the individual fragments [Su et al. 2012; Shrestha et al. 2010]. Building an aggregate story is made relatively easy by the fact that individual shots could be selected nearly at random from the source media available: the intention of the composite video is to give an overall impression of the event and not to convey the personal experiences of any one attendee.

Contrast this with the setting in Figure 3(b). Here, we see our parent (the redheaded mother in the green dress) who, together with other parents, is capturing part of a school concert event. This content-in-the-small event differs from the large-scale concert in a number of ways. First, the mother is not recording the content to advertise that she was there, but because she wants to maintain a memory of her child. Second, it is not the quality of the music or the exclusivity of the event that is the main motivation for content recording, but the emotional attachment to a particular performer. Third, in the long term, this one event will be less important than its place in a series of life events that involve the persons recording and being recorded.

As with the situation in Figure 3(a), the content from all parents in Figure 3(b) could also be aggregated and processed as part of a larger narrative. One important difference between the two settings, however, is that in the school concert case, random shots cannot be selected for inclusion in a personal video. The mother of the trombone player may have only limited emotional attachment to, say, the child playing clarinet. This limited attachment also applies to viewers of the content. In a mash-up of the concert, each parent—and each member of the extended social circle of the parent—will likely want their own favorite performer to be the star of the show, but probably at varying levels of detail. Note, however, that relationships (and thus, content affinity) can change over time: if the trombone player and the clarinetist establish a more defined personal relationship, a future viewer may want to see more of both (and less of earlier partners for each player). As a result, the complexity of the media management task for user-producers has increased, as has the volume of personal media content.

This increased complexity can lead to a number of interesting research questions, such as the following.

—How do you efficiently manage a family video archive, when contributions come from a number of different members, each with their own sharing communities?

—How do you efficiently label the relationships among persons, places, and events in the context of a life-long evolvement of relationships for participants? (Consider the notion of 'home' for a 16-year old person: during the expected following 70 years of life, this notion will likely change multiple times.)

—How do you balance the needs of content-owner protection (both in terms of the physical bits of content recorded and the logical/semantic content represented by those bits) and the innovative use of content during the lifetime of the media clip?

This (non-exhaustive) list focuses on issues of content archiving, content labeling, and content control. It does not, however, address the issue of *content interaction*: how does an end-user who is not the original content creator interact with content available from shared sources. Note that the qualification "from shared sources" is very important: most commercial content is laden with licensing restrictions that make meaningful interaction (such as incremental selection, linking, copying, extending, enhancing, and even redistribution) legally difficult. The situation for shared content is fundamentally different. Shared community content is less restricted on potential reuse and intelligent content navigation, at least within some notion of accepted communities of use. The father of a young musician probably doesn't want to see his daughter teased via content in which she makes a musical mistake, but otherwise he might not have many 'creative control' concerns on the reuse of content uploaded to a sharing site. This provides unprecedented freedom for end-users to create their own personal media compositions, using multiple sources.

From our perspective, creating new versions of content from an existing baseline is an authoring activity. This authoring activity can be made richer is late binding of content to any particular narrative is supported. It can also be made richer if at least some of the content is pulled on-demand by the viewer rather than solely determined by the biases of one content author. It is inherently social, since social interactions drive content (fragment) selection, composition, and viewing.

3. HISTORICAL PERSPECTIVE

At the 1993 ACM Multimedia Conference, we presented a paper on structured multimedia authoring [Hardman et al. 1993]. Just over a decade later, we revised this study for the initial issue for ACM TOMCCAP [Bulterman and Hardman 2005]. At that time, multimedia authoring was seen by many as a seminal research topic. As described in these publications, several paradigms existed for compositing (or binding) media objects, including the following.

—*Structure-Based Composition*. Composition where the (often hierarchical) logical structure of the components serves as the basis for generating a particular presentation instance timeline.

—*Timeline-Based Composition*. Composition in which a particular presentation instance determines the content relationships among objects.

—*Graph-Based Composition*. Composition in which the relationships among objects have cause/effect relationships, but limited logical structure.

—*Script-Based Composition*. Composition in which the inherent logical structure of elements is hidden as side-effects of a procedural execution model.

All of these methods (of which structure-based remains the most compelling) are examples of relatively formal models in the sense that there is a need for an explicit authoring activity to take place in creating a presentation. This explicit activity was intended to manage the inherent complexity of selecting, editing, combining, and positioning media in temporal and spatial dimensions. In many ways, the process was similar to early text processing systems, in which formatting codes and layout directives needed to be directly and overtly inserted into a content stream.

One of the first Grand Challenges to the multimedia research community was to develop media authoring tools that would make creating complex media titles less formal and as easy as using a WYSIWYG word processing system [Rowe and Jain 2005]. Since that time, a number of consumer-level video editing tools have been developed that would lead a casual observer to believe that multimedia editing is a solved problem: even relatively novice content editors using tools like iMovie or Windows Movie Maker (or even more sophisticated tools such as Adobe Premiere or Apple's Final Cut Pro), were able to create media productions. The process was further simplified by modern content capture tools, such as smartphones, in which recording, (simple) editing and integrated uploading were combined into a single task. In many ways, these tools have reduced 'authoring' to capturing content, throwing out unwanted segments and uploading the result to a sharing site.

While it is indisputable that media sharing is much easier than at any time in the past, we wonder if the resulting products of such authoring interfaces have provided any significant advances for the viewers of media content. It is even questionable if there have been significant advances for content authors. Recent data suggests not. In spite of the ubiquity of video capture and sharing options, more than 86% of Internet users have never uploaded even a single video [Purcell 2010]. Although over 80% of YouTube uploads are amateur content, over 90% of the YouTube views are for professional content [Kruitbosch and Nack 2008]. The reason for this is that creating compelling videos—videos that meet the needs and desires of the viewer, not only the producer—is a complex task [Eisenstein et al. 1949]. Although a number of research efforts have addressed content creation from different perspectives [Adams et al. 2005; Kirk et al. 2007; Cattelan et al. 2008; Shipman et al. 2008], it is clear that new paradigms are needed in order to enable end-users to create and share personal media with others.

Formal authoring systems all are based on an implicit model in which an editor is assumed to understand the basic aspects of content production. These include understanding;

(a) the content alternatives available,

(b) the interests (and attention spans) of the intended audience, and

(c) the formal or informal narrative and cinemagraphic principles required to build a compelling story.

While significant steps have been made in better understanding the encoding of narrative structures [Sundaram and Chang 2000], the management of content and the management of viewer-driven interests provide fruitful areas for new work. We argue that there are two primary reasons that personal content viewers are unresponsive to nonprofessional content. The first reason is that the opportunity

to home-editors represented by (b) is largely unexploited by formal authoring systems. In many professional editing situations, all three of these aspects have been well understood, albeit for (b) at an aggregate level of detail. For more personal content, home editors would seem to have a tremendous advantage: they typically know the person or persons for whom a particular content object is being created. Sometimes the intended audience is relatively diffuse (such as one's set of Facebook friends), but other times it can be highly focused: the grandmother of a young high school musician. The second reason that personal content viewers are unresponsive to nonprofessional content is that conventional formal authoring systems do not support a pull model in which a content viewer is intimately involved in the process of content selection and personalization. This means that the author/editor (and not the viewer) determines all of the choices related to end-user personalization at the detail level.

