Multimedia Systems (2007) 12:423–438
DOI 10.1007/s00530-006-0065-6

REGULAR PAPER

User-centered control *within* multimedia presentations

Dick C. A. Bulterman

Published online: 15 November 2006

Abstract The focus of much of the research on providing user-centered control of multimedia has been on the definition of models and (meta-data) descriptions that assist in locating or recommending media objects. While this can provide a more efficient means of selecting content, it provides little extra control for users once that content is rendered. In this article, we consider various means for supporting user-centered control of media within a collection of objects that are structured into a multimedia presentation. We begin with an examination of the constraints of user-centered control based on the characteristics of multimedia applications and the media processing pipeline. We then define four classes of control that can enable a more user-centric manipulation within media content. Each of these control classes is illustrated in terms of a common news viewing system. We continue with reflections on the impact of these control classes on the development of multimedia languages, rendering infrastructures and authoring systems. We conclude with a discussion of our plans for infrastructure support for user-centered multimedia control.

1 Introduction

At first glance, the notion of *user-centered multimedia* seems to be a tautology. Multimedia objects are created, encoded, transported, decoded and rendered for one reason: to meet the needs of the user experiencing that media. Without a user (i.e., a viewer, reader or listener), multimedia is useless. It is difficult to imagine a more user-centered enterprise within the realm of information systems. Upon closer examination, however, it is clear that the user has played anything but a central role during the first decade of wide-scale multimedia systems deployment.

Rather than seeing the user as an active, controlling element in the process of presentation creation, selection and rendering, nearly all current commercial and research media players see the user as a passive element with no more influence over content than has been available since the advent of general radio broadcasts in about 1925. At that time, the "user" had the following options to control content on a rendering device:

- to turn the rendering device on (assuming it was off);
- to tune in one of the multiple content streams, based on either random selection or the use of a published program guide;
- to replace one tuned content stream with another;
- to turn off the rendering device (assuming it was on).

Missing from this list is an ability to directly influence the content within a stream or to support content-based navigation. We call this the "tune-and-play" control paradigm.

Not much has changed in this age of networked digital media. While the types of media available to a user have expanded to include video, animation and streaming text feeds, the amount of runtime control over content remains minimal. This is not to say that there has been no progress: the formalization of meta-data descriptions

D. C. A. Bulterman (✉)
CWI: Centrum voor Wiskunde en Informatica, Amsterdam,
The Netherlands
e-mail: Dick.Bulterman@cwi.nl

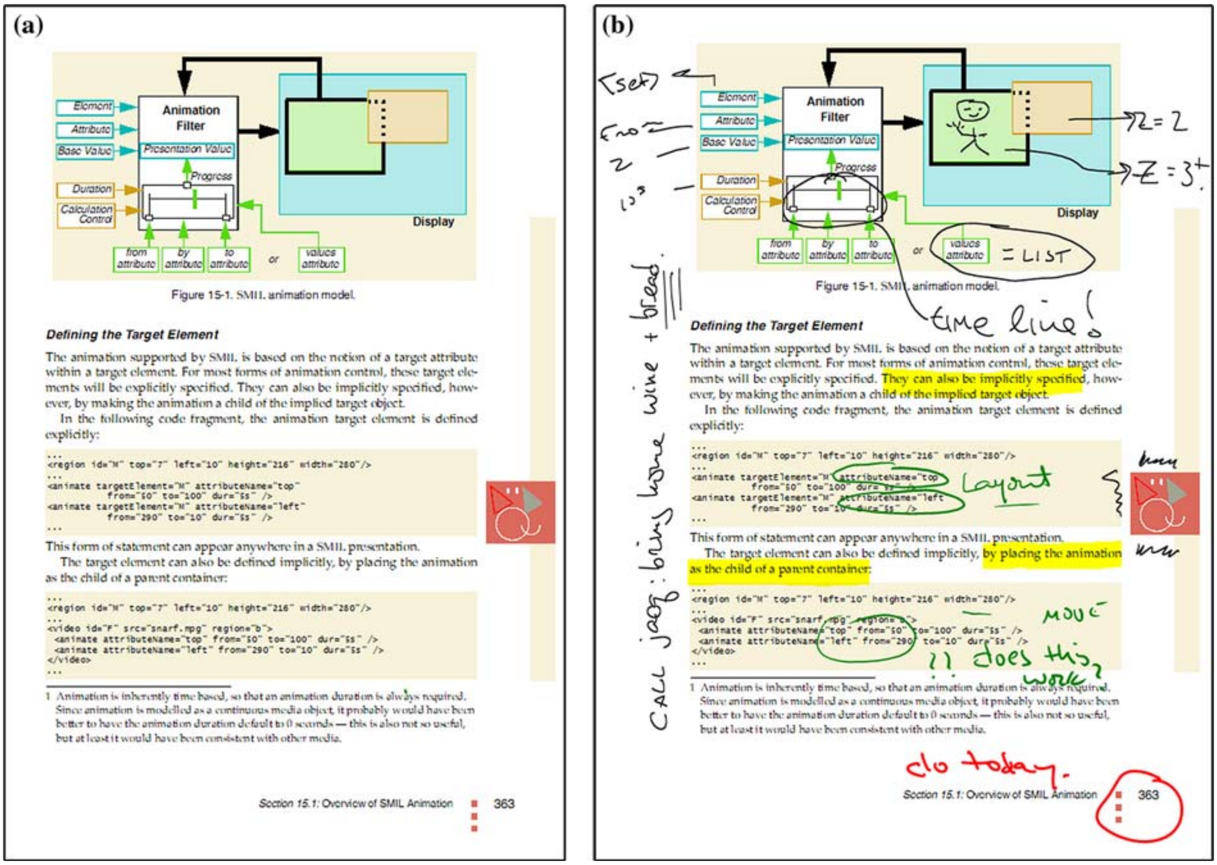


Figure 15-1. SMIL animation model.

Defining the Target Element

The animation supported by SMIL is based on the notion of a target attribute within a target element. For most forms of animation control, these target elements will be explicitly specified. They can also be implicitly specified, however, by making the animation a child of the implied target object.

In the following code fragment, the animation target element is defined explicitly:

```
...
<region id="M" top="7" left="10" height="216" width="280"/>
...
<animate targetElement="M" attributeName="top"
         from="50" to="100" dur="5s" />
<animate targetElement="M" attributeName="left"
         from="290" to="10" dur="5s" />
...
```

This form of statement can appear anywhere in a SMIL presentation.

The target element can also be defined implicitly, by placing the animation as the child of a parent container:

```
...
<region id="M" top="7" left="10" height="216" width="280"/>
...
<video id="F" src="snarf.mpg" region="b">
 <animate attributeName="top" from="50" to="100" dur="5s" />
 <animate attributeName="left" from="290" to="10" dur="5s" />
</video>
...
```

[1] Animation is inherently time based, so that an animation duration is always required. Since animation is modelled as a continuous media object, it probably would have been better to have the animation duration default to 0 seconds — this is also not so useful, but at least it would have been consistent with other media.

Section 15.1: Overview of SMIL Animation 363

Figure 15-1. SMIL animation model.

Defining the Target Element

The animation supported by SMIL is based on the notion of a target attribute within a target element. For most forms of animation control, these target elements will be explicitly specified. They can also be implicitly specified, however, by making the animation a child of the implied target object.

In the following code fragment, the animation target element is defined explicitly:

...
<region id="M" top="7" left="10" height="216" width="280"/>
...
<animate targetElement="M" attributeName="top"
         from="50" to="100" dur="5s" />
<animate targetElement="M" attributeName="left"
         from="290" to="10" dur="5s" />
...

This form of statement can appear anywhere in a SMIL presentation.

The target element can also be defined implicitly, by placing the animation as the child of a parent container:

...
<region id="M" top="7" left="10" height="216" width="280"/>
...
<video id="F" src="snarf.mpg" region="b">
 <animate attributeName="top" from="50" to="100" dur="5s" />
 <animate attributeName="left" from="290" to="10" dur="5s" />
</video>
...

1 Animation is inherently time based, so that an animation duration is always required. Since animation is modelled as a continuous media object, it probably would have been better to have the animation duration default to 0 seconds — this is also not so useful, but at least it would have been consistent with other media.

Section 15.1: Overview of SMIL Animation 363

<set>

2

Z = 2

Z = 3+.

= LIST

time line!

CALL jaap: bring home wine + bread.

Layout

move

?? does this work?

do today.

