Fragment, Tag, Enrich, and Send: Enhancing Social Sharing of Video

PABLO CESAR, DICK C. A. BULTERMAN, JACK JANSEN
Centrum Wiskunde & Informatica, The Netherlands
DAVID GEERTS
Centre for Usability Research, Belgium
and
HENDRIK KNOCHE and WILLIAM SEAGER
University College London, UK

The migration of media consumption to personal computers retains distributed social viewing, but only via nonsocial, strictly personal interfaces. This article presents an architecture, and implementation for media sharing that allows for enhanced social interactions among users. Using a mixed-device model, our work allows targeted, personalized enrichment of content. All recipients see common content, while differentiated content is delivered to individuals via their personal secondary screens. We describe the goals, architecture, and implementation of our system in this article. In order to validate our results, we also present results from two user studies involving disjoint sets of test participants.

Categories and Subject Descriptors: H.4.3 [**Information Systems Applications**]: Communications Applications—*Information browsers*; H.5.1 [**Information Interfaces and Presentations**]: Multimedia Information Systems—*Audio*; *video*; I.7.2 [**Document and Text Processing**]: Document Preparation—*Format and notation*; *hypertext/hypermedia*; *languages and systems*; *multi/mixed media*

General Terms: Design, Documentation, Experimentation, Languages

Additional Key Words and Phrases: Asynchronous media sharing, differentiated content enrichment, secondary screens

ACM Reference Format:

Cesar, P., Bulterman, D. C. A., 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, Article 19 (August 2009), 27 pages.
DOI = 10.1145/1556134.1556136 http://doi.acm.org/10.1145/1556134.1556136

1. INTRODUCTION

Online multimedia content sharing systems have slowly grown in popularity over the past decade. Initially, shared content consisted of photographs or short video clips that contained repurposed studio

This work was funded by the following projects: NL-NOW BRICKS PDC3, ITEA Passepartout, FP6 IST SPICE, and FP7 IP TA2. Development of the open source Ambulant Player was funded by the NLnet foundation.
Authors' addresses: P. Cesar, D. C. A. Bulterman (corresponding author), J. Jansen, CWI (Centrum Wiskunde & Informatica), Amsterdam, The Netherlands; email: {Dick.bulterman@cwi.nl; D. Geerts, Centre for Usability Research, Leuven, Belgium; H. Knoche, W. Seager, University College London, London, UK.

content, such as music, music videos, and individual news items that were intended to reach a wide, anonymous audience. During the past two years there has been a growing awareness of the power of personal media productions that are geared to a more limited community of viewers. These include holiday videos, focused product reviews, and intraproject demonstration videos. Video sharing Web sites have captured the imagination of a broad community of users and investors: Sharing personal media is "hot."

While the production model for digital media has undergone a fundamental shift in the form of user-generated content, the UI model for manipulating media has evolved much more slowly. The dominant model for Web-based content sharing and social interaction is still centered around a single user with a personal display, mouse, and keyboard. Even modern mobile devices remain very limited with respect to interacting with content. Both reflect that users are expected to consume media, not socially interact with it.

This article describes an architecture and implementation of an inherently more social approach to viewing and sharing media. The goal of our work has been to study languages and interfaces that allow casual users not sitting at a PC to socially interact with media in single and group contexts. Building on top of popular online video services, our system enables copyright-safe personal recommendation and forwarding of fragments of content in a family/friends network, as well as the further personalization of content with text, audio, and line art using an enhancement interface. The work reported in this article concentrates on two sharing models: *asynchronous content sharing* among members of a social group that are distributed in time and space, and *synchronous sharing* of content among members of a social group that are spatially and temporally colocated. A major difference from previous approaches is the distribution and previewing of content over a collection of personal control devices (secondary screens), in which media viewing is dynamically differentiated for members of a social group sharing a common display device.

This article is structured as follows. Section 2 provides a motivational user scenario for our work. Section 3 surveys existing systems, highlighting features that we feel are ill supported. Section 4 focuses on the architecture and implementation effort of the system. Section 5 reports on two qualitative studies that we have used to evaluate both the feature set supported by our work and the underlying interaction architecture. Lastly, Section 6 discusses the results and contributions of this work, highlighting the architectural implications for next-generation video sharing systems.

2. MOTIVATION

Mark and Katrina have two children. During a business trip to Vancouver, Katrina views a documentary about South America, the location of the family's upcoming vacation. She wants to share relevant parts of this presentation with her family back home in Amsterdam. This sharing consists of identifying the video, overlaying a personal navigation structure highlighting portions of interest for different family members, and adding a number of voice-over annotations to some of the fragments, such as "I want to go there" or "we have to visit this spot." She then sends the family a message with a pointer to "her" version of the video. Note that such enrichments do not generate a new version of the video (this would violate the content owner's copyright); instead, they are encoded as a set of overlay content wrappers containing a link to the base video, along with a navigation map and set of annotations. Katrina uses asynchronous sharing: The 9-hour time difference and the differentiated content make synchronous sharing (such as via a chat-based system) impossible.

Back home, Mark and the kids receive the message sent by Katrina. While they have multiple viewing options, they choose to watch it as a family on their high-definition television set (Figure 1(a)). During the presentation, Mark uses his mobile device as a secondary screen (Figure 1(b)): This screen exposes the personal navigation structure Katrina designed for him only. The navigation stream was packaged as

Fig. 1. The hybrid viewing environment, with one common display and multiple personal secondary devices.

Fig. 2. Creating a navigation/recommendation poster.

differentiated content in Katrina's message. Each of the children also received personalized information, which they may view on *their* personal devices.