3.1 Interactive Storytelling and End-User Authoring

During the past decade, various AI approaches have been suggested for the creation of configurable and interactive storytelling [Ibanez et al. 2009]. One representative example in this direction is Vox Populi [Bocconi et al. 2008], in which rhetorical documentaries are created from a pool of media fragments. Another example is the Narrative Structure Language (NSL), a production-independent framework for the authoring and delivery of configurable and interactive video narratives [Ursu et al. 2008]. More recently, a system capable of creating different story variants from a baseline video was presented [Piacenza et al. 2011].

Generally, such systems sequence video fragments, while maintaining local video consistency. In order to support the automated generation of the interactive story, extensive use is made of metadata annotations on media objects. These systems typically use well-defined (and generic) content and story descriptions. Our views on socially-aware multimedia authoring differ from these interactive storytelling approaches in two important ways. First, the community content that we consider is not professionally produced and annotated. While a reasonable degree of person and object recognition is possible, the poor lighting and overall moderate quality of the content often requires user intervention to classify and locate content fragments. A second difference is that the narrative definition used with commercial systems is typically plot-based, using a well-developed semantic story framework. In our initial work, the emphasis was on structure-based narratives, using a catalogue of shot styles to select clips rather than an evolving plot line.

3.2 Community Video Mash-Ups

A second thread of more general story development is represented by work on video mash-ups and content repurposing. In this respect, it is interesting to note the current shift from local-based home videos management systems [Lienhart et al. 1999; Abowd et al. 2003] to global-based video sharing internet services.

Kennedy and Naaman [2009] describe a system for synchronization and organization of user-contributed content from live music events, creating an improved representation of the event that builds on the automatic content match. Shrestha et al. [2010] report on an application for creating mash-up videos from YouTube recordings of concerts. They present a number of mechanisms (including temporal alignment and content quality assessment) that are used for creating multi-camera mash-ups. Naci and Hanjalic [2007] report on a video interaction environment for browsing records in music concerts, in which the underlying automatic analyzer extracts the instrumental solos and applause sections in the concert videos, and also the level of excitement along the performances. Lately, crowd-sourcing has been proven to be a good basis for content analysis, for example, fans of a band can be useful for improving content retrieval mechanisms. Video search engines have been shown to allow

user-provided feedback to improve, extend, and share automatically detected results from video footage recorded during a rock n' roll festival [Snoek et al. 2010].

Other interesting works propose community video remixing [Shamma et al. 2007], video repurposing [Pea et al. 2004], video enrichment systems that enable reciprocity [Cesar et al. 2009]. In this direction, it is important to mention current practices around news stories, where users can reuse fragments of video clips for expressing opinions [Xie et al. 2011]; and tools that are capable of producing audiovisual media show based on events, people, locations, and time, taking into account the interpersonal ties [Singh et al. 2012].

4. SUPPORTING USERS-IN-THE-SMALL: AN EXAMPLE

The surveyed approaches provide important contributions to manipulating community content, but they do so at an abstract and impersonal level. Individual fragments are drawn from an equivalence class of scenes, in which there is no single correct selection. In contrast, socially-aware multimedia authoring intends to help end-users generate stories in which social bonds between people play a major role. Scenes are selected because they have a high relevance to the underlying story being developed and the particular viewer of that story. Relations between the viewer and the content become an important determinant for shot selection. During the past two decades, we have undertaken a series of projects that have tried to understand how authoring systems can better support the process of creating compelling content for delivery in a network setting. In this section, we report on one such project: *MyVideos*.

4.1 Participant Motivations

Our work on MyVideos is part of a multi-year study with performers, parents, and family members who create and consume small-scale concert content. Potential users have been involved in the design and evaluation process since the beginning of the project, starting with interviews and focus groups, leading up to the evaluation of a prototype. Starting in December 2009, the parents were invited to a focus group that took place in Amsterdam; in April 2010 they recorded (together with some researchers) a concert of their children. From July–September 2010, these parents used our prototype application with the video material recorded in that concert. Based on the feedback and results, the software was redesigned in a second phase. This second time, we involved a high school in Woodbridge (U.K.), where a concert was recorded in November 2011. During these years, we have systematically investigated mechanisms for helping end-users explore assets from a community collection of videos and to automatically generate "stories" from these assets based on a narrative model [Guimaraes et al. 2011; Zsombori et al. 2011].

Our approach in MyVideos has been to deploy an initial test environment for interim evaluations (after about one year) and to then use the results of this evaluation to design a refined system for the second phase of the project. A general timeline is shown in Figure 4.

After a ten-month evaluation that included video recording sessions and field trials, we concluded that participants were in general quite enthusiastic about our socially-aware approach. They were proud of their own productions. They actively looked for video clips of their close friends and relatives and complained when the quality of the video was not good enough or when the video annotations were wrong. Most importantly, they wanted to share video clips featuring members of their close circle. "Can I send it now?" was the reaction of one of the participants after seeing a video clip he especially liked. In particular, participants largely agreed that the system would help them in recalling memories from social events and that our system would allow them to create more videos and to share them with others. Such results were encouraging, since they reinforced and validated basic social theories (such as social connectedness), which were the basis for designing the prototype implementation.

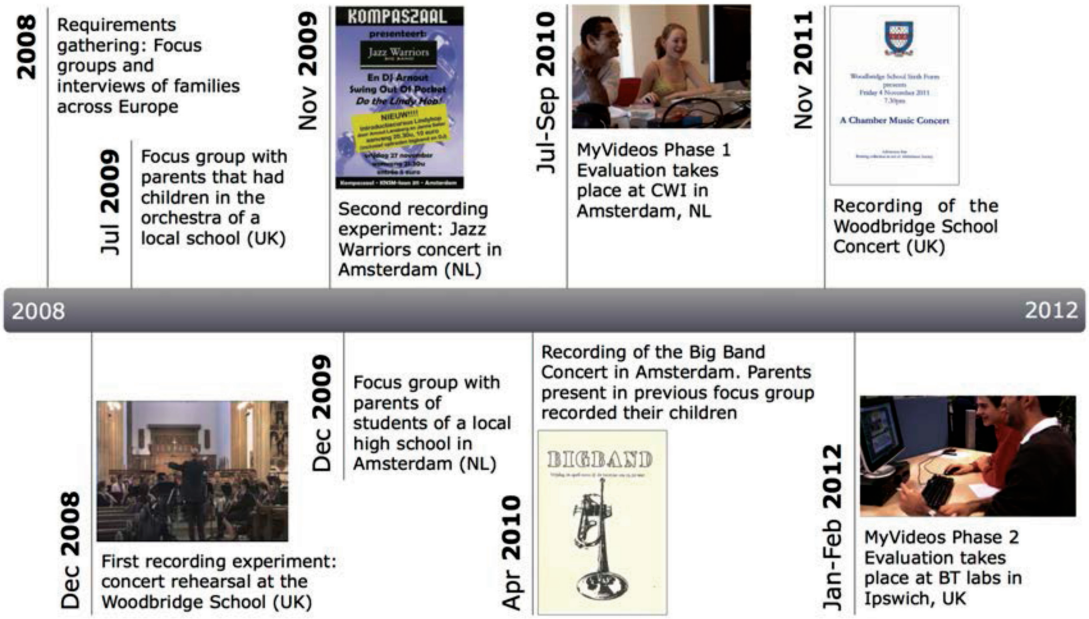


Fig. 4. General timeline of our study on socially-aware authoring principles.