Fig. 1 The author's (**a**) and user's (**b**) views of a textbook (content courtesy Springer-Verlag)

and the advent of disk-based media players have added the following rendering control options to the list above:

- to select content streams via an electronic program guide integrated into the media player interface,
- to pause a content stream, allowing it to be buffered locally and then restarted under user control,
- to seek within a content stream using a timeline interface that is based on the granularity of the source object's encoding.

Even with these new convenience features, however, users have control functionality that is essentially equivalent to that offered magnetic tape users in the 1950s. While meta-data enhancements have made it possible to obtain a significant amount of information *about* a media object, it remains nearly impossible for a non-technical user to exercise anything but the most basic kind of control *over* that object during rendering.

This article considers various aspects of increasing user control over the media presentation experience. We are inspired, in part, not by new technology, but by control operations that have stood the test of time. This is illustrated by the page from a multimedia textbook [10] shown in Fig. 1. In Fig. 1a, we see the content as the author intended: a collection of in-line and out-of-line body text, images and code fragments, all carefully packaged to provide a coherent message. Figure 1b illustrates a more user-centered view of the same page: here, user-centered content additions have been provided in the form of margin notes, links to related content and user-directed navigation information. These additions allow a generic work to be transformed into a personal one that meets the particular needs of a content consumer. Text documents have supported this type of user-centered control for decades, including meta-notions of the management and redistribution of copyright-protected material. The disadvantages of augmenting paper publications include the destructive nature of the augmentations and a general inability to effectively manage the needs of multiple users of a single document. Although these disadvantages are substantial, they are more related to the characteristics of the publication medium — paper — than to more abstract constraints of media transfer. One goal, then, of our work has been to combine the best characteristics of paper documents with the substantial advantages in presentation flexibility of rich media electronic presentations.

The sections below discuss four approaches to increase the amount of user-centered control available when viewing and navigating through content *within* multimedia presentations. We present each approach in the context of a common application example and then reflect on the impact of the four approaches on the major phases of the digital media life cycle: during content creation/authoring, content transport and delivery, and actual rendering of content on a local device once it has been selected. Section 2 provides an overview of the environmental aspects that determine the scope of user-centered multimedia control, including the characteristics of multimedia presentations and the multimedia processing pipeline. It concludes with the definition of four classes of user-centered control that can reasonably be expected to be supported in modern multimedia systems. Section 3 provides a description and examples of each of the four control classes, drawing on the common thread of an interactive news example. It also provides a set of reflections on the implications of the four control classes on the design of multimedia languages, multimedia presentation infrastructures and multimedia authoring systems. Section 4 reviews a set of implications for the user interface when user-centered control is provided. Section 5 ends with conclusions of our experiences with user-centered media control and provides an overview of our current research agenda to enhance the control capability of non-technical end users.

2 Environmental aspects of user-centered multimedia control

This section considers the characteristics (and constraints) of practical multimedia environments, since these ultimately determine the types of control classes that are available to media users.

This article uses a newscast presentation as a common content paradigm. One frame from a 40-minute newscast is shown in Fig. 2. Television news is a highly studied domain of multimedia presentations [14,16,21,24,25,29,30]. One of the reasons that news is popular is that it is universally familiar: it represents, even more so than TV sports, a class of media content that appears in all cultures across the globe.[1] Other reasons that news is heavily studied include the fact that news programs are highly structured and that they follow strict formatting guidelines (at least within a family of newscasts). This structure is helpful in segmenting a news program into components [28].

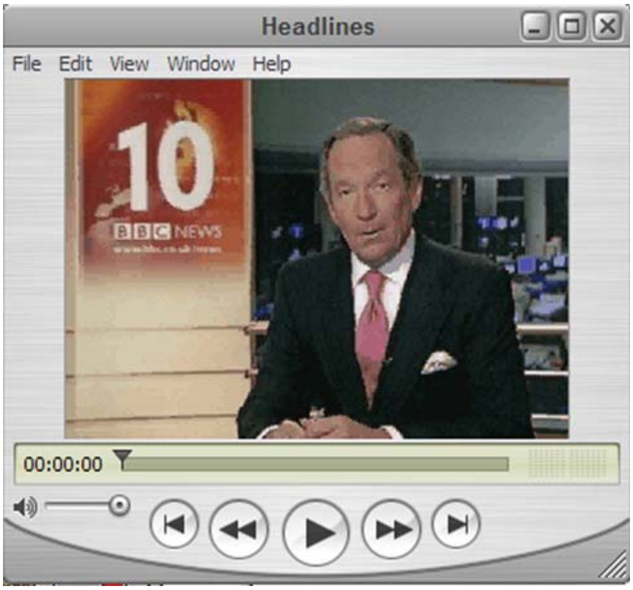


Fig. 2 Fragment from a 40-min newscast (content courtesy BBC News)

2.1 Control points within the multimedia processing pipeline

Before defining individual actions to allow user-centered control, it is useful to define a set of control points within the multimedia pipeline that can be influenced by end-user actions. A typical processing pipeline for digital content is shown in Fig. 3.

The source content is available as input into the pipeline. This input must be encoded in one or more container formats (such as DIRAC [4], Ogg Vorbis [44], Timed Text [37] and SVG [23])[2] and stored in a media repository system (which can either be an informal network-accessible file system or a multimedia database system such as Monet [26]). The content may be made available as a single composite media object or as a discrete collection of individual objects. The content may also be semantically modelled, so that the tuning portion of the "tune-and-play" paradigm can be supported in a personalized manner [3].

For content that is packaged as a single media object — such as MPEG-4 [27] — the entire presentation is

[1] This is not to say that all news programs use the same format or follow the same conventions: the cultural variations can be considerable, although this is more an issue for content analysis than content control.

[2] The references in this section cite some interesting non-proprietary formats. They are not intended to be complete.

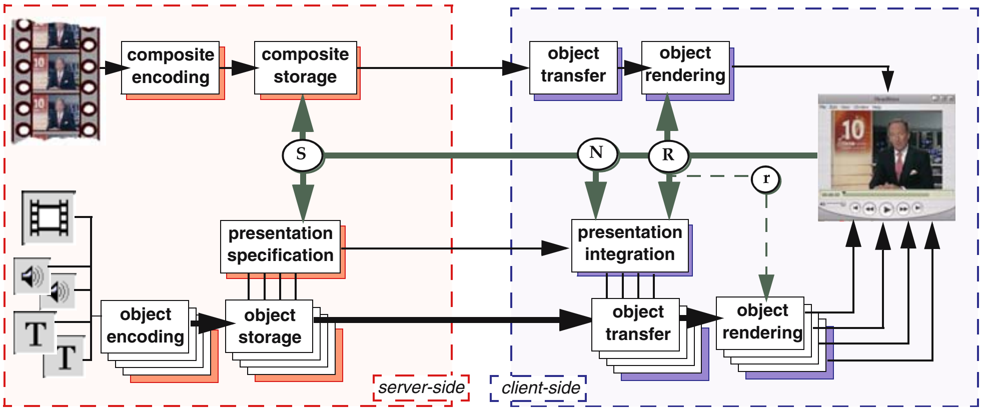


Fig. 3 Media processing pipeline and user control points for an on-line version of a newscast

stored in a composite 'sealed container': all of the content is accessed via a single transfer control stream. This is illustrated in the upper portion of Fig. 3. The apparent simplicity of using a composite object has many advantages, not the least of which is control over the content composition by the presentation author. It comes at a cost, however, in that the amount of user-centered control over the presentation at access time will be limited.

It is also possible to define and deliver a presentation as a set of discrete media objects — either because multiple media components within a story are independently managed, or because a presentation is structured as a collection of related independent stories. This is illustrated in the lower portion of Fig. 3. In this case, a separate phase of presentation specification and integration is required using a language that makes any spatial and temporal relationships among objects explicit, such as SMIL [10,11]. It can also define conditions under which objects participate in a presentation, as well as any rules for content dispatching (including cost functions and relative access restrictions), and the capabilities for event-based or link-based interaction within the presentation. The development of the presentation container model can be done manually, using an authoring tool [6,32,42] or automatically [2,25].

In our segmentation of composite-object versus discrete object presentation structuring, we make a distinction between selection and interaction functionality sealed within a composite container (such as an MPEG-4 encoding) and the abstracted collection of presentation relationships defined by a language such as SMIL. While both can provide the basis for conditional content and extensive user interaction, the options within the MPEG-4 file are all defined and controlled by the content creator, while a SMIL specification can provide a set of alternatives — including dynamic presentation extension — that can be interactively determined during presentation rendering. In this sense, we feel that the encodings like MPEG-4 control the user, while encodings like SMIL provide a potential for the user to control the content.