Neither Katrina nor Mark are in the video editing or postproduction business. When Katrina made the original recommendation, she used a portable device without keyboard or mouse, but only a touch interface. The interface is shown in Figure 2: a poster image is selected for each navigation point, which is added to the collection of posters for that program based on its temporal positioning in the content. Katrina may add optional text captions, voice-over descriptions, and even line-art overlays if she so desires (either for everyone or for a particular user's secondary screen), but this isn't required. She may even time-limit some of her annotations, knowing (in this case) that they won't be important after she returns home.

This scenario highlights two of the major contributions of our work:

(1) an interface that supports the direct recommendations of content to others in a social network, using lightweight [Kindberg et al. 2005; Kirk et al. 2007] or full-feature editing systems. The

recommendations can reference all or part of a base piece of media by using nondestructive fragmenting of content;

(2) the development of a personal remote control model that allows users to manipulate and view metainformation, and preview content in a mixed social setting, providing a private space in a socially crowded livingroom.

In contrast, consider other approaches to content enhancement. Current technology allows a degree of content recommendation and personalization, but it does so using Web-based solutions that are separate, nonintegrated components: in the context of our example, Katrina first needs to find the video on a Web server, then she has to download it and possibly reformat it for use with the authoring tools on her PC, and she then needs to recode and upload the video to her personal server and forward a link to her family in Amsterdam. As discussed further in Section 3, these represent unnatural solutions to the sharing problem and that do not integrate well in environments such as a shared livingroom.

Previous work on authoring systems has concentrated on the provision of full-fledged software tools with complex functionality. Work in this area includes the provision of integrated solutions for the full creation process [Adams et al. 2005], tools for creating directed collections of family content [Abowd et al. 2003], visualization tools for viewing and creating hypermedia documents [Shipman et al. 2008], and collaborative support authoring tools for stored [Sgouros and Margaritis 2007] or live content [Resende Costa et al. 2006]. In contrast, our objective is to provide a personal and lightweight authoring tool that can be easily integrated into a number of contextual environments. The proposed tool is personal because it allows end-users to enrich the media content as it is watched. It is lightweight because the authoring process is done while watching, and with the devices used for watching, so there is no need to shift to another environment for the authoring process. Note that our goal is not to organize collections of media for personal use, but to share media in an informal, nearly transient manner with members outside of the family circle.

Similar aspects of research can be found on video editors for mobile phones [Jokela et al. 2008], in the *Watch and Comment* paradigm [Cattelan et al. 2008], and the *DJ DreamFactory Platform* [Liu et al. 2007]. However, Jokela's work is limited to mobile authoring, Cattelan's work requires the usage of gesture-enabled portable PCs for watching and authoring, and Liu's work is only intended for Web-based collaborative broadcasting of media. Interestingly enough, all three solutions are based on the same underlying concept as our application: the usage of structured multimedia documents [Bulterman and Hardman 2005] for the development of rich media presentations. Alongside research on authoring environments, multimedia annotations [Stamou et al. 2006] play an essential role on the new media landscape. Even though we consider annotating, we prefer the term *tagging* [Marlow et al. 2006], since it is done by the end-users as one of the lightweight activities provided by our application. We go one step forward from conventional tagging by providing actual enriching features, such as the addition of a personal audio commentary at the beginning, during, or at the end of a sharable video.

A notion of fundamental importance in our work is the (relative) temporal and spatial scope of the comments being made. When a viewer makes a recommendation for a friend, this recommendation is intended to be asynchronous: the receiver is not engaged in a concurrent chat session with the sender, but will receive the recommendation message only when he/she is receptive. On the other hand, several messages about the same content may be created synchronously by multiple parties who are interacting together in the same physical and temporal space: in order for Katrina's two kids to be able to create messages concurrently for their (disjoint) social groups, each needs a personal device that allows them to manipulate content without blocking other family members. In this manner, our work extends the sharing interface from the desk to the couch.

3. BACKGROUND

Modern online video services provide social features such as posting comments about a specific video, rating them, and sharing video material with others by embedding a fragment of HTML code that includes the video's location. In spite of their success, there are a number of serious restrictions in such interfaces. First, videos are addressed as atomic objects, without any partitioning in time or space. (We call such partitions *fragments.*) Second, the lack of content-based fragmentation brings with it a lack of intraobject navigation. Third, the lack of user-defined fragmentation results in an inability for users, rather than producers, to share bounded portions of an object among subgroups of viewers. For shorter videos this might not be needed, but for longer videos it is often useful to define a short fragment of the base video that can be used to illustrate a particular point. Finally, the user cannot customize the recommended video by including, for example, a voice commentary or strategically placed line-art overlays.

In order to categorize basic and innovative features provided by current video sharing systems we have selected four representative examples: YouTube[1] , Asterpix[2], Yahoo! Videos[3], and Lycos Cinema[4]. Together these systems provide video description and manipulation functionality both in a synchronous and asynchronous manner [Chorianopoulos 2007]. The intention of this article is to focus on asynchronous media manipulation capabilities for sharing and situationally synchronous media manipulation for creation; we find that this provides the most realistic operational use of a content enrichment facility.

3.1 Asynchronous Manipulation Features

There are several activities performed by the participants in a media sharing system. We differentiate the functionality required by content owners and content users. *Content owners* are defined as the initial parties to share a piece of media. They require the following functionality.

—*Upload:* a facility to add media to the content server;

—*Describe:* a facility to describe the entire media object; to be used for searching and for display during viewing; and

—*Tag:* a facility to add keywords about the media content.

Content viewers are defined as parties that reference the owner's content; they may send others pointers to the content. The basic functionality required by viewers are as follows.