These insights from users helped us design a second iteration of our system. Some of the more general findings of our user groups (performers, parents and potential viewers) can be summarized as follows.

—*Jump-Starting the Authoring Process*. Users indicated their preference for an authoring paradigm in which a first compilation would be created in their behalf. This baseline production could then be extended or altered later, either by the parent authors or the family viewers. Such approach would simplify their task and increase their productivity.

—*Interpersonal Relationships*. Participants assumed that the system could automatically identify and process their interpersonal relationships with the performers. This approach would simplify the preference selection mechanism for triggering the narrative engine.

—*Personal Imprint*. End-user participants requested that automatically generated compilations of video clips could be modified. They wanted to remain in control over the final production, being able to make small changes and tweaks. This approach would allow them to create more personal stories.

The initial phase of our project provided interesting insights into creating composite presentations. Unlike many conventional approaches, authoring was no longer a process of throwing away unwanted bits of "known" content, but an iterative process of content discovery and integration. These same insights applied to end-user viewers (in our experiments, family members who were not at the original concert). Here, the same desires for content discovery and high-level tailoring focused on a single individual was the norm.

4.2 Issues in Supporting Socially-Aware Capture and Sharing

It is difficult to frame any meaningful discussion on the facilities required for socially-aware multimedia authoring in abstract terms. As a result, this section describes the design an implementation of

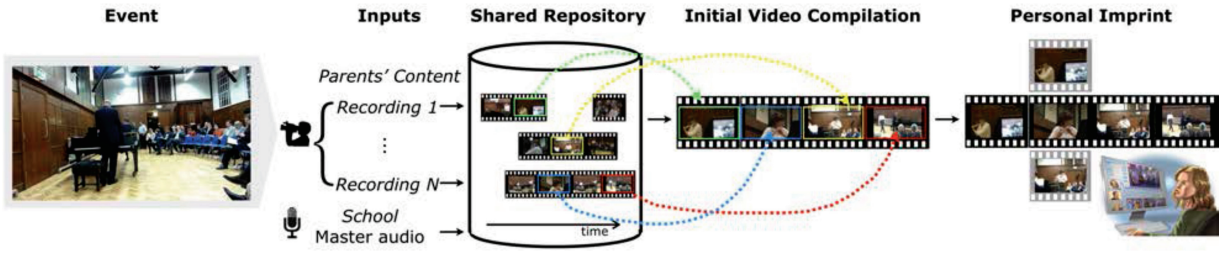


Fig. 5. High-level workflow of the MyVideos hybrid authoring tool.

one of several instances of our MyVideos work. The purpose of this section is not to describe any one component in detail, but to illustrate the facilities required for improving socially-aware multimedia video creation. The high-level workflow of our system is shown in Figure 5.

The processing within MyVideos starts after the concert with the uploading of content to a central server. Each media item is automatically associated with the person who uploaded it, and mechanisms are provided for participants to restrict sharing of certain clips. Trying to recreate realistic situations, there are no specific requirements imposed on users while recording: they could record as many clips as they wished, using their own preferences for content framing, panning, and zooming. This flexibility came at a cost, however, since most existing solutions for automatic video analysis did not work well for content captured with ad hoc devices (e.g., mobile phones), with low quality and low lighting conditions, and with nonstandardized length. It is clear that analyzing user-generated content is challenging and deserves more investigation [Shrestha et al. 2010].

Content was tagged and annotated using a combination of automatic tools and manual approaches [Guimaraes et al. 2011]. In particular, our automatic tools allow for temporally aligning all of the source video clips, and for assigning temporal annotations to each video clip, including the performers that are portrayed, the instruments, and the song that is being played. A content curator double checks the annotations and fine-tunes them so they are consistent and valid. The content analysis part of our system is out of the scope of this article, but it is clear that even high-level personalized annotation of general content is anything but a solved problem. A combination of manual and automatic approaches will be required for capturing any meaningful highly-personal annotations: in our example, seating charts, names of soloists, even the program booklet for the concert provided meaningful external information that was used to help disambiguate automatically-recognized content. All of this still requires manual intervention. It is clear that until automated tools are developed that are steered by viewer interests and social context (rather than only object recognition), the effort required for manual annotation will remain a significant barrier to the development of general solutions.

As shown in Figure 6, the individual clips are processed and sorted in a number of dimensions. Concert material can be accessed by performer, instrument, or piece performed. In this view, the performers have a prominent place in the screen, and the pieces performed are organized in two columns. Two recommend fragments for each song/performer combination are provided automatically by the system. Once a viewer starts browsing content, additional clip options are presented that follow their personal interests. In addition to clip-based browsing, a second facility was available that allowed a personalized compilation to be delivered via an interface with was named the *Director's Cut*. The contents within this compilation are tailored to user tastes during viewing. The resulting dynamic presentation can be saved for iterative sharing.

Fig. 6. Interface for browsing and accessing individual video clips.

4.3 Insights from User Evaluations

Several evaluations took place of the two phases of our project. In this section, we highlight results that relate to the concepts of the socially-aware authoring paradigm. As reported in Guimaraes et al. [2011], the evaluation of the initial system involved three social events. While the first two recording experiments focused mainly on the evaluation of the annotation processes and narrative structures, the third one, a school concert in Amsterdam, allowed us to engage a group of parents, relatives, and friends of performers for later on evaluate the initial version of our system. For the evaluation of the second prototype implementation, new recordings took place in the Woodbridge high school (U.K.) in November 2011. The concert lasted around 1 hour and 20 minutes, in which 18 students performed in 14 songs. A total of twelve cameras were used to capture the concert. The master camera was placed in a fixed location, front and sideway to the stage, set to capture the entire scene (a 'wide' shot) with no camera movement and an external stereo microphone in a static location. Eight cameras were distributed among parents, relatives, and friends of performers. Members of the research team used the other three cameras. In total about 331 raw video clips were captured, some of which were recorded before or after the event.

Thirteen people (from six families) participated in the evaluation of our second prototype implementation. The participants consisted of performers, parents, and other relatives of the teenagers that performed in the Woodbridge school concert. All the participants were English speakers and were currently living in the U.K. Seven participants (∼54%) were 40+ years old; the other six people were in the 11–20 age range, four of which performed in the concert. Six participants were female.