When the presentation (in what ever form) becomes active, actual media access, transfer and rendering takes place. For a composite-object presentation, only one object needs to be located and delivered via a rendering client. In a discrete-object presentation, multiple objects may be stored at several locations in a distributed environment; each must be located and synchronized based on the presentation specification using an integrating media player (such as Ambulant [1] for SMIL) and then rendered via the user's client. The media object access can be facilitated by a streaming media infrastructure built on top of RTP/RTSP [33,34].

Each of the pipelines shown in Fig. 3 are influenced by a series of user control points, which are abstracted onto the center presentation control bar. In an ideal world, a user could have influence over the entire media processing pipeline, but for many reasons, this is not practical in a real world of limited resources and legal constraints on media use and redistribution. We identify three user-centered control points:

- ***S***: This is the *presentation selection control point*. For the composite object case, a single media item will be identified at a media server. For the discrete object case, the initial selection abstraction is the integrating presentation specification.
- ***R***: This is the *rendering control point*; it supports the virtual video tape controls of start/stop/pause and fast-forward/reverse. For the composite object case, a single presentation timeline is manipulated. For the discrete object presentation case, it is possible to manipulate both the composite presentation timeline and the individual object timelines (denoted by

control point ***r***); from a practical perspective, the manipulation of multiple timelines usually is associated with the display of optional content streams that are played together with the primary set of content elements.

- ***N***: This is the *logical navigation control point*. Since the presentation specification for the discrete object case contains a logical sub-structuring, it is possible to use this structure to guide presentation navigation. This logical structure is usually defined during presentation authoring but it can also be generated dynamically [20].

In general, the further upstream the definition of presentation relationships is placed from the user, the fewer user-centered control operations will be available. As with broadcast media, a user has little influence over the objects at the ***S*** control point and has only limited legal abilities to subset, recombine or alter the content within a given media encoding. At the ***R*** control point, actual rendering takes place, which can easily be directly influenced by a user. At the ***N*** control point, logical relationships among the base media are exposed, but (more importantly) a given media player may also provide interaction facilities to allow user-centered control of the packaged presentation. As a result, we feel that this is the most fruitful area for studying user-centered — rather than producer-centered — control. This control point will be the focus of the remainder of this article.

2.2 Presentation selection versus intra-presentation control

Historically, the most fundamental aspect of media control given to a user in the "tune-and-play" paradigm is that of presentation selection. One of the focal points of current media selection research is the labelling of content to assist in the automatic selection of media objects on a user's behalf [19,35,43]. These labels can be gathered in a standard format, such as TV-Anytime [36], and then either presented to the user via an electronic program guide or filtered by a user agent [3,13]. The result of the selection process can either be a pre-produced presentation or a custom presentation created on demand based on a manipulation of various selection criteria.

While there are many important advances to be made in the area of content labelling and selection, this work has only a limited impact on the control given to a user *during* a presentation. While the automated selection of content provides an important high-level navigation resource that assists users in resolving content alternatives, it does little to improve the actual viewing experience for any given media object once that content is actually rendered. Therefore, instead of targeting object selection, our work focuses on expanding the intra-presentation control options given a user during the viewing/listening/reading experience. In particular, we are interested in extending the "tune-and-play" user control paradigm that has dominated media consumption since the days of the crystal tube.

2.3 Classes of intra-presentation control

The study of the role of the user in general information systems has long been the province of the human–computer interaction community. A portion of this work has concentrated on developing a model of the 'user' and then defining classes of interactions in terms of this model [17] that allow user wishes to be fulfilled. It is tempting to adopt a similar approach for studying user-centered multimedia — that is, to define a model of the user from several perspectives, including that of content creation, content selection and content rendering, and to then define a set of media operators based on this model — but doing so would simply lead to new methods to automate the "tune-and-play" paradigm. Instead, we see the challenge of user-centered multimedia to transform the user from a passive element in the media pipeline to an active element that can influence control over how and when media objects are accessed, composed, shared and rendered within the boundaries of the legal constraints on media manipulation. In this respect, the key to user-centered multimedia is not the construction of a model of the user, but the construction of an effective model of media content control.

In our study of user-centered multimedia control, we define four control classes that a user can manipulate while interacting within a presentation that extend beyond the existing "tune-and-play" paradigm:

- *Intra-presentation personalization*: Control over media sub-selection within a presentation, where the user is able to select a subset of content to meet the constraints of a particular usage context.
- *Intra-presentation navigation*: Control over media activation and sequencing within a presentation, where the user is able to access the content within a presentation based on logical content-based navigation instead of only timeline-based seeking.
- *Intra-presentation augmentation*: Control over the enrichment of a given presentation with personal extensions such as margin notes and pointers to related content.
- *Intra-presentation archiving and sharing*: Control over personal archiving and sharing of content

within a peer group, including the selection of subparts of a presentation and the access control information within the peer group.

Each of these classes is defined in terms of a media *presentation* rather than a single media *object*. This is not intended to limit the generality of the proposed user-centered control since, in the trivial case, a presentation may consist of a single media object.

2.4 Legal constraints on user-centered control

We conclude this section with a short consideration of the legal constraints imposed on the manipulation of third-party media objects. This is not a gratuitous gesture: it is a recognition of a fundamental restriction on how media can be processed and manipulated. The copyright owner for a piece of content can dictate how that content is to be used. User-centered approaches that assume that a particular content encoding can be altered — either by editing or augmentation — are, at best, self-limiting and, at worst, illegal. Throughout any study of the user's influence on media, each media object needs to be considered to be an atomic object than cannot be altered. While there are also instances where users are allowed to alter objects that they do not own, this is a limited special case.

3 Intra-presentation user-centered control

This section describes multiple aspects of supporting the four classes of intra-presentation user-centered media control defined in Sect. 2.3. We present this discussion in the context of manipulating the newscast abstraction defined at the start of Sect. 2.

3.1 Intra-presentation personalization

Intra-presentation personalization allows conditional content to be controlled during rendering based on a particular user. Three examples of this control are: the conditional selection of optional content streams (such as open/closed captions [31]), the selection of individual natural language encodings for content streams (such as selecting Dutch or English for audio and text in a presentation) and the selection of semantic level of content within a presentation (such as a highly-detailed content stream versus a stream with summary information [2]). Such personalization is often not static: a given user may have many different usage contexts, depending on the particular device used or the personal circumstances that exist when the presentation was activated. As a result, any given presentation can be personalized in multiple ways for the same user, depending on the access context.

Consider the example shown in Fig. 4. Here we see the news story presented earlier in Fig. 2. In Fig. 4a, the content encoding is shown as a monolithic object: the video data, the audio data and the captions information are packaged inside a single unit. The amount of selectivity in such an object is minimal; the sub-streams are intended to be presented together. An alternative to the virtual videotape model of Fig. 4a is shown in Fig. 4b. Here, each of the main media components are defined (and sent) as separate objects. This allows individual sub-stream selection — such as by a blind user, who may not need video. (It also allows a quality of service infrastructure to perform resource management when multiple, scaled encodings of objects are available, but this is not a user-centered issue). The utility of the independent sub-stream approach is further shown in Fig. 4c; here, multiple versions of audio and text captions — perhaps in different languages or at different semantic levels of detail — are referenced. The user can control

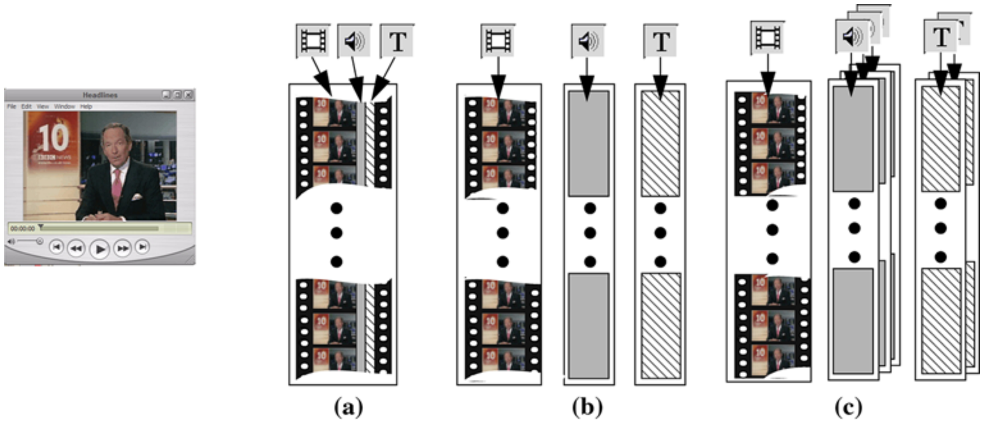


Fig. 4 Improving intra-presentation selectivity by de-coupling media components

which level is of interest, or which encoding meets his or her needs. It is important to note that the level of control offered is not restricted to only selecting content steams provided by the original content creator. Additional pointers to content can be added to an existing file — if an intermediate packaging format is used — or, as discussed below, a user could choose to add his or her own content extensions.