—*Share:* a facility to send recommendations to others. This can be done via an email with a link to the media or as an embedded HTML fragment on a social Web site;

—*Comment:* a facility to post comments about an object, in whole or part; and

—*Rate:* a facility to indicate the popularity of a video. Users might explicitly rate it or favorite it, or the site might use nonintrusive metrics such as implicitly counting the number of views.

Advanced user features are those extensions to conventional sharing behavior that allow personalized, focused sharing and enhancement of content by nonowners (without compromising the rights that owners have).

[1]http://www.youtube.com/
[2]http://www.asterpix.com/
[3]http://video.yahoo.com/
[4]http://cinema.lycos.com/

—*Fragment:* a facility that allows a user to define one or more ranges of clips within a base media object. These fragments can be explicitly or implicitly exposed to parties with whom the user shares content.

—*Annotate:* a facility to add user-generated notes or comments to a particular media fragment. The annotations may be audio, text, image, or line art in nature. The annotations may be exposed to all parties sharing the media, or only to a user-defined subset.

—*Enrich:* a facility to add new temporal links, subtitles, captions, remixing [Shaw and Schmitz 2006; Shamma et al. 2007], repurposing [Pea et al. 2004], overlaid media, or voice introduction to a baseline object. These enrichments may be layered; that is, a particular media object may expose a history of enhancements by various parties.

The purpose of our work has been to design and implement a prototype environment for studying advanced user features, and to evaluate these features in two independent user trials.

3.2 Synchronous Manipulation Features

In traditional Web-based media sharing systems, many thousands of users could potentially be adding annotations about a particular object concurrently. Each user sits at his or her own display, using his or her own keyboard and pointing device. This approach works well when each of the users are physically separated, but it is much less appropriate when two or more users attempt to add, edit, or view fragments simultaneously in the same physical space. Consider the family livingroom: assume that multiple family members are sitting together on the couch. While watching a movie together, each finds content of interest that they would like to share in their personal social networks. All members could try to grab the shared remote control, or each could reach for their personal secondary screen devices. In this latter model, the users share a common control environment in that they all view the common content on the main screen. Each may perform individual edits concurrently, however, either browsing or contributing new annotations for or from friends.

The synchronous manipulation that this article investigates is not based on synchronized communication between users, but synchronous manipulation among users who independently manipulate common content within a shared physical space. As we will discuss in Section 4, this requires a layered approach to not only media content extensions, but also for interlaced control among participants viewing content together.

3.3 Selected Examples

Before describing our sharing architecture, we consider features available in current-generation sharing systems. As shown in Figure 3, these include functionality such as title-based search, simple annotation of a full clip, popularity tracking, content rating, third-party commenting, and an external referencing interface for embedding source material in (other) social Web sites such as FaceBook[5] or MySpace[6]. Other features include flagging content, the possibility of responding to a video by uploading a new video, the ability to create personal playlists, and community features for inviting friends or forming groups.

Asterpix is a Web service that provides access to media content from different sources such as DailyMotion[7]. It includes functionality similar to YouTube, but adds an important feature: The viewer is capable of enriching the video by adding temporal links such as commentaries or related videos.

[5]http://www.facebook.com/
[6]http://www.myspace.com/
[7]http://www.dailymotion.com

Fig. 3. Two YouTube media sharing interfaces.

Fig. 4. The Asterpix media interface.

As shown in Figure 4, the temporal links are rectangular shaped and when surrounding an object the system uses a tracking system to automatically follow such object. After the enrichment process, Asterpix uploads a new version of the video under a unique URI.

Yahoo! Videos provides an interface similar to YouTube, but includes a "tag it" feature, for facilitating the indexing and searching of videos. In addition, personalization functionality such as selecting a different thumbnail for a video embedded in a social Web site is provided. While watching a clip using Yahoo! Videos the user can start Jumpcut[8], a Web-based video editor, to edit the clip.

In contrast to the other examples, Lycos Cinema provides a synchronized experience. Lycos Cinema is a virtual theater in which the user can invite people to join for watching a movie together. It offers synchronous features such as chat, presence awareness, and join invitations, similarly to other social interactive television systems such as Joost[9] and Motorola's Social TV/TV2 [Metcalf et al. 2008]. Lycos Cinema also includes an asynchronous feature in which the user can clip a video. Moreover, similar

[8]http://jumpcut.com/
[9]http://www.joost.com/

Table I. Comparing Selected Functionality Across Systems

	Share	Annotate	Fragment	Enrich	Multi-Layered Enrichments	Differentiated Social Sharing
YouTube	+	-/+	-	-/+	-	-
DailyMotion	+	-	-	-	-	-
Asterpix	+	+	-	+	-	-
Yahoo!	+	-	-	-	-	-
Lycos	+	-	+	-	-	-
CollaboraTV	-	+	-	+	-	-
Zync	+	-	-	-	-	-
Joost	+	-	-	-	-	-

to Zync[10], it allows synchronized watching of online video. Another application along the same lines is CollaboraTV [Nathan et al. 2008], which allows a user to add temporal comments that will be shown during the selected video fragment when the video is watched by the peers of the user.

3.4 Undersupported Features

Many sharing sites provide institutionalized support for community building around their media. The effectiveness of such support is questionable [Halvey and Keane 2007]. Recent data suggests that "...users are directed to YouTube by friends sending them specific videos" [Gill et al. 2007] and that "the aggregate views of these linked videos account to 90% of the total views" [Cha et al. 2007]. This confirms our belief that people are directly guided by other people in the media selection process. Our work illustrates two major directions in which sharing facilities of media can be improved: media manipulation and the ability to support shared but differentiated social viewing.

Table I shows a comparison among current systems for a selected set of functionality. The table assumes that all the systems already provide the basic functionality of media description support for facilitating searching. Only a minority of the systems allow fragmentation, annotation, and enrichment of media. Moreover, such systems do not provide support for differentiated shared viewing.