We used a semistructured approach for data collection. Individual interviews started with an explanation about high-level goals and discussion on current video recording and sharing practices.

This was then followed by interaction with the prototype system and questionnaires (before and after).

A detailed analysis of the results of this evaluation is beyond the scope of this article, but there are some general conclusions that can be drawn about the utility of the socially-aware paradigm that can be distilled from our experiences.

—*Event Recall*. One of the most interesting results of the MyVideos process was that users typically do not remember the individual content fragments that they capture during a recording session. Users typically adhered to a write-only practice of capturing content, but not viewing it. Having a system that ingested this content and aligned it was useful, but doing so required more support than simple encoding. Information on the event (structure, participants, program) needed to be imported in order to make sense of the content they created.

—*Personalization*. Authors realized that raw content alone was not compelling, but they felt that manually editing stories in multiple versions for diverse audiences was not a practical option for personalization. The fully-automated generation of tailored stories was a desired goal, but the provision of personalized navigation and interaction was seen as a useful intermediate step.

—*Personalized Supplemental Content*. One of the facilities of MyVideos was the provision of a facility to comment on videos that were created via the system [Guimaraes et al. 2012]. This form of commenting supported in-line comments that allowed a user to add temporally-aligned comments for specific audiences. This facility was highly rated by users.

—*Manual Tweaking*. It was interesting to note that the ability to adjust previously created stories was considered valuable by viewers, but only marginally used by authors. From our perspective, this validates the idea that an end-viewer is more interested in tweaking content received than a content author is in providing multiple custom versions of content for each intended user group.

In general, our findings indicated that the general concept of the viewer-centric socially-aware authoring paradigm were valid, that significant challenges remain before complete solutions can be provided. These challenges are summarized in the following section.

5. RESEARCH ISSUES FOR SOCIALLY-AWARE MULTIMEDIA

Providing support for socially-aware multimedia will significantly impact the support required for effective encoding, storage, classification, selection, transmission, protection, and sharing of (potentially composite) media objects. The principal reason for this is that the context in which media is used will strongly determine how it is classified and accessed. Annotations and metadata will become multifaceted and dynamic, and will be determined by use rather than by design.

The following sections highlight some of the issues that we feel will need study.

5.1 Customized Media Selection

Media selection in socially-aware multimedia is not a case of 'find as many essentially equivalent videos of an event as possible', but 'find the relatively few videos within that event that are relevant to me now, and structure them into a story based on my context (and that of the people in the video)'. A key aspect of this process is to support interactive content selection.

To understand the progress that has already been made in managing a user's ability to obtain media, it is interesting to follow the development through the last 80 years of broadcast content selection. In the period up through the mid-1960's, if users wanted to experience broadcast content, they would need to consult a published program guide, compare its information with a calendar and a clock, and then physically position themselves to view the content when it was available, as is illustrated in

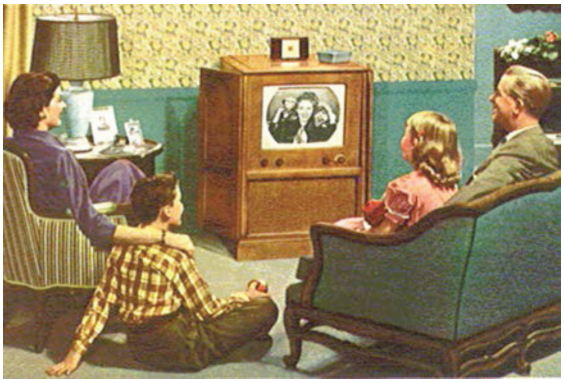


Fig. 7. Conventional media scheduling—the clock on the TV set was the most important scheduling element in this picture.

Figure 7. Note that this is essentially the same process that had consisted since 1927, when the first widely-available radio broadcasts started in the U.K.

Perhaps the most significant innovation in (broadcast) content selection occurred with the introduction of the video tape recorder. For the first time in history, it was the viewer that determined when content would be watched, on the precondition that it had been broadcast and recorded earlier. A next but more minor innovation came with the introduction of the digital set-top box, which included an embedded program guide, providing the opportunity for more automated content selection and recording. After the introduction of the set-top box, the next logical development is to remove the TV guide altogether and to have the system itself recommend content for the family, which it found based on metadata encoded by the content providers.

One drawback of many home content systems is that a set-top box is typically not aware of who is actually watching TV. Some form of personalization is supported, but at a fairly impersonal level. At present, much research is being expended on recommender system technology. These systems depend heavily on producer-generated metadata for determining available candidate content. For socially-aware multimedia, the granularity of the metadata needs to be refocused to personal content. Another change in focus is that content selection will need to move from selecting 'programs' to selecting fragments of content. For a given viewing experience, several fragments typically will need to be dynamically combined to support end-user engagement.

5.2 Media Encoding and Storage

At present, media encoding is based on an agnostic view of content. This has been used to great advantage on sharing websites and physical distribution media. The assumption has been, however, that all of the fragments related to a single story are compressed into a single fixed media object. There are usually no facilities for packaging custom versions of content from a single base encoding. Each personal version of a video (or video fragment) must be re-encoded in a new document.

One important difference required for supporting the end-user dynamic composition implied by the socially-aware multimedia paradigm is that small logical groups of media would be stored on several servers, each as individual fragments. These fragments could be mixed/matched dynamically at viewing time to support the interests of the viewer. In terms of our concert example, this would mean that all of the individual assets captured by all of the parents could be saved in a cloud over servers. Individual presentations could then be stitched together on demand, as shown in Figure 8.

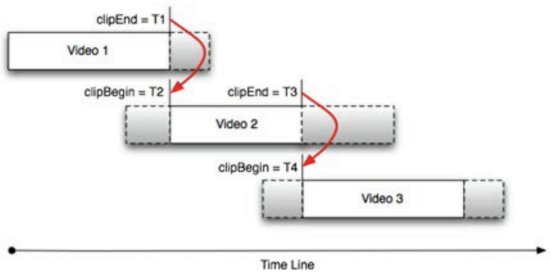


Fig. 8. Each socially-aware media object would be composed of separately-addressable individual fragments, which would be dynamically combined based on end-user interests.

Fine-grained selectivity of content within fragments is impeded to a large degree by the need for the efficient coding of media content. The size of GOP (group-of-pictures) blocks in modern encoders typically results in a temporal decoding distance of nearly ten seconds between major frames in a media object. Our experience in providing systems-level support for efficient navigation and frame-level selection is discussed in Gao et al. [2011]. At present, special-purpose languages, decoders, and media renderers are required to support fine-grained composition. As a result, for most user-level applications, support for late binding of content is nonexistent.