The use of multiple independent objects to represent a composite media item was first advocated in the GR*i*NS *channel* model in our early work on presentation structuring [5]. As shown in Fig. 5, the channel model defined a presentation in terms of a collection of independent media objects, each of which could be selected (or deselected) by the user at presentation time. This approach presents a multimedia author with an extra structuring and coordination burden, although several systems exist that support such compositing at the language level (such as SMIL) and at the tools level (such as GR*i*NS [6], HyperProp [32] or LimSee [42]). Some of the concepts have also been integrated into user-level control of multi-channel formats, such as DVD disks, where multiple language subtitle tracks can accompany a single audio/video object. And, as we will consider in Sect. 3.4, the utility of a multi-channel approach, when combined with the concept of user-centric layering of content, extends beyond user-centered content selection to user-centered content augmentation and sharing as well.

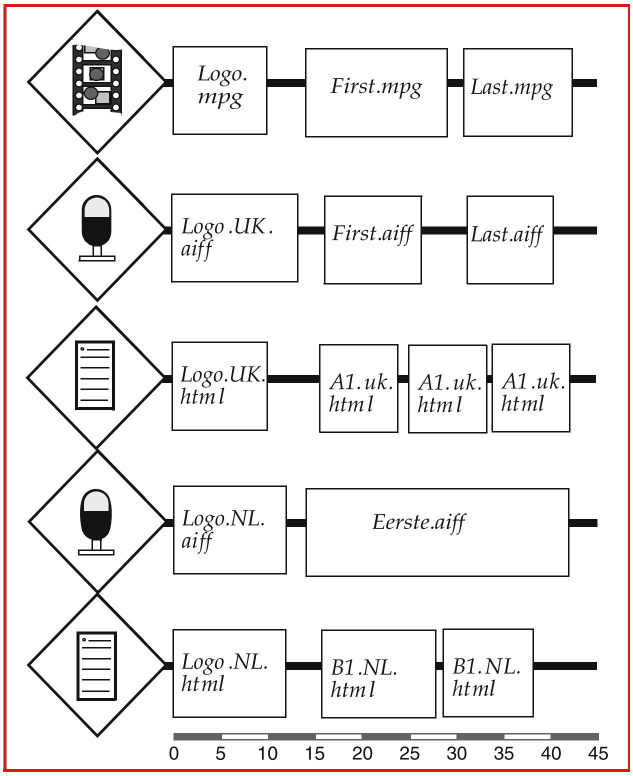
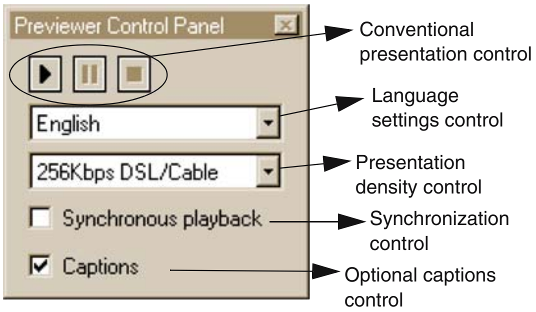


Fig. 5 Presentation fragment containing multiple logical resource channels (from [5])

There has been a trend within industry and standardization groups to provide support for intra-presentation selection only in special versions of content or with special tools developed for specific target groups, such as talking book readers for the blind [15] or captioned players for the deaf. Unfortunately, such ‘special’ treatment brings with it ‘special’ costs, effectively creating a barrier for supporting a ‘special’ market segment. It is important to realize, however, that support for intra-presentation selection does not only serve the needs of a special group, such as the blind or deaf. There are many instances when, out of consideration to office-mates or in the noisy context of the factory floor, a substitute encoding for a video or audio stream could be useful. (One example of this is watching a network newscast in a noisy airport departure lounge.)

The use of a channel-like approach to structure media for user-selection exposes a wide range of user control options, but it also requires that a user understands the options available for exercising that control. Figure 6 shows the dynamic control panel used in the GR*i*NS player to expose the content options within a presentation. In this illustration, a user is presented with the ability to perform conventional start–stop–pause control operations, but also to control language, presentation density, synchronization and captions rendering. For some use cases — and for some users — such control operations could be defined in profiles, so that a consistent selection is always made, but other users may want to have the control options made explicit by having them included in the control panel. In the case of the GR*i*NS implementation, the options presented in the control panel are extracted dynamically from the source SMIL file and a custom control panel is created for each presentation.


Previewer Control Panel

English

256Kbps DSL/Cable

☐ Synchronous playback

☑ Captions

Fig. 6 GR*i*NS dynamic control panel for exposing multiple content streams

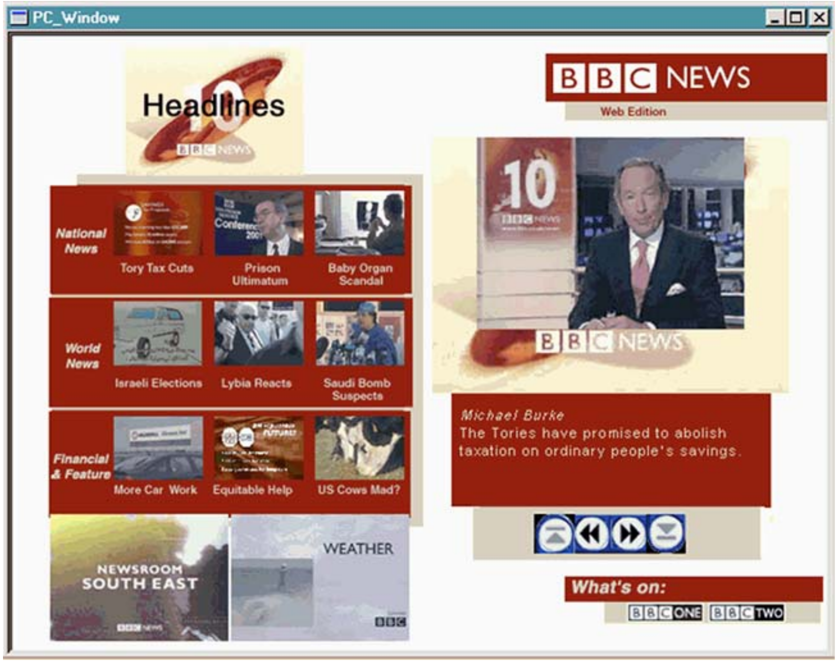
The generation of custom intra-presentation control illustrates the advantage of using an intermediation integration format such as SMIL over an embedded control format such as MPEG-4. The SMIL file can be independently parsed before a presentation begins to extract control options to be presented to a user. These options, which may vary during the runtime of the presentation, are more difficult to extract when they are interlaced in the encoding of the actual media object. Another benefit of using an intermediate integration format is that it opens the possibility for adding custom content to a file (rather than simply selecting among the content streams in the original encoding). We consider this in Sect. 3.3

3.2 Intra-presentation navigation

Support for intra-presentation selection, as defined in Sect. 3.1, addresses the selection of a subset of media objects in a presentation by substituting one object for another. Another form of user-centered control is exposing logical relationships among diverse components within a presentation. We call this intra-presentation navigation. This form of navigation can be seen as a user-centered extension to the timeline-based navigation of content.

Consider the semi-automatic reformatting of the 40-minute newscast [7] as shown in Fig. 7. In this view, the individual stories in the newscast have been segmented and presented in menu form to the user. Rather than using a time-slider, the user can directly select stories from the menu board. The standard elements in each newscast are given fixed navigation blocks (such as the headlines, the weather and the local news). Individual stories within a newscast are partitioned into classes, as are links to associated programming content. There are also logical navigation buttons (below the captions block) that allow a user to jump to the next story within a group, or to different groups. The point of this prototype was not to define a definitive news presentation interface, but to explore various types of user control within an otherwise unstructured (from the user's perspective) news broadcast. Our goal was not to implement new story segmentation techniques, but to determine the types of navigation primitives that different kinds of users would need to manipulate a custom partitioning of content.