Of the characteristics in Table I, the facilities for sharing, annotation, and fragmentation have been covered earlier in this article. By *multilayered enrichment*, we mean a facility that allows multiple enhancements that are logically layered to be managed from a single user interface. The opposite of this functionality is that each enrichment gets published as a separate, unrelated object. Such multipublishing is not uncommon: many videos on YouTube already have one to four aliases [Cha et al. 2007]. Since these are all published as independent objects, no content management user interface is available to aid in fragment searching.

The final characteristic in Table I is labeled *differentiated social sharing*. This facility allows one viewer to obtain (either by direct request or as a property of the recommendation) extra information that may not be visible to others, even when all viewers are watching a single common media stream on a shared display such as a TV. By using a secondary display device, users can obtain additional information on the displayed content without disturbing others, they may be able to manipulate a

[10]http://timetags.research.yahoo.com/zync/

separate control interface (for example, Mark's use of a secondary display for navigation in Section 2), or it may allowed targeted content to one of the shared viewers, such as a personal hint in an otherwise shared-experience game. Current research indicates that "our devices should collaborate to support a notion of user-centric activities that span multiple devices]" [Dearman and Pierce 2008].

4. ARCHITECTURE AND INFRASTRUCTURE

The goal of the research described in this article has been to evaluate the usefulness and feasibility of providing media manipulation functionality as a spontaneous activity in a social environment, such as the livingroom. To support this evaluation, we implemented a prototype that allowed users to view third-party media, to construct content fragments and to add annotations, and to share these annotated fragments within a small-scale social network. In this network, one size did not need to fit all: individual users could have tailored messages, and all users had the option of using a secondary screen for supporting navigation.

This section describes the architecture of our implementation, with special focus on content modeling, the system software, and the social network aspects. The main contribution of this prototype is the development of an open-ended testbed system in which the creation of media fragments is easy to perform, in which such fragments can be enriched and targeted to specific viewers, in which video manipulation does not result in a new encoded version of the media, and in which and all such activities can be performed in the context of differentiated social sharing. The implementation of our architecture is integrated into the open-source Ambulant player.[11]

The results of the UI aspects of this work are summarized in Evaluation Results. There are also more generic results to report. One of these results is that, by making media annotation a nondestructive, layered task, the number of video servers required to store media content can be significantly reduced. At the same time, real personalization of video content, such as a postcard sent by a friend, is provided. We find these aspects to be significant.

4.1 Design Goals

The primary design goals that guided the work reported in this article were as follows.

—*Investigate Distributed, Concurrent Control*. A home media environment is a complex combination of people, devices, and content. The collection of people who consume content will vary from a single person to a local collection of family members to a distributed collection of remote viewers. Rather than assume a single point of control (with a single hand-held device in one room), we wanted to study a broader base of content and control interaction in which multiple control pointers were active simultaneously.

—*Separate Rendering of Control Information from (Shared) Content*. Given the diverse user, content, and rendering environments, we wanted to explicitly separate out the viewing and interacting with control information from the viewing and interacting with media content.

—*Focus on Individual Users Instead of Shared Devices*. Most digital television systems are centered around a set-top box: This box is connected to an external content stream, it puts content on a TV display, and it interacts with the user's remote control. Architecturally, the user is an appendage to the system. In many households, there are multiple users who each have their own content needs. We wanted the user to be central in our system.

—*Enable a Framework for Micro-Recommendations*. Content recommendations are typically managed both at a device level (that is, figuring out which content the set-top box should store) and at the full-program level. We wanted to develop an environment where individual programs could be partitioned

[11]http://ambulantPlayer.org

(by the user or system) into a collection of fragments of interest, and in which collections of programs could be grouped in to packages. This framework would be the basis for future study on automated microlevel recommendations.

—*Enable a Framework for Sharing of Recommendations*. It is becoming common for recommender systems to gather content for a user, and for a user to rate content. We also wanted to investigate ways of having individual users send personal recommendations within their family or social network.

—*Interface with Practical User Environment*. We wanted our system to build on existing use models for broadcast content. Rather than assuming that conventional broadcast outlets will disappear, we wanted to study ways of working within a common home consumer electronics framework.

The last design goal inspired us to focus on relatively passive viewing of content in a mixed personal/social setting. The mixed setting provides a wealth of interesting interaction problems that are not encountered when it is assumed that a user is sitting behind a personal computer. We feel that results from the family couch can scale to the PC world, but that PC-based solutions may not scale to the passive couch-top environment. A nontechnical goal was to frame the technical progress in our research in a context that will likely have broad impact on the way real people consume real media. This is the motivation for the integration of several external user tests.

4.2 Content Modeling

Currently, online video systems widely support the capability of sharing video material among users, either by sending email links or by embedding the content into a social network. As indicated before, the major form of accessing video on the Web is via this sharing capability [Gill et al. 2007]. When content is shared by migration, the generation of new videos has negative implications: it saturates already overloaded video servers [Cha et al. 2007] and it introduces an impossibility to manage composite video edits performed on various instances of the same high-level video object. The saturation of video servers not only has obvious economic implications, it presents authorship/copyright implications as well; that is, the owner of a video does not have any idea of who and how other people are manipulating his/her production. The duplication of video implies that a downstream viewer (in the temporal sense) cannot selectively recall enrichments made by others unless these were encoded within the local instance being manipulated. Moreover, essential information (such as who has manipulated what when), very useful for archival and indexing purposes, is lost when a new encoded version of the content is provided.