Having a logical media object be constructed out of dynamically combined physical fragments allows customized navigation to be supported. One approach to implementing such dynamic combination is supported by DASH, a system for HTTP-based streaming [Stockhammer 2011]. At present, DASH is typically used for storing predefined fragment encodings, nearly always based on support bitrate-adaptive resolutions. (During presentation, the quality of the content can be adjusted based on environmental factors, such as available bandwidth or enduser screen size.) Adaptivity in our work could leverage this support, but our main interests are in supporting a more abstract form of content selection: providing more trombone content to the father of the trombone player and more clarinet content to the mother of the clarinetist. This is a matter of dynamic content selection rather than (or at least in addition to) dynamic encoding selection. The selection of dynamic content requires more illusive criteria for content selection, such as a profile of the viewer in addition to profiles of the available content, and a content-wide temporal model that exposes logical divergence and convergence points for creating content streams. It also requires a container format that allows differential segment length to exist across candidate segments. To support this, the current model of content streaming would need to be revised: the seamless integration of individual content fragments (as opposed to encoding fragments) into a logical whole is a composition concept that most media servers and media container languages are as-of-yet ill-equipped to support.

5.3 Media Classification and Annotation

Personal media classification and labeling remains a challenge for supporting effective content sharing. For professional content, content is often highly segmented along the lines of established commercial distribution models. For personal content, the situation is vastly different. This shift in emphasis is new for multimedia, but there are many established examples in music, art, and literature where the intentions of the composer, artist, or writer are decoupled from the applications of the media itself. Consider the manuscript shown in Figure 9. Here we see an instance of an ancient Chinese document that provides an excellent example of end-user (instead of author) annotation via red stamps placed on the document. The custom was evidently that each reader would attach his stamp to a manuscript

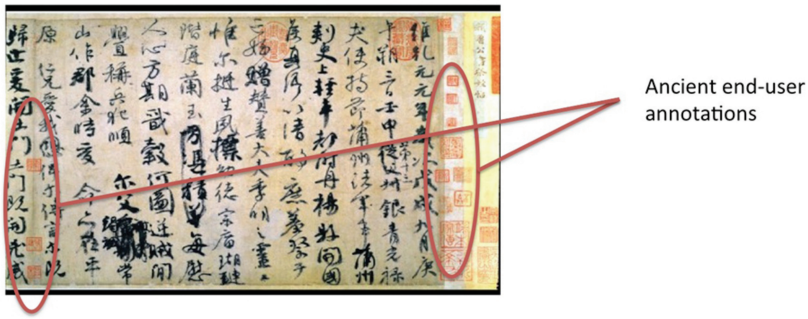


Fig. 9. Annotations as end-user content.

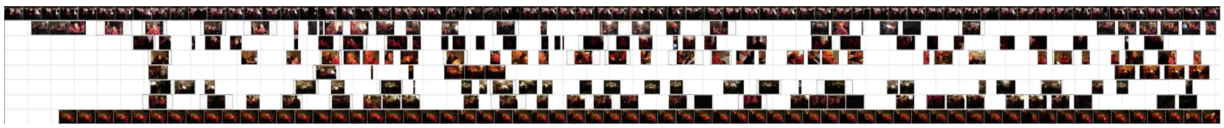


Fig. 10. An overview of media fragments captured at a student concert.

that he considered significant. This form of annotation says simply, I (Mr. X) am now associated with this document.

In many ways, the *stamp of approval* model shown in this document is similar to techniques, such as Facebook's *Like* button: it provides a simple model of end-user association with content (and personal profiling). There is a compelling simplicity to "I was here" annotations: if a system knows something about Mr. X, it could easily attach this knowledge to the content that Mr. X viewed. This is much easier than discovering information on the context itself. Although it will always be necessary to determine semantic information on content fragments, the sheer volume of these fragments (driven by the ubiquity of media capture devices), coupled with the relative poor performance of automatic analysis tools on personal content, will mean that a substantial burden of supporting socially-aware presentation classification will need to come from indirect analysis of content fragments (such as understanding the viewer rather than the viewed media). At present, personal content annotation is driven by device-supplied metadata (clocks, GPS coordinates, file names). For socially-aware multimedia, it is also necessary to encode relative social relationships among interested parties, plus to maintain those relationships over time. As with any large software system, the long-term maintenance costs of media will dominate the short-term development costs. This will require a new generation of iterative, socially-aware media classification tools. The analysis of content becomes a continuing task, not an import activity.

Within the recent capture experiment performed since the MyVideos system discussed in Section 5, it was found that parents at a high school concert captured approximately 150 media fragments at a 60-minute, two set student concert. Some of this content overlapped temporally, but much of it was disjoint in terms of its semantic content. This was the consequence of disjoint views on the notion of "content of interest": there was little overlap across parents each capturing content related to their own children.

One obvious way to organize the content is to analyze the audio signals on each of the content streams and to align the audio tracks of each object with the audio from one of the always-on main cameras. The result of this alignment, which works surprisingly well, is shown in Figure 10.

Fig. 11. Content-free frame-based navigation as supported by YouTube. This is only the tip of the navigational iceberg.

While the temporal alignment of clips is a necessary step for understanding and annotating content, it is clearly only a partial solution. New and efficient domain-specific approaches are required to better classify content. There are several possibilities. A screen grammar could be defined to classify appropriate shots for concerts (either highly public or highly personal). If we were instead filming a soccer game or a wedding, much of the infrastructure would be the same, but the structuring paradigms might be quite different.

In addition to content structuring, logical story structuring could also be developed based on narrative models or other content cues. The main goal of all of this analysis is to delay the moment at which individual content objects are bound to a general narrative structure. The later the binding takes place, the greater the flexibility that exists in creating differentiated, tailored presentations. Of course, later binding also requires a heightened level of user interaction to supply binding context. In many ways, this can be seen as a process of recommendation-based content selection at the scene level, rather than at the program level.

5.4 Improved Content-Based Navigation

One of the challenges with temporal searching along the timeline in Figure 10 is that it is a highly unstructured activity. The time axis provides no information on the logical structuring of the event, let alone the performers in the concert or their relationships. Still, in the absence of any semantic structuring of content, it is often all that is available.

Consider the screen image shown in Figure 11. Here we see a conventional YouTube interface for navigating through a video object. Of course, there is no navigation at a logical level taking place: a user can only select key frames without any higher-level narrative guidance. We note that even 1980's generation DVD technology provided a more significant structure navigation control paradigm through its chaptering interface.

It will be necessary to replace timeline searching with navigation based on an overlay of structure components. One approach to providing this structure in our concert application is based on graphs of performers, instruments, songs, or solos. It could also be based on cinemagraphic classifications, such as long shots, pan shots, tight shots.