The navigation structure illustrated in Fig. 7 was determined from an analysis of the content at presentation authoring time. A visual content navigation map was added as a set of user-controlled events, which were then presented along with the based audio/video object. Although the navigation structure used in this example was created during the presentation authoring phase, it is logically separate from the actual news media objects. This means that an alternative navigation map can be created without changing the original content. Most monolithic media encoding formats do not provide the necessary support to allow this form of flexible user-centered navigation of content. Within text files, various XML formats [38–41] provide both internal and model-based hyperlink support and/or local flow-based event model specifications, but no direct support for temporal event processing. Among open web formats, SMIL provides the most unified and comprehensive set of linking



Michael Burke
The Tories have promised to abolish taxation on ordinary people's savings.

Fig. 7 News presentation exposing a logical navigation structure (image from a screenshot generated by the GR*i*NS player)

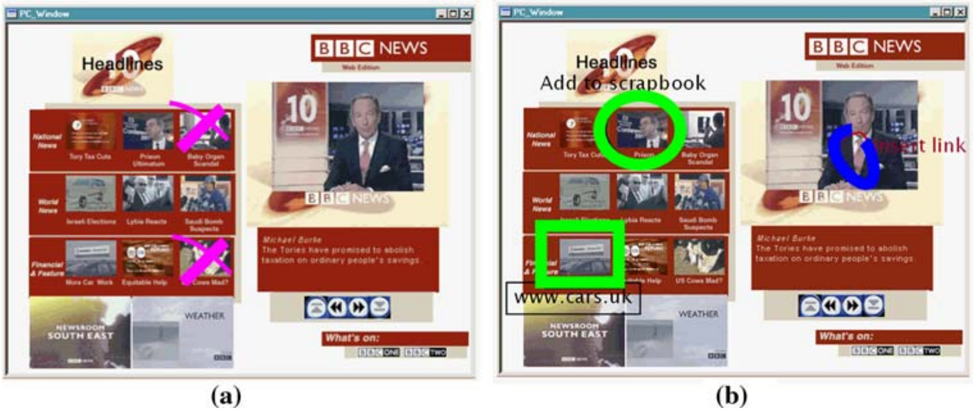
and event primitives that allow non-invasive support for logical control of navigation within a presentation. This means that the individual content objects do not need to be changed in order to support an external navigation structuring. From a legal perspective, this is an important advantage.

3.3 Intra-presentation augmentation

The facilities for intra-presentation selection and navigation discussed above enhance a user's ability to interact with media content based on a logical structuring and partitioning that is typically provided by an external party (from the user's perspective) during either base content creation/encoding or presentation authoring. The effective use of these two classes of control can significantly expand the possibilities for a user to gain increased control over the media viewing experience, but they remain limited to insights of external parties. A third class of control operations provide the user with the ability to augment the content or navigation structure by providing additional content or links into an external presentation.

Figure 8 illustrates two types of user-centered content augmentation that we have studied within the context of extended user control within a digital news viewer application. In Fig. 8a, a menu storyboard is presented that in most respects is identical to that illustrated in Fig. 7, with one exception: the user is able to perform supplemental edits on the collection of media objects available for view. In the example shown, a user (in our case, a parent) determines that a medical story about a doctor who was charged with performing unnecessary operations and subsequently stealing organs from babies would not be an appropriate content in a family in which one of the children was about to go to the doctor. In a similar manner, a story on the spread of disease among animals — and the decision to kill and incinerate them as a precautionary measure — was determined to be suitable only for the vegetarian members of the family. Note that unlike recommender systems that compile news programs from fragments, this system puts the end user in control of the editing operation. The key to the user interface in this instance was the fact that editing occurred as a control mode within a viewer application, rather than in a separate content management mode. In Fig. 8b, a similar transparent editing approach was used to allow end-user augmentation of content. Three activities are illustrated: first, a user is allowed to select individual parts of a newscast and save them in a clippings folder. Second, a user is given the opportunity to add links to associated content (in this case, an audio file that contains an embedded link to a web site on English cars). Third, a drawing tool is provided that allows a hot-spot to be created and associated with a link (in this case, the father inserts a link to a page that says: I would like this tie for my birthday!).

One approach to support content augmentation is to edit and extend a particular media object encoding, but this brings with it a host of rights management issues. Instead, we use an approach that creates a new intermediate file containing a set of user control layers that extends the control semantics of the presentation. Figure 9 illustrates this in the context of our news example. Each of the control operations illustrated in Fig. 8 are given a separate control layer in a document. (The top layer contains a link art and control information, while the layers below contain other content navigation and augmentation primitives.) In our SMIL base, this means that both a spatial and temporal alignment of the control operation is made with the base content. The control information can be further refined to contain access list options, so that they can be customized to apply to individual users or classes of user groups. (This



Michael Burke
The Tories have promised to abolish taxation on ordinary people's savings

Fig. 8 Examples of user-centered content augmentation and personalization

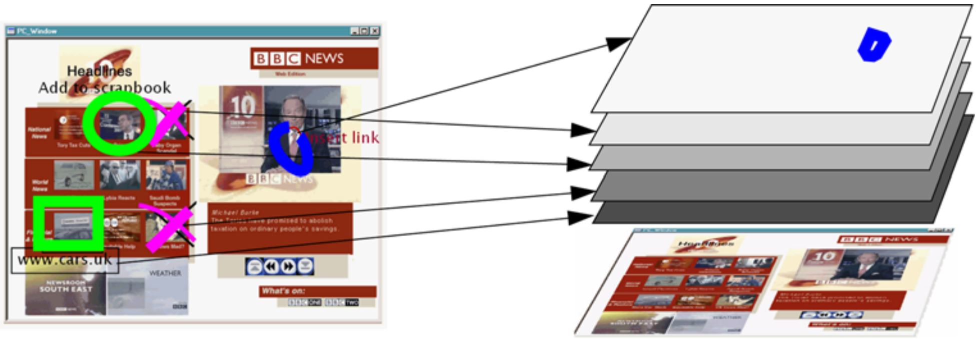


Fig. 9 Using a layered model for augmented user-centered control

latter aspect is important when considering augmented content sharing in Sect. 3.4.)

In addition to allowing control content to be added to a presentation, we have also experimented with adding new media content into a presentation (also within the context of a SMIL layer) by allowing a user to pause a base media object, add SVG-based line art and then provide movement and scaling to the base media. This was done in the context of a medical example [9], as shown in Fig. 10.

There are several important control aspects to the provision of intra-presentation augmentation. First, the user-created edits and content extensions do not occur automatically, but under the direct control of the content viewer. This control may be hierarchical (as in Fig. 8a, where a parent plays a coordinating role for managing in-coming media) or at a peer-level (as in Fig 8b, where multiple members of a family may independently augment content). This does not replace the functionality of recommender systems, but it adds a level of user fine-tuning locally. Second, the edits and augmentations do not need to change the actual base content: they may be stored as a series of content layers within (in our case) a SMIL presentation, differentiated by user. Third, the ability to subset a portion of the source media (as in the clippings example in Fig. 8b) does not require that the source media be altered or even stored locally. A logical pointer is used, along with a temporal clipBegin/clipEnd specification within a SMIL layer to provide an indirect editing reference.

All of these augmentations need to be managed by the user at the time the content is viewed. This implies the need for an integrated viewing/editing environment. Where much of our previous work has focused on providing dedicated authoring tools for content creation, our current activities are geared to understand the needs of incremental authoring support by non-expert users. We have termed this “couch-top authoring” [12]; it will be discussed further in Sect. 5.

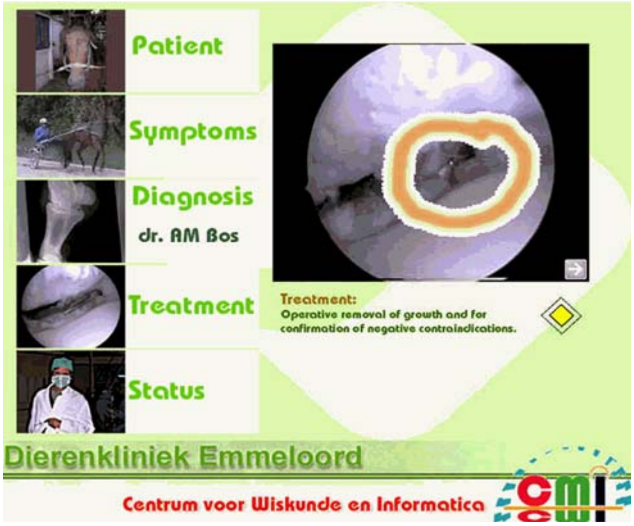

Patient
Symptoms
Diagnosis
dr. AM Bos
Treatment
Status
Treatment:
Operative removal of growth and for confirmation of negative contraindications.
Dierenkliniek Emmeloord
Centrum voor Wiskunde en Informatica
CWI

Fig. 10 Allowing users to add conditional line art to a medical dossier (from [9])

3.4 Intra-presentation archiving and sharing

The previous three sections illustrate how users can be given extended control over the media viewing experience, but a significant question of motivation can be posed: why would any user actually bother with this kind of control when viewing essentially transient media objects? In the most pessimistic case, a user would spend an entire day navigating, sub-setting, extending and augmenting a particular content object, only to have all of that work be nullified by the following evening’s newscast! Even in a less pessimistic case, it is clear that much video material is highly dated and has only a limited ‘shelf-life’. Still, we feel that there is significant motivation for including user-level control within a media viewer application.