Implementing video manipulation activities such as fragmenting, annotating, and enriching as nondestructive operations linked to a single master copy of a video provides a solution to these problems. Current research provides two major directions: temporal URIs [Pfeiffer et al. 2003; Rutledge and Schmitz 2001] and structured media documents [Goularte et al. 2003; Hua and Li 2006]. The first approach permits sharing video fragments by replacing a base URI with an annotated URI that indicates the starting and ending point of the fragment. Such a solution can be easily implemented and enhances the video sharing functionality. Nevertheless, because the amount of information that can be fitted into a URI is limited, such a solution is not powerful enough for more complex video manipulation capabilities. Furthermore, essential metadata such as who has created the fragment, when, why, and for whom might be very difficult to include. Hence, our system relies on structured documents for providing enhanced media sharing features in the form of SMIL 3.0 [Bulterman and Rutledge 2009] and TV-Anytime Phase II.[12]

In our environment, the starting point is an unstructured, or raw, video. We then define a structured shell in the form of a SMIL description by using authoring templates. If an automatic scene selection

[12]http://www.tv-anytime.org/about/phase2.html

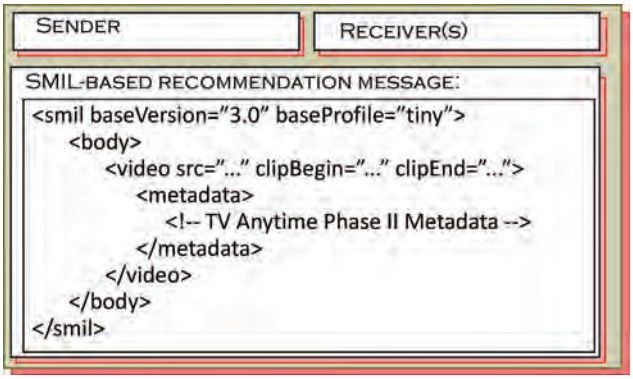

Fig. 5. Modeling micropersonal recommendations.

tool is available, the video can be further structured into sections of interest. All navigation points are encoded as a series of content events that are placed in a dynamically created SMIL file. Each of these functions is captured via a *poster interface*. Each poster contains the following information.

(1) *temporal moment of the navigation point*: the time at which the navigation point appears in the content.

(2) *enrichments*: text, ink, audio, and link enhancements that have been added by the user. Note that similar content is generated for studio-produced navigation points if they are distributed as TV-Anytime markup.

Our content hierarchy allows navigation/recommendation points to be described at three levels: the package level, where collections of programs are stored and grouped by package name; the program level within packages, where individual programs are identified; and the fragment level within programs, where individual navigation points are identified. It is a longer-term interest to have the partitioning of content between packages, programs, and fragments occur automatically. However, at every level, a user should be able to exert a personal influence over content scheduling.

When video manipulation takes place, the SMIL file is used to control the interactive display at the client. The SMIL file is transformed into a TV-Anytime description at the point that one or more recommendations are transformed into a recommendation message. (Details are available in Cesar et al. [2008].) At no time is actual content integrated, manipulated, or shared among users. All communication is done via an abstracted SMIL file or as a set of TV-Anytime markup. Hence this is a nondestructive solution for the video manipulation challenge. One very important advantage is that viewers might select to see the manipulations performed by a specific user or turn off all enrichments at once.

Figure 5 shows the structure of the SMIL-based recommendation message resulting from our lightweight manipulation of video content. As shown in the figure, SMIL 3.0 is used as the host language; the end-user manipulations on the video are encoded in the SMIL file, using SMIL constructs. They do not change the base media objects in any way. For example, the end-user can fragment the video by using *clipBegin* and *clipEnd* attributes. Moreover, the end-user can insert audio overlays by using the audio element or any other kind of multimedia overlays. The actual process of SMIL creation/manipulation is hidden from the end-user, who simply sees an interactive UI on the secondary device.

An important feature included in SMIL 3.0 is the expansion of the metainformation facilities. In previous versions, metainformation was restricted to the head element, meaning all metainformation referred to the whole document. SMIL 3.0 now allows placing metainformation on any element within

the document body. This makes it possible to provide information on semantic intent within the presentation by making the binding of that information with relevant nodes more local. In our example, the video fragment indicated in the video element is annotated with the universal TV-Anytime content descriptors to help the localization process. Each fragment of video could be further annotated by using, for example, Friend of a Friend (FOAF) for indicating the creator of the fragment or the person intended as recipient.

4.3 Home Network

Another challenge addressed by this article is the dynamic distribution of media content and control to the most suitable device at home. The goal is to dynamically monitor the end-user environment for available rendering/interactive devices. Moreover, such devices should provide an accurate description of their physical, rendering, and interactive capabilities. Based on these descriptions, the context of the user, and the nature of the media to be watched, we can take a decision on the actual distribution mechanisms to be utilized. It is out of the scope of this article to describe the device description mechanisms and decision algorithms; the interested reader can refer to Hesselman et al. [2008].

Device discovery is done by an exchange of invitations to join the network using Bluetooth, Wireless LAN, or IP Multimedia Subsystem (IMS). After devices have been discovered, they provide the descriptions of their physical, rendering, and interactive capabilities. In order to support a wider set of devices than mobile handsets, we have extended UAProf[13] for describing devices' capabilities. Applications are described as a Software Oriented Architecture (SOA), in which the input and the output model are described using XML. Based on device and service descriptions, a matching algorithm together with basic recommendation facilities are used to determine where to render the media and how to provide user interaction in the most suitable way depending on the context of use.