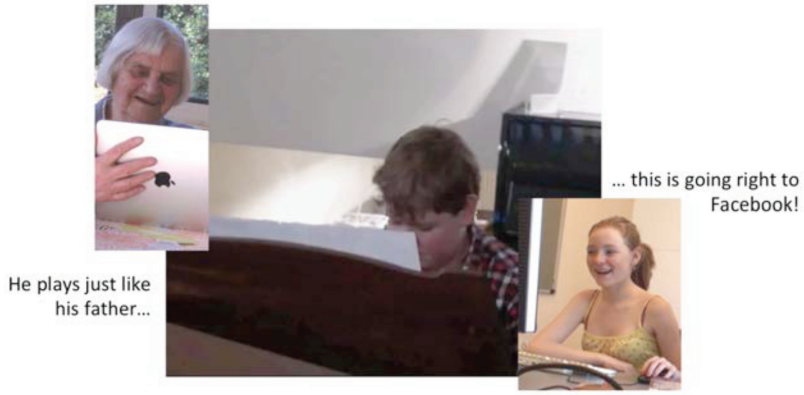


Fig. 12. A botched piano solo could be a source of commiseration when viewed by the piano player's grandmother, but it could also be used as a weapon to taunt the artist when placed in the hands of his sister.

5.5 Security and Privacy Concerns

A 'fun' picture shared with friends can become less fun when those friends turn into enemies. As shown in Figure 12, content could be used or misused by various members of a user community, depending on their (possibly time-variant) relationships. Research is required to support content access and content protection that reflect these time-variant social and personal relationships. In the same vein, content sharing and content recommendation needs to be sensitive to the context of use: are you watching alone, with your spouse, with your children, with your friends? New management and recommender models are needed to manage this situation.

One aspect of security and privacy of socially-aware multimedia is that personally metadata will likely become too sensitive to simply place on a third-party storage system (like Facebook or Google): all of us will want to take back our identity and maintain our own control of our life-long information. This will require convenient interfaces. It will also probably require users to become accustomed to paying for media access and sharing services.

5.6 Incremental Content Authoring

Most current media authoring is predicated on the notion that content creation is a one-time event. In socially-aware multimedia, content authoring becomes an incremental process of content refinement, sharing, and repurposing. "Old" assets remain living entities. This should foster a new generation of create-view-refine-share authoring systems. A key element of this approach is that media gets integrated into some larger narrative story, rather than that the media object is the story itself.

One of the foundations of incremental authoring is that content can take on new meaning based on the insights of downstream users. For text media, this is not a new concept. Figure 13 illustrates how end-user content layering can be applied to a page from a noted technical manuscript. The question is how can a common base page be transformed into a personal page (from the user's perspective, a page in which own commentary has been added, and thus own value). The answer for conventional books is pretty simple: take a pen and mark up the base media.

Changing content brings with it questions of ownership. In printed documents, this is a solved problem: even though the base content is copyright protected, there is a clear distinction between user media and that of the original authors. For webpages and online content, the relationship is less simple.

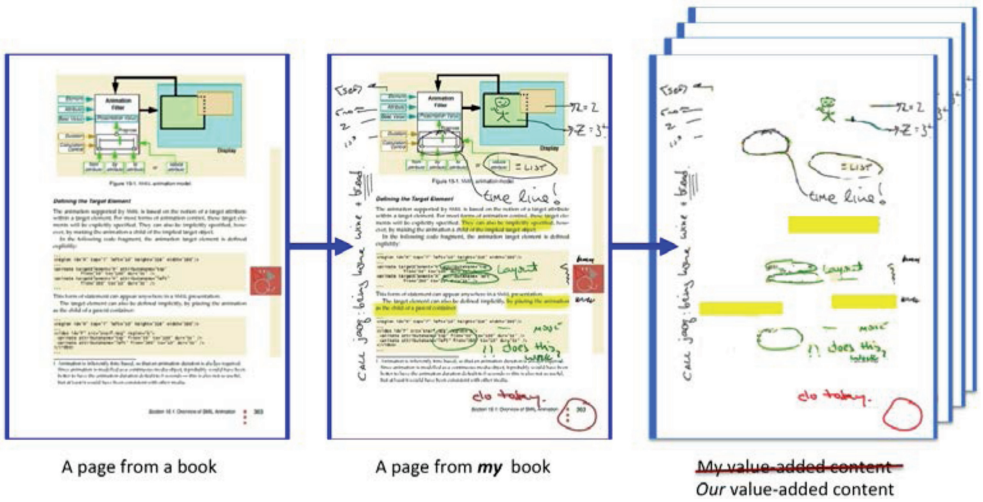


Fig. 13. An example of layering (collections of) personal annotations onto existing content. Socially-aware systems should support similar facilities for all types of temporal/a-temporal media.

If transparent sheets had been placed between all of the pages, we could take all of the user's comments and distribute them as separate items—all fully within current law. The content added could be further aggregated with the context created across a social network (or across the world), and analyzed. What are the most marked-up pages in the book? Does these represent the most interesting or most unclear sections of text? Do the markup patterns change over time? Which comments are appropriate for which users?

A significant authoring challenge is to provide this markup and analysis functionality for all media. When annotating a piece of media—whether it be text, audio, images, or whatever—the implication has been that, as with the book pages shown in Figure 13, the annotations are of a highly personal nature. Of course, if many of these personal notes are collected and analyzed, they could provide valuable insights into the reusability of personal media assets. Even a simple density analysis of multiple media annotations could provide interesting clues for socially-aware recommender systems.

6. ADDITIONAL ISSUES

The entire premise of socially-aware multimedia is that the context of the viewer is more important—or at least equally as important—than that of the content producer. This involves considering context drawn from the experience world of each potential viewer, and realizing that one semantic media size does not fit all users. This brings up many issues of secondary importance that need to be considered and supported. We discuss three of these here.

6.1 Measuring and Maintaining Content Engagement

In the early days of radio, listeners starved for news and entertainment were happy to be able to receive a distortion-tainted broadcast signal. In the early days of television, reflection-induced ghosts and intermittent reception were part of the game of watching TV content. As transmitter and receiver technology improved, a focus shifted from quality reception to content engagement.

Measuring content engagement is a difficult social and sociological problem that transcends coding and delivery. Any potential value for engagement is based on understanding not only the content in question, but also the viewing history and personal context of the receiver. This includes not only

(a) (b)

Fig. 14. Socially-aware systems should encourage rather than derail engagement.

long-term personal data, but also information on where, when, how, and with whom the content experience takes place. This context—or more correctly, these contexts—are themselves not static, but they will vary over time. While parts of any context can undoubtedly be shared among users, much of the context definition will depend on a combination of {source, user, access} information tuples.

Understanding engagement has many practical benefits, some of which we current study within the EU project Vconect.[2] Two examples of topics of interest are the following.

—*Distributed Engagement*. How can audience engagement be measured in a distributed theatre environment, where a performer in one location and his/her audience are spread over multiple other locations. Is aggregate audience behavior (such as the average level of laughing or coughing) a reliable guide, or are more personalized measurements necessary?