One of the main motivations for empowering users to augment third-party content with extra control or incidental media augmentations is the explosive growth in third-party content via personal media repositories such as Flickr [18] or YouTube [45], or the (expected) explosive growth in personal and commercial media objects (such as podcasts or on-demand IPTV items). The goal of all of this content is to facilitate awareness by group sharing. Pointing is certainly one form of sharing, but subsetting, highlighting and augmenting provide compelling extensions that will enable the building of local or global communities based on shared content experiences. A fundamental challenge is to support these sharing operations within the context of legal access to media.

Consider the architecture illustrated in Fig. 11, which we have developed within the ITEA project Passepartout [22]. In this environment, a hybrid home-server and personal digital recorder (PDR) is defined that contains a substantial hard disk, a high-density optical storage device (in our case, using Blu-Ray technology) and a set of network connections to devices within and outside the home. The PDR serves as a content storage, management and sharing device, supporting three sets of external content connections:

- *Conventional Broadcast Content:* This is the primary interface for receiving broadcast content. It is expected that such content will be protected by some form of digital rights management (DRM) mechanisms.
- *P2P 'Family& Friends' Content:* This is a secondary content interface used for the sharing of content within restricted groups of external users. The group may consist of family members who access content from remote sites, or members of 'friends groups' who access comments based on some form of content notification protocol.
- *Home-based Content (Re)Distribution:* This is a local content distribution system based on fixed network or wireless technology that allows content to be viewed on various types of home devices. Some of these devices may be conventional TV's, while others may be low- or full-powered interactive viewing devices (such as TabletPC's).

One of the key aspects of the PDR is that it provides a storage and sharing mechanism which, in turn, brings with it an incentive for users to manipulate the content stored in the device. While it may, indeed, be unrealistic to expect users to invest in exercising control options over transient content, it is much more realistic to expect that users will invest in controlling content that they will personally archive (including broadcast, third-party content — including information on optical storage — and their own media objects) and then share within a formal or informal peer network.

The major incentive for studying this work in the broadcast context is the current shift in content distribution (and redistribution) from the conventional hierarchical model in Fig. 12a to the more peer-oriented model illustrated in Fig. 12b. The advent of an environment where consumers of media will also be transformed into producers and distributors of content provides a compelling reason to study increased user-centered control within these shared presentations in a manner that does not violate the content license of a particular media object encoding.

As with our detailed discussion of individual options for controlling local content flow, one of our primary concerns in developing user control models for the PDR is the management of access to media in light of rights

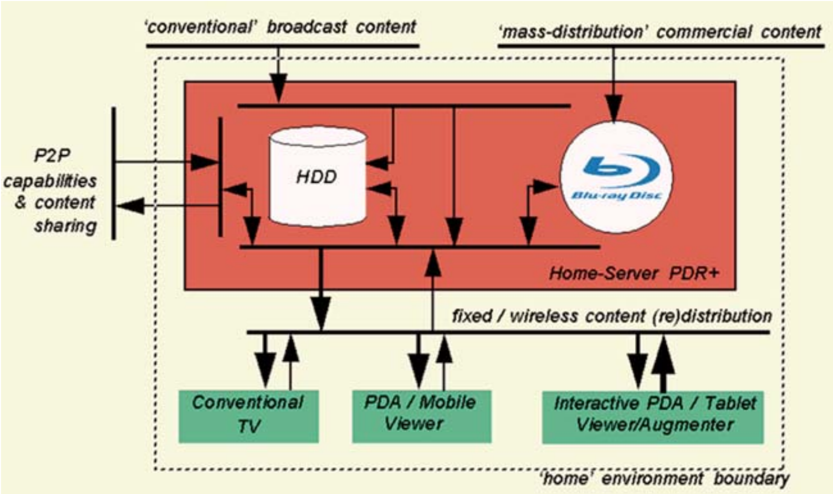

'conventional' broadcast content
'mass-distribution' commercial content
P2P capabilities & content sharing
HDD
Blu-ray Disc
Home-Server PDR+
fixed / wireless content (re)distribution
Conventional TV
PDA / Mobile Viewer
Interactive PDA / Tablet Viewer/Augmenter
'home' environment boundary

Fig. 11 A home network media server architecture (from [22])

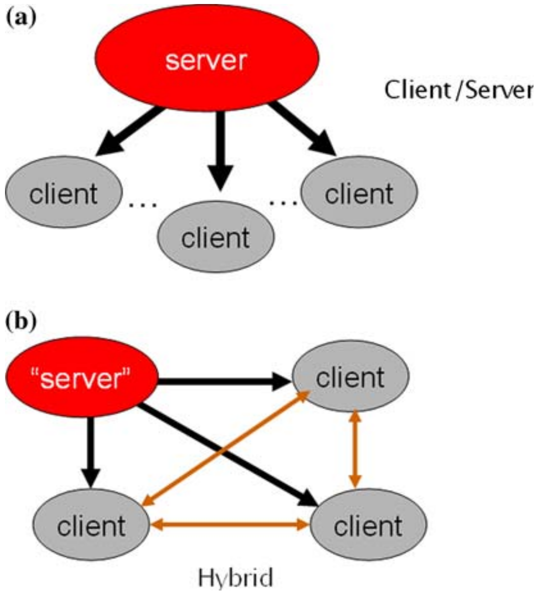


Fig. 12 Changes in information distribution infrastructures

management constraints. One the one hand, the ability to share media objects within a family or across a peer-to-peer (P2P) network provides a compelling incentive to invest in content control. On the other hand, it is unrealistic to expect that users will be allowed to subset and then share portions of restricted third-party content with peers on a P2P network. Our solution to this process is to place all control operations in an intermediate file (in our case, an extended SMIL file [8]), and to use this file as the basis for content sharing. If a reference is made within this file to protect contents (such as a commercial movie stored on a protected optical disk), it is the target partner's responsibility to obtain a (legal) copy of this content. The augmentations will remain valid because they are based on logical references rather than subsetted media objects.

4 User interface implications

During our experiments with user-centered control of multimedia, we have concentrated on various aspects of presentation authoring and language structuring. It is clear, however, that the efficient encoding and transfer of media does not in-and-of-itself guarantee that end-users will be motivated to become active media consumers. This section considers two media interface implications that have grown out of our work on media control systems: the architecture of the media player and the nature of content control devices. We continue to frame this discussion in terms of the news example defined in Sect. 2.

4.1 The architecture of the media player

While it is tempting to place the user at the center of the media delivery experinece, in all practical implementations of media delivery systems (from broadcast radio and TV, through the PC to personal media devices), it is media player that is actually the central point of control. As the physical integrating element of the media processing pipeline, the media player controls all aspects of actual media rendering, including providing a captive control architecture for the user.

Section 13 shows a historical progression of player-based media encapsulation for our news example. The media player — much more than the media content — defines a control abstraction that holds the user captive during the presentation. The control abstraction provides not only an operational context for delivering the media (at home on the couch, at work on the desk, on the tram in you hand), it also defines a bounding box outside of which control operations are impossible.

Consider the following scenario: you have five spare minutes before a meeting starts at work. To fill the time in a responsible manner, you tune in a personalized edition of last night's evening news via your desktop player. Your desktop software selects appropriate stories for you, and exposes all of the user centered intra-presentation control operations that we discussed in Sect. 3. After viewing three items, it is time to go to the meeting, which is in another building. You leave your desk, pull out your mobile telephone and once again ask to see the news presentation you were just watching. Once again, a personalized presentation is created, but you need to start at the beginning of the news and navigate to the point that you had been at on your desktop player. Not only will the control paradigms within each player be different, but the control scope over the presentation is also different, forcing the user to adapt to the player.