Within the home, the central content storage/management and service publishing component is a home media server. This server can ultimately be implemented in many different forms (as a PC Media Center, as a conventional set-top box, as a network controller hidden in a utility closet). The result is a hybrid architecture in which devices can be directly connected to each other in a Peer-to-Peer (P2P) fashion within the home. Our main concern was not to study the commercial models for home media servers, but: (1) to study a model in which multiple control clients could be managed in a home environment and (2) dynamically distribute the media content to the most suitable rendering device(s). For this reason, we made the pragmatic decision to use a small-size personal computer (in our case, a Mac-Mini) upon which our server infrastructure could be implemented.

Apart from device discovery, dynamic content distribution, and service publishing, the server performs the following functions.

—*It Connects to an External Content Pool*. This pool consists of a connection to (digital) broadcast content, to a peer-to-peer content infrastructure for sharing nonprofessional content, and to a collection of physical optical devices such as DVD and BluRay disk players.

—*It Caches Content that is Differentiated per User*. Each of the users of the environment is managed separately. Each maintains their own content preferences and their own user group. For each user, nonprotected content is cached using PVR-like functionality.

—*It Implements a Content Recommender System*. The server manages the recommender environment for the home. This includes communication with external recommender systems, forwarding and receiving recommendation messages, and enabling users to add and share personal recommendations within program content.

[13]http://www.openmobilealliance.org/

—*It Provides a Home Management Interface*. Not all content recommendations generated by external systems will actually be suitable for all members of a family or social network. In addition to simply storing recommendations, our server environment provides a management system in which a hierarchy of control allows privileged users to manage (override, augment) the recommendations provided for others.

—*It Communicates with the Client Devices within the Home*. This includes communicating with the distributed remote control devices and actually sending information to clients.

Our architecture has been designed to be aware of DRM issues in the home. All operations on actual media content are abstracted from the actual media encoding into a higher-layer structure. A portion of this structure uses the TV-Anytime specification for program and package descriptions. The local user operations are implemented by dynamically generating SMIL presentations that describe the transient structure of content modifications and annotations within a program. The work reported in this article does not focus on recommender strategies (other than gathering and forwarding personal micro-recommendations). Our focus is on the communication aspects of distributing control in a home environment. This is a function of the communication with the collection of client devices.

Content navigation is performed based on the set of posters that have been defined for a particular program. These posters may be defined by the content owner (in this case, the BBC), they may be automatically induced, or they may be defined by a viewer. When the user fragments a video, a screenshot poster is taken from the primary screen. The user may enrich the poster or the video fragment with additional information (annotations), line art, or an audio voice-over.

4.4 Social Network

Current online video sharing systems are based on two models: person-to-person and person-to-world. The first makes use of an active email address in which a link to the video is included, while the second provides an HTML fragment to be posted in a Web page in which a link to the video is embedded. In this article, we argue that a broader social network architecture is needed. Our architecture is based on social network capabilities in which individuals can be gathered as solitary users (in front of their personal devices) or as a group of people such as families that interact with a common social medium such as the TV. We use XMPP[14] and the Google extensions[15] to retrieve contact information about the peers.

Our work divides the social communication model into three categories: immediate communication/mobile, immediate communication/static, and world communication. In Figure 6(a), the home server is connected to a message gateway[16] that generates a Push WAP message in the form of a SMS, so the recipient can display the video fragment using his/her mobile phone. In Figure 6(b), the message can be sent in the form of an email to the recipient's home server, which then informs the user using his/her active connection device (TV/PC). Finally, the sender can post the message to a Blogger[17] account (as illustrated in Figure 6(c)), allowing the enriched fragment of video to be shared with the world [Cesar et al. 2008].

4.5 System Architecture

Previous sections have discussed specific aspects of our application such as: (1) how the multimedia content is modeled, (2) how media content and control are distributed among different devices in the livingroom, and (3) how we realize the sharing of the resulting micropersonal recommendation. This

[14]http://www.xmpp.org/

[15]http://code.google.com/apis/talk/jep_extensions/extensions.html

[16]We use Clickatell (http://www.clickatell.com/)

[17]We use the Blogger API (http://code.google.com/apis/blogger/)

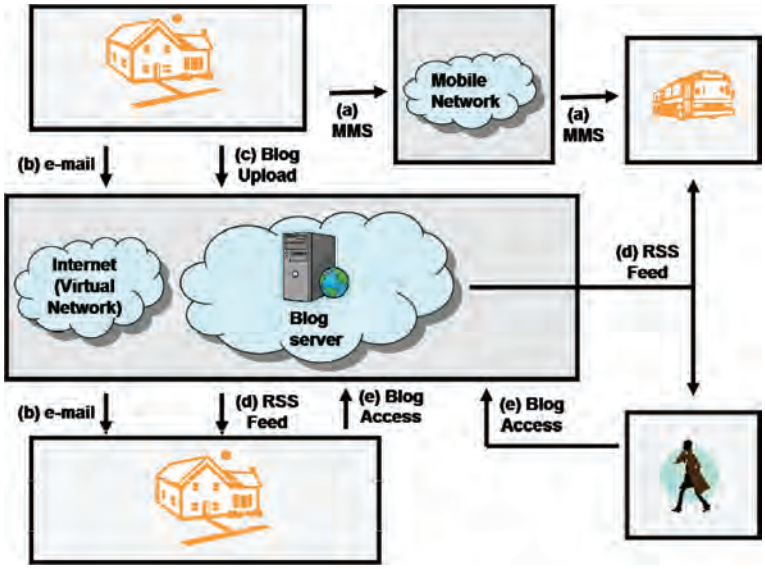

Fig. 6. Distributing/sharing micropersonal recommendations.

section introduces the overall system architecture of our application and describes the most relevant composing subblocks.

The overall architecture is composed of the following five components.