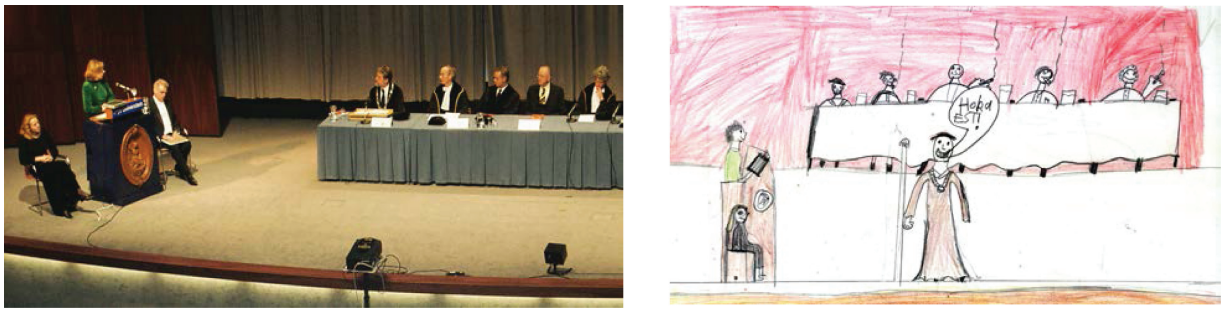
—*Presentation Effectiveness*. Can student engagement be determined within the context of a high school mathematics class. Can a teacher in such a class (even locally) be given feedback information that can help determine who is understanding material and—more importantly—who is not?

Sensor-based engagement measurement provides both opportunities and dangers. How can subjects know when they are being measured, and how can they opt in or out of such a system?

In addition to measuring engagement, socially-aware multimedia system should also assist in maintaining engagement: keeping a user involved in a conversation flow. Many current-generation Web systems are very poor at this sort of engagement management. Consider the situation illustrated in Figure 14(a), where two colleagues are sitting on a train in the Paris Métro system, arriving at the Gare d'Austerlitz station.

Assume that neither are from Paris, and that one asks the other: where does the name 'Austerlitz' come from? Assume further that this question leads to an animated discussion between the two, ranging from historical conjectures to cultural stereotypes, which would transcend any factual resolution of the original question. The 'social quotient' of the encounter would undoubtedly be quite high. Now, consider the same situation, but with the following twist: upon hearing the question, a helpful passenger in the same metro car pulls out her smartphone, types in a search command and proudly shows the Wikipedia page shown in Figure 14(b). While the factual information on Wikipedia goes a long way toward resolving the original question, it also stops the social engagement dead in its tracks. The point of this example is that social media systems need to facilitate the posing of questions and the browsing

[2]http://www.vconect-project.eu/.



(a) An event as seen by a socially-agnostic camera.

(b) An event as seen by a socially-aware 12-year old.

Fig. 15. Representing emotional information.

of alternatives, rather than the resolving of simple content queries. This requires a significant change of focus within information retrieval systems.

6.2 Encapsulating Formats for Media

One of the most important technological bases for supporting dynamic composition and content/presentation adaptation is the use of a declarative specification language for encoding composition information. Such a format allows presentations to be generated dynamically, to be verified, and to be transformed. One such language is SMIL, which was largely developed by our group at CWI. Other alternatives have also been developed, such as the Brazilian NCL language [Soares et al. 2010] for use in interactive TV applications. Whatever the language choice, the most important requirement for dynamic control is a unified notion of time and that is not based on a scripting metaphor. In this sense, it is disappointing that HTML-5[3] only supports placing a single media object in a static webpage without any temporal context, and supports the most basic VCR-like controls.

6.3 Encoding Emotional Content

Consider the very formal academic ceremony illustrated in portion of Figure 15(a). Here, the redheaded Mom from Section 3, still in her green dress, defending her Ph.D. dissertation. A reasonably-informed content analysis system may be able to capture many important aspects of this situation: it may be able to recognize the mom (as long as she keeps wearing the same dress!) or other prominent people, it may be able to use subtle location cues such as university logos, and it may even be able to identify the degree to which participants are engaged with the presentation. But from a personal perspective, this is kid stuff.

The real power is describing the same scene through the emotional eyes of a 12-year old boy in Figure 15(b), who is seeing his mother being grilled by a set of self-assured academic fat cats (in his opinion, of course!), smoking virtual cigars and having a grand time at Mom's expense. In the historical context of the university in question, Figure 15(a) is a valid and persistent model of the event, but for the long-since-grown son, the lasting impression—and the basis of his social context for this event—will be the presentation in Figure 15(b).

[3]http://dev.w3.org/html5/spec/.

7. CLOSING THOUGHTS

Much has changed in the world of multimedia. Who would have expected 20 years ago that within two decades, it would be commonplace to not only listen to music via your computer, but to buy it there as well? That books would not only be written on a PC, but that the PC and its technological cousins would become a handy way to read them, or to have them read aloud. That the computer would threaten to replace not only the television, but also the movie theatre as a venue for the shared watching of content. And, perhaps more significantly in the long term, that the computer would not only render a wide range of real and artificial images, but that it would attempt to understand them as well.

In the previous sections, we have outlined what we mean by socially-aware multimedia. Using several examples, we have argued that the impact of supporting socially-aware multimedia transcends the incremental and provides a number of (fascinating) new challenges that require fundamental research results across a wide range of multimedia disciplines.

This article has presented the idea of socially-aware multimedia as a next step in the evolution of media authoring. By introducing the notional of a temporally-variant social content into media storage, access and sharing, we hope to stimulate a new generator of media research in which the multimedia user is given the central role that she deserves.

ACKNOWLEDGMENTS

The ideas presented in this article have developed over time based on our work with many talented people at CWI, at the VU University, and in over a dozen European projects. They also stem from very fruitful interactions with our research partners in North and South America. We are indebted to the fine interactions that we have had with Jack Jansen, Bo Gao, and many others. A bit further from home, we acknowledge our long-term partnerships with Luis Fernando Soares Gomes, Maria de Graca Pimentel, and Andy van Dam. Finally, we would like to thank the reviewers of this article, who made excellent suggestions for improvement, many of which we integrated into the final text.

REFERENCES

ABOWD, G. D., GAUGER, M., AND LACHENMANN, A. 2003. The family video archive: An annotation and browsing environment for home movies. In *Proceedings of the ACM International Workshop on Multimedia Information Retrieval*. 1–8.

ADAMS, B., VENKATESH, S., AND JAIN, R. 2005. IMCE: Integrated media creation environment. *ACM Trans. Multimedia Comput. Commun. Appl. 1*, 3, 211–247.

BOCCONI, S., NACK, F., AND HARDMAN, L. 2008. Automatic generation of matter-of-opinion video documentaries. *J. Web Semantics 6*, 2, 139–150.

BULTERMAN, D. C. A. AND HARDMAN, L. 2005. Structured multimedia authoring. *ACM Trans. Multimedia Comput. Commun. Appl. 1*, 1, 89–109.

CATTELAN, R. G., TEIXEIRA, C., GOULARTE, R., AND PIMENTEL, M. D. 2008. Watch-and-comment as a paradigm toward ubiquitous interactive video editing. *ACM Trans. Multimedia Comput. Commun. Appl. 4*, 4, Article 28.