One of the current research goals of the Ambulant player [1] is to define a player control architecture that will allow a presentation to migrate with the user through various rendering contexts, from home through desktop to mobile systems. The focus of this work is making a particular presentation portable, instead of simply restartable. This work not only investigates the decentralized rendering of a single presentation at a macro scale (that is, in terms of high-level player contexts), it also investigates the further decentralization of the actual media player into a non-monolithic of cooperative rendering components, as is illustrated in Fig. 14. Here, a distributed set of rendering agents are dynamically associated and dispatched to render appropriate portions of a media presentation. In this example, the viewing goggles could be used for the video content,

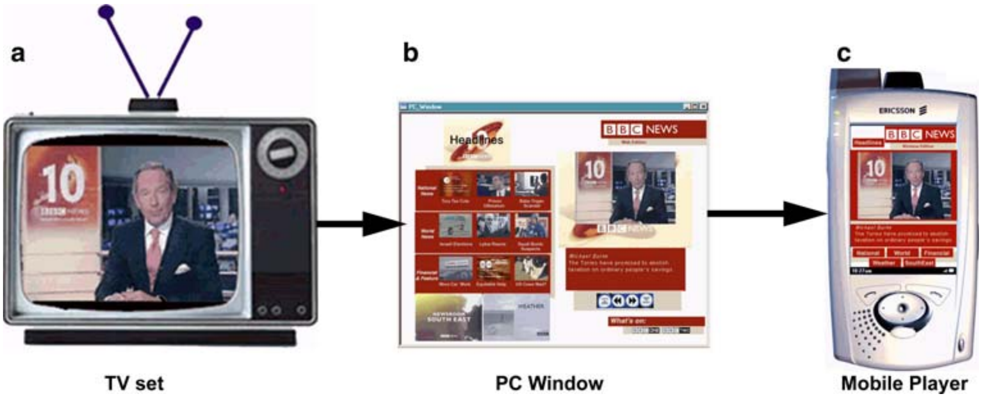


Fig. 13 Media player evolution for rendering the evening news

the personal media player could be used for audio, the PDA could be used for presentation storage and augmentation and the mobile telephone could be used for selectively updating contents or for synchronizing with the remote player controller. If a different set of devices were available, a different partitioning could take place. Note that if no viewing goggles were available, the user would essentially be blind; having a presentation object that supports accessible rendering would be very useful in this case.

The importance of presentation migration is not the development of multiple static views of a presentation (such as transforming a presentation using some sort of styling mechanism), but in supporting presentation migration at run-time. A given presentation — including access state, progressive content caching information and content-based logical (rather than time-code) navigation — would follow a user across various access contexts, rather than having each content define an separate, disconnected access instance. The core of our approach to support content migration is to define non-monolithic media players: players in which dynamic collection of renderers on scalable disconnected devices can be used to migrate content during its runtime.

4.2 Architecture of the control interface

The running example in this article has been a television news broadcast that is accessed and controlled via a desktop (and mobile) player. From a research perspective, it is interesting to note that the migration from TV to PC (illustrated in Fig. 13, 14) has come full-circle: after several false starts, it appears that a new focus on supporting broader interaction within a digital television environment is regaining interest.

The home television architecture typically consists of a television set and a remote control unit, as well as a collection of digital decoders and recorders.Within this context, part of our current work is investigating the integration of multiple scalable remote control devices to allow differentiated content delivery and differentiated content control to the living room. The operational assumption is that each user has his/her own personal remote control (rather than having all persons fight over one common remote). During common content viewing (such as an evening newscast), individual viewers can perform secondary content access operations — such as searching an electronic program guide or buying a necktie — without disturbing other viewers in the family. The personal remote control is also used to augment content during the news, or to simply read one's e-mail (depending on the control device).

At present, we support a controller hierarchy consisting of a conventional remote control device, a PDA, a Nokia 770 Internet Tablet and a pen-based TabletPC. Figure 15 illustrates two sets of control operations using the Nokia device. The player control function within the personal remote communicates with a media server that controls all of the remote devices in the home (See Fig. 11).

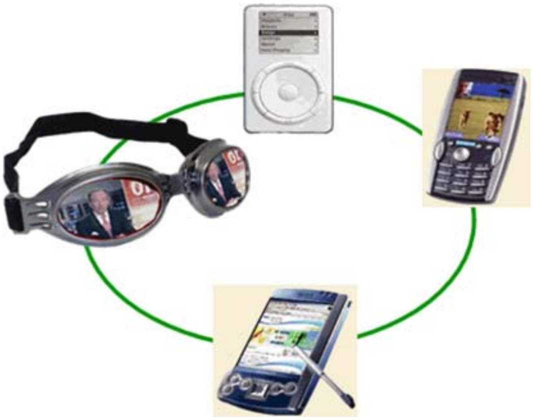


Fig. 14 A non-monolithic media player

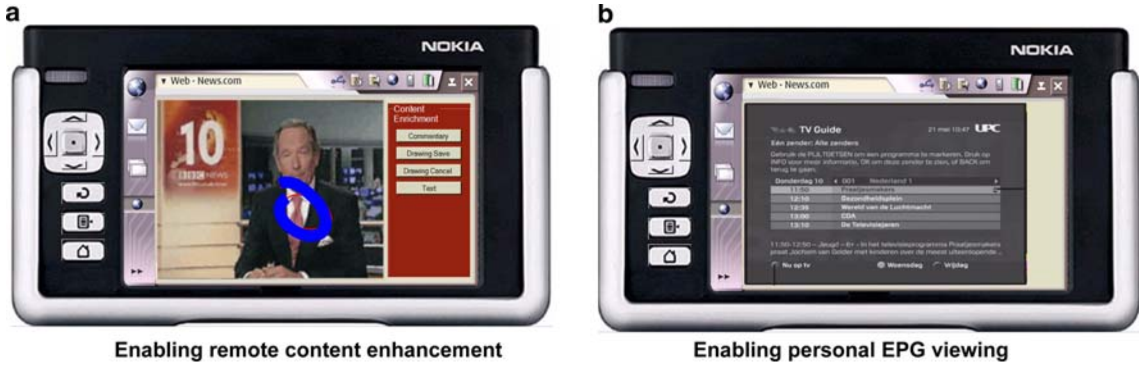


Fig. 15 Migrating TV control operations to a personal remote control device. **a** Enabling remote content enhancement. **b** Enabling personal EPG viewing

5 Conclusions and directions for future work

The key concept that has guided our work in the area of user-centered multimedia control is that all forms of digital media should be able to be manipulated as if they were conventional books, but with the additional facilities that currently only electronic media can support. This means that users should be allowed to insert logical bookmarks at user-determined points in the content, add augmented content to the base material (including ink and audio comments), attach links and pointers to related content in other presentations and to be able to archive and share content within the limits of the user rights of that material. They should also be able to extend the base content with full multimedia augmentations, including structured collections of audio, video and text objects that can be packaged for personal or group sharing.

There are several implications of this work that provide both constraints and opportunities for the future development of a rich, interactive control paradigm for digital content. In this section, we reflect on our experiences investigating user-centered control in the context of presentation languages, authoring tools and distribution infrastructures.

In terms of support for multimedia languages, we have been fortunate that many of the basic concepts that we have developed for user-centered control have been integrated into international standards. The most notable impact that we have been able to have has been on the various versions of the SMIL language and within the accessibility community. The GR*i*NS channel approach, with its integrated spatial, temporal and activation control model, forms the basis for the custom test attribute facility within SMIL, and has been adopted as a fundamental part of the Daisy consortium's format for digital talking books [15].

At the same time, however, it has been disappointing to note that there has been little exploitation of this facility beyond that used within specialized user communities. Part of this is probably a result of the difficulty of creating baseline multimedia presentations: many authors struggle to complete one version of a presentation and are not likely to expand their content to meet additional needs. Another part is the lack of user demand for differentiated content streams, and an associated lack of a commercial model to support this work. We are confident, however, that these are only transient limitations.

One of the means of encouraging greater user-centered control of content — especially for intra-presentation navigation and augmentation — is to better integrate presentation authoring with presentation viewing. This integration has, in our opinion, two important aspects. The first is the realization that content viewing and (incremental) content augmentation must be combined within a unified interface. The second is that the focus of presentation control needs to shift from a media-player-centric view to a media-experience-centric view: an active presentation is itself a logical entity (complete with a defined user context and operational control state) that should be able to migrate freely among a set of candidate rendering players without having to be restarted. In our view, the combination of these two aspects argues strongly for the development of a broad range of scalable pen-based devices (such as PDAs, mobile communicators and full-featured Tablet PCs), to serve as content augmentation devices, with the focus of the control resting with the person accessing the presentation rather than within the artificial constraints of a single physical media player.