—*Presentation Module*. It is responsible for parsing and rendering the structured multimedia documents. It includes functionality for managing the layout and timing information of the document, and for rendering on the most appropriate user device at a given moment. In addition, it provides extra functionality over traditional media players for showing/hiding multiple layers of enrichments.

—*User Interaction Module*. It is responsible for handling user inputs originated in any of the available end-user devices. The user interaction module is responsible for parsing the user input and converting it into instructions for either the Control or the Lightweight Authoring modules.

—*Control Module*. It is responsible for managing the user actions that affect a running presentation such as pausing, playing, or rewinding a video displayed in the high-definition television. This module is responsible also for actions resulting in a change on the media rendered in a secondary screen, for example, a request to show the metadata associated to the current video fragment being played in the high-definition display.

—*Lightweight Authoring Module*. It is responsible for realizing the user actions that manipulate the running presentation. As indicated before, such manipulations do not alter the media content but the describing media document. For example, this module implements functionality such as *add_media*, *add_annotation*, *fragment_media*.

—*Sharing Module*. This module is the interface with the outside world. It is responsible for creating the personal micro-recommendation, including the information of the receiver(s), and for sending it to others.

In order to better explain how the application works, we provide the flow diagram between the components for two differentiated activities: content control in Figure 7 and content enrichment in Figure 8.

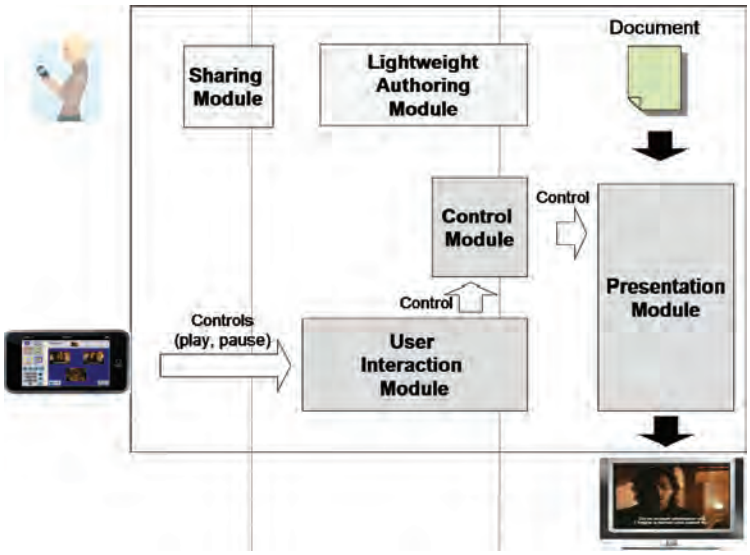

Fig. 7. System architecture: content control.

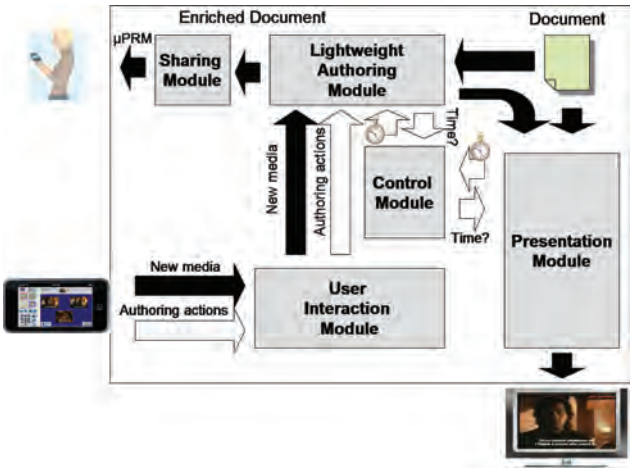

Fig. 8. System architecture: content authoring.

In both cases we assume a running presentation, such as a video being displayed, where the Presentation Module is rendering a multimedia document. It is important to notice that the Presentation Module includes a scheduler and a running clock for controlling the temporal synchronization of the presentation. In addition, since the multimedia document includes references to the media, either local or global, the Module includes a data source resolution submodule. Finally, it provides functionality for managing the layout of the presentation.

The content control phase involves three of the five system components. The User Interaction Module receives control commands from the available devices, such as the remote control or a hand-held device. Then, it converts the control into an action that it is handled by the Control Module, which informs the

Presentation Module in case the action affects the running presentation. For example, when the user wants to pause the Presentation Module, it needs to pause the running clock and all the other active media elements.

Figure 8 shows a more complex use case, in which actual authoring commands, such as to fragment a running video or to add new media to the video, are requested from the user. Once again the commands are converted into actions that are, in this case, handled by the Lightweight Authoring Module. In order to realize the appropriate action, the Authoring Module needs to request the current presentation time, through the Control Module, to the Presentation Module. For example, when fragmenting a video clip or adding new media, it is necessary to know the timing at which the fragment should be started. Then, the Lightweight Authoring Module creates the enriched multimedia document, based on the running presentation and taking into consideration the right timing information. The resulting document is the input for the Sharing Module, which constructs and sends the micropersonal recommendation based on the social network information.

The architecture hides authoring issues such as determining when the content manipulation takes place. This information is based on the internal clock of the running presentation. In addition, since the manipulation is done at the document level, it is easy to implement and nondestructive. It is easy to implement because content enrichment translates into the simple addition of XML text to the existing base-document. For example, by adding the following SMIL fragment to the document (<audio src= '...'>), we can include a new audio commentary. Other actions, such as fragmentation, only require the deletion or the modification of already existing SMIL fragments in the base-document. It is non-destructive because the enrichments take place at the document level, without actually manipulating the underlying media.