CESAR, P., BULTERMAN, D. C., JANSEN, J., GEERTS, D., KNOCHE, H., AND SEAGER, W. 2009. Fragment, tag, enrich, and send: Enhancing social sharing of video. *ACM Trans. Multimedia Comput. Commun. Appl. 5*, 3–19.

DING, J. R., YANG, J. F., AND LIN, J. K. 2006, Motion-based adaptive GOP algorithms for efficient H.264/AVC compression. In *Proceedings of the Joint Conference on Information Sciences (CIS'06)*.

EISENSTEIN, S. 1949. *Film Form: Essays in Film Theory*. Hartcour, New York.

GAO, B., JANSEN, J., CESAR, P., AND BULTERMAN, D. C. A. 2011. Accurate and low-delay seeking within and across mash-ups of highly-compressed videos. In *Proceedings of the NOSSDAV*. 105–110.

GUIMARAES, R. L., CESAR, P., BULTERMAN, D. C. A., ZSOMBORI, V., AND KEGEL, I. 2011. Creating personalized memories from social events: Community-based support for multi-camera recordings of school concerts. In *Proceedings of the ACM International Conference on Multimedia*. 303–312.

GUIMARAES, R. L., CESAR, P., AND BULTERMAN, D. C. A. 2012. Let me comment on your video: Supporting personalized end-user comments within third-party online videos. In *Proceedings of the WebMedia*. 253–260.

HARDMAN, L., VAN ROSSUM, G., AND BULTERMAN, D. C. A. 1993. Structured multimedia authoring. In *Proceedings of the ACM International Conference on Multimedia*. 283–289.

IBANEZ, J., AYLETT, R., DELGADO-MATA, C., AND MOLINUEVO, E. 2009. On the implications of the virtual storyteller's point of view. *Knowl. Eng. Rev. 23*, 4, 339–367.

KENNEDY, L. S. AND NAAMAN, M. 2009. Less talk, more rock: automated organization of community-contributed collections of concert videos. In *Proceedings of the International Conference on WWW*. 311–320.

KIRK, D., SELLEN, A., HARPER, R., AND WOOD, K. 2007. Understanding videowork. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 61–70.

KRUITBOSCH, G. AND NACK, F. 2008. Broadcast yourself on YouTube: Really? In *Proceedings of the 3rd ACM International Workshop on Human-Centered Computing (HCC'08)*. ACM, New York, NY, 7–10.

LIENHART, R. 1999. Abstracting home video automatically. In *Proceedings of the ACM International Conference on Multimedia*. 37–40.

NACI, S. U. AND HANJALIC, A. 2007. Intelligent browsing of concert videos. In *Proceedings of the ACM International Conference on Multimedia*. 150–151.

PEA, R., MILLS, M., ROSEN, J., DAUBER, K., EFFELSBERG, W., AND HOFFERT, E. 2004. The diver project: Interactive digital video repurposing. *IEEE Multimedia 11*, 1, 54–61.

PIACENZA, A., GUERRINI, F., ADAMI, N., LEONARDI, R., PORTEOUS, J., TEUTENBERG, J., AND CAVAZZA, M. 2011. Generating story variants with constrained video recombination. In *Proceedings of the ACM International Conference on Multimedia*. 223–232.

PURCELL, K. 2010. The State of Online Video, The Pew Research Center's Internet & American Life Project. http://www.pewinternet.org/Reports/2010/State-of-Online-Video.aspx. (Last accessed June 2010).

ROWE, L. A. AND JAIN, R. 2005. ACM SIGMM retreat report on future directions in multimedia research. *ACM Trans. Multimedia Comput. Commun. Appl. 1*, 1, 3–13.

SCHMANDT, C. 1993. Phoneshell: The telephone as computer terminal. In *Proceedings of the ACM International Conference on Multimedia*. 373–382.

SHAMMA, D. A., SHAW, R., SHAFTON, P. L., AND LIU, Y. 2007. Watch what I watch: using community activity to understand content. In *Proceedings of the International Workshop on Multimedia Information Retrieval*. 275–284.

SHIPMAN, F., GIRGENSOHN, A., AND WILCOX, L. 2008. Authoring, viewing, and generating hypervideo: An overview of Hyper-Hitchcock. *ACM Trans. Multimedia Comput. Commun. Appl. 5*, 2, Article 15.

SHRESTHA, P., DE WITH, P. H. N., WEDA, H., BARBIERI, M., AND AARTS, E. H. L. 2010. Automatic mashup generation from multiple-camera concert recordings. In *Proceedings of the ACM International Conference on Multimedia*. 541–550.

SINGH, V. K., LUO, J., JOSHI, D., LEI, P., DAS, M., AND STUBLER, P. 2011. Reliving on demand: A total viewer experience. In *Proceedings of the ACM International Conference on Multimedia*. 333–342.

SNOEK, C., FREIBURG, B., OOMEN, J., AND ORDELMAN, R. 2010. Crowdsourcing rock n' roll multimedia retrieval. In *Proceedings of the ACM International Conference on Multimedia*. 1535–1538.

SOARES, L. F. G., MORENO, M. F., NETO, C. S. S., AND MORENO, M. F. 2010. Ginga-NCL: Declarative middleware for multimedia IPTV services. *Commun. Mag. 48*, 6, 74–81.

STOCKHAMMER, T. 2011. Dynamic adaptive streaming over HTTP-Standards and design principles. In *Proceedings of the Multimedia Systems Conference (MMsys'11)*. 133–143.

SU, K., NAAMAN, M., GURJAR, A., PATEL, M., AND ELLIS, D. P. W. 2012. Making a scene: Alignment of complete sets of clips based on pairwise audio match. In *Proceedings of the ACM International Conference on Multimedia Retrieval*. Article 26.

SUNDARAM, H. AND CHANG, S.-F. 2000. Determining computable scenes in films and their structures using audio-visual memory models. In *Proceedings of the ACM International Conference on Multimedia*. 95–104.

URSU, M. F., THOMAS, M., KEGEL, I., WILLIAMS, D., TUOMOLA, M., LINDSTEDT, I., WRIGHT, T., LEURDIJK, A., ZSOMBORI, V., SUSSNER, J., MYRESTAM, U., AND HALL, N. 2008. Interactive TV narratives: Opportunities, progress, and challenges. *ACM Trans. Multimedia Comput. Commun. Appl. 4*, 4, Article 25.

XIE, L., NATSEV, A., KENDER, J. R., HILL, M., AND SMITH, J. R. 2011. Visual memes in social media: Tracking real-world news in Youtube videos. In *Proceedings of the ACM International Conference on Multimedia*. 53–62.

ZSOMBORI, V., FRANTZIS, M., GUIMARAES, R. L., URSU, M. F., CESAR, P., KEGEL, I., CRAIGIE, R., AND BULTERMAN, D. C. A. 2011. Automatic generation of video narratives from shared UGC. In *Proceedings of the ACM Conference on Hypertext and Hypermedia*. 325–334.

Received October 2012; revised April 2013; accepted July 2013