Of course, all of the presentation and interface technology available cannot guarantee that a content producer (and an augmenting content consumer) will be motivated to construct open-ended rich media presentations. In this regard, the limiting factor is probably not so much technical as emotional and financial. Creating compelling media — whether in the large, such as a studio production, or in the small, such as a collection of personal images that are packaged for family/friends viewing — is a time consuming task that still requires a great deal of skill. Unless adequate compensation mechanisms are deployed (either financial or emotional), and unless adequate personal content protection systems are developed, an end user will most likely remain a passive element in the media pipeline. Here-in lies a seminal challenge for future research.

Acknowledgements This work was supported in part by Stichting NLnet under the AMBULANT-I and AMBULANT-II projects, by the ITEA Passepartout project, by the FP-6 project SPICE and by the BRICKS/PDC-3 grant. Much of the design and implementation work described in this article was done by Jack Jansen, Sjoerd Mullender, Pablo Cesar and Kleanthis Kleanthous. I thank them and this article's editors and anonymous reviews for their useful comments.

References

1. The Ambulant SMIL 2.1 Player. http://www.cwi.nl/projects/AmbulantPlayer/
2. André, E., WIP, PPP: A comparison of two multimedia presentation systems in terms of the standard reference model. Comput. Stand. Interfaces 18, 555–563 (1997)
3. Ardissono, L., Kobsa, A., Maybury, M. (Ed.): Personalized Digital Television. Targeting programs to individual users. Kluwer, Dordrecht (2004)
4. Borer, T., Davies, T.: DIRAC — Video Compression using Open Technology, EBU Technical Review (2005)
5. Bulterman, D.C.A.: User-centered abstractions for adaptive hypermedia presentations. In: Proceedings of the ACM Multimedia (1998)
6. Bulterman, D.C.A., Hardman, L., Jansen, A.J., Mullender, K.S., Rutledge, L.: GR*i*NS: A GRaphical INterface for creating and playing SMIL documents. Comput. Netw. ISDN Syst. 30, 519–529 (1998)
7. Bulterman, D.C.A.: Repurposing Broadcast Content for the Web, EBU Technical Review, 287, pp. 1–10 (2001)
8. Bulterman, D.C.A.: Using SMIL to encode interactive, peer-level multimedia annotations. In: Proceedings of ACM DocumentEngineering 2003 pp. 32–41. Grenoble, France (2003)
9. Bulterman, D.C.A.: Animating peer-level annotations within web-based multimedia. In: Eurographics Multimedia 2004, Nanjing, China, 27–28 October 2004
10. Bulterman, D.C.A., Rultedge, L.: SMIL 2.0: Interactive Multimedia for the Web and Mobile Devices. Springer, Berlin Heidelberg New York (2004)
11. Bulterman, D.C.A., Gassel, G., et al.: Synchronized Multimedia Integration Language (SMIL 2.1). http://www.w3.org/TR/2005/REC-SMIL2-20051213/
12. Cesar, P., Bulterman, D.C.A., Jansen, A.J.: An architecture for end-user TV content enrichment. Proc.EuroITV: 4th European Conference on Interactive Television, Athens, Greece, pp. 39–47, May 2006
13. El-Beltagy, S., DeRoure, D., Hall, W.: The evolution of a practical agent-based recommender system. In: Proceedings of Workshop on Agent-Based Recommender Systems, Autonomous Agents (2000)
14. Cooper, M., Foote, J.: Scene Boundary Detection Via Video Self-Similarity Analysis. In: Proceedings of IEEE International Conference on Image Processing (2001)
15. Daisy Consortium, Specifications for the Digital Talking Book, ANSI/NISO Z39.86-2005
16. Dowman, M., Tablan, V., Cunningham, H., Popov, B.: Web-assisted annotation, semantic indexing and search of television and radio news. In: Proceedings of the 14th international Conference on World Wide Web (Chiba, Japan, May 10–14, 2005). WWW '05. ACM Press, New York, pp. 225–234 (2005)
17. Fischer, G.: The Importance of models in making complex systems comprehensible. Mental Models and Human-Computer Interaction 2. Elsevier, North Holland (1991)
18. Flickr. http://www.flickr.com/
19. Foote, J.T.: Content-based retrieval of music and audio. In: Proceedings of SPIE Multimedia Storage and Archiving Systems II, vol. 3229, pp. 138–147 (1997)
20. Geurts, J., Bocconi, S., van Ossenbruggen, J., Hardman, L.: Towards ontology-driven discourse: from semantic graphs to multimedia presentations. In: Second International Semantic Web Conference (ISWC2003) Sanibel Island, Florida, USA, pp. 597–612, 20–23 October 2003
21. Haas, N., Bolle, R., Dimitrova, N., Janevski, A., Zimmerman, J.: Personalized news through content augmentation and profiling. In: Proceedings of ICIP'02, pp. 9–12. IEEE Press, Rochester (2002)
22. ITEA Project Passepartout: http://www.hitech-projects.com/euprojects/passepartout/
23. Jackson, D., Northway, C.: Scalable Vector Graphics - 1.2 Specification. http://www.w3.org/TR/SVG12/
24. Li, F.C., Gupta, A., Sanocki, E., He, L., Rui, Y.: Browsing digital video, in CHI '00: Proceedings of Human Factors in Computing Systems, ACM Press, pp. 169–176 (2000)
25. Merialdo, B., Lee, K.T., Luparello, D., Roudaire, J.: Automatic construction of personalized TV news programs. In: Proceedings of the Seventh ACM international Conference on Multimedia (Part 1) (Orlando, Florida, United States, October 30 - November 05, 1999). MULTIMEDIA '99. ACM Press, New York pp. 323–331 (1999)
26. MONET: Extending databases for multimedia. URL: http://www.cwi.nl/∼monet/modprg.html
27. MPEG-4 Specification. ISO/IEC JTC1/SC29/WG11
28. NIST, TREC Video Retrieval Evaluation Home Page. www-nlpir.nist.gov/projects/trecvid/
29. Patel, N.V., Sethi, I.K.: "Video shot detection and characterization for video databases", Pattern Recognition, Special Issue on Multimedia, vol. 30, pp. 583–592 (1997)
30. Qi, Y., Hauptman, A., Liu, T.: Supervised classification for video shot segmentation. In: Proceedings of IEEE International Conference on Multimedia & Expo (2003)
31. Robson, G.D.: The Closed-Captioning Handbook. Focal Press/Elsevier (2004)
32. Rodrigues, R.F., Soares, L.F.G.: Inter and intra media-object QoS provisioning in adaptive formatters. In: IV ACM Symposium on Document Engineering - DocEng2003, Grenoble, France, November 2003
33. Schulzrinne, H.: RTP: real-time transport protocol URL: http://www.cs.columbia.edu/∼hgs/rtp/
34. Schulzrinne, H.: A real-time stream control protocol(RTSP). URL: http://www.cs.columbia.edu/∼hgs/rtsp/draft/draft-ietf-mmusic-stream-00.txt (11/26/96)

35. Tsinaraki, C., Polydoros, P., Kazasis, F., Christodoulakis, S.: Ontology-based Semantic Indexing for MPEG-7 and TV-Anytime Audiovisual Content. Special issue of Multimedia Tools and Applications Journal on Video Segmentation for Semantic Annotation and Transcoding (2004)
36. TV-Anytime Forum website: http://www.tv-anytime.org/
37. W3C, Timed Text Home Page. http://www.w3.org/AudioVideo/TT/
38. W3C, XForms Home Page. http://www.w3.org/MarkUp/Forms/
39. W3C, XHTML Home Page. http://www.w3.org/MarkUp/
40. W3C, XPath Specification. http://www.w3.org/TR/xpath
41. W3C, XML Pointer, XML Base and XML Linking Home Page. http://www.w3.org/XML/Linking
42. Weck, D.: LimSee2: The Cross-Platform SMIL 2.0 Authoring Tool. http://wam.inrialpes.fr/software/limsee2/
43. Wold, E., Blum, T., Kreislar, D., Wheaton, J.: Content-based classification, search, and retrieval of audio. IEEE Multimedia 3(3), 27–36 (1996)
44. XIPF.org Foundation, Vorbis I Specification. http://www.xiph.org/vorbis/doc/Vorbis_I_spec.pdf
45. YouTube - Broadcast Yourself. http://www.youTube.com