4.6 Content Lifecycle

The new media landscape, in which the end-user becomes an active node that might affect multimedia content, requires a revision of traditional views on the multimedia content and metadata lifecycle [Kosch et al. 2005]. These views focus on how to provide enough information for helping the user to find interesting media content, assuming the end-user as an active entity that only consumes the media. Our system provides the mechanisms to empower the user with viewer-side enrichment of content. Thus, the lifecycle of multimedia content does not end at distribution time, but effectively starts at distribution time. Figure 9 shows the resulting media document after fragmentation and enrichment of a given video has taken place. In this example, the user has fragmented the active presentation (*clipBegin*, *clipEnd* modifications in the <video> element), the user has added an audio commentary in parallel to the video (new <audio> element), and he has included some ink overlay (new <ink> element). As introduced in Section 4.5 all these simple manipulations at the document level translate in an actual personalization of media in which the user authors his personal view on the video content.

4.7 Current Status and Limitations

In order to gather useful information via user studies, a full working implementation of the system was developed. The implementation is composed of a number of elements, as introduced previously. The software base for the implementation of the fragmenting/enriching tool is the Ambulant Player,[18] an open-source media playback engine. Functional extensions have been made to support the functionality presented in this article. Individual clients such as Nokia 770 make use of local applications that perform enriching operations. The range of operating systems and media libraries make the development of

[18]http://ambulantPlayer.org/

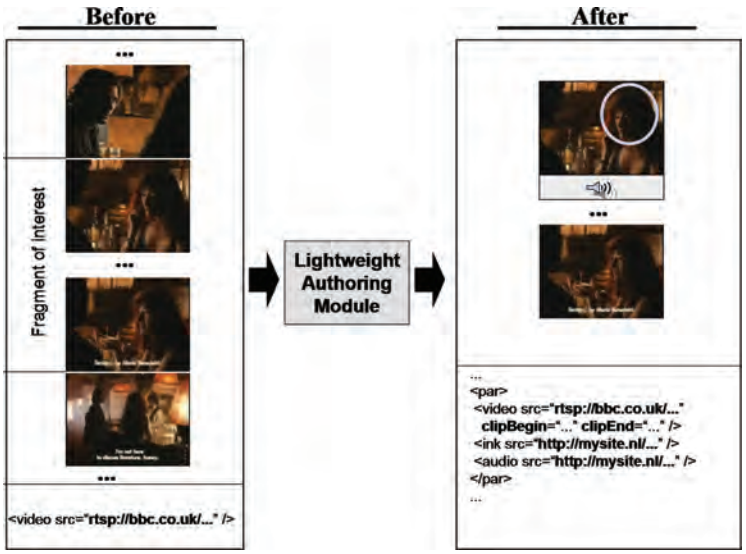

Fig. 9. Modeling content manipulations.

custom clients unavoidable, but all control operations have been harmonized in a layer of interface specifications.

Currently, the system is a working prototype, suitable for use in guided field trials. The initial encouraging results from the user testing described in this article have motivated an extended investment in the demonstration system to improve its stability and performance. Developments of the prototype are updated on the Ambulant Web site.

5. EVALUATION RESULTS

In order to evaluate the functionality provided by our architecture, we submitted our prototype to user testing. Our tests focused on a qualitative analysis and took place in two countries: the UK and Belgium. The tests were temporally separated, allowing the systems to be improved and more functionality included based on the results from the first test.

The intention of the tests was not so much to evaluate the specifics of the prototype's user interfaces, but to get a better understanding on the benefits of end-user content manipulation, as perceived by end-users. Interestingly, both studies provide similar results on the user expectations towards media sharing systems. We can identify a number of commonalities between the results.

Video fragmentation and enrichment. Users welcomed the capability to fragment and enrich content. The qualitative tests indicate that people want such functionality and that, when available, it will be widely used. Users like the capability of creating clips from videos, either for better navigation or to be able to send specific parts of a program to someone. Moreover, they enjoyed being able to enrich such clips, and sending them to other people. This is much in line with the current social practices of television watching, for example, talking during a program but also discussing it afterward. The fundamental implication of both sets of tests was that, rather than simply writing comments about media on social sites, users wanted to be able to subset the content as well (and to highlight particular sequences in the content).

Secondary screen. It was clear from both tests that users liked the idea of a secondary screen in multiparty social settings that allow usages other than controlling media. In both locations, participants

Fig. 10. Evaluators recommending content.

said that a secondary screen which allowed them to edit and send content without interrupting coviewers who are watching television would be very useful. As watching television in a social setting with multiple people (either friends or family) is still a dominant paradigm, this is an important conclusion that supports the design choices for our system.

By inspecting the recorded video sessions, the notes taken during the sessions, and the participant's answers to the questionnaires, we gained insights into how a system like ours could expected to be used once deployed. Results indicate that participants will use it to share content with people in their inner circle such as friends and family. They will share enriched content for maintaining relationships and for informative reasons. In most cases they will do it while watching, even though support is needed for bookmarking a specific time code for later enrichment. At the same time, a number of concerns were raised about the type of devices to be used. Participants preferred using light devices with finger-based interaction over stylus. From a technological perspective, they highlighted that a notification mechanism for incoming recommendation together with a clear identification on who sent the recommendation should be incorporated in the system. The following sections provide information on the methodology followed during the tests, the settings, and the characteristics of the participants. More detailed analysis on the obtained results is included as well.

5.1 UK Study

We modeled a representative user community by constructing 12 groups of paid subjects. Each group held up to 3 friends. A total of 27 participants took part in the study (average age: 28; meridian age: 23). Of these, 18 participants were male and 9 were female. The study was segmented into individual test sessions which lasted 90 minutes. Each session was video recorded. After each session, questionnaires were administered to collect feedback on individual preferences of the features of our system.

In terms of user experience, we constructed a home-like environment consisting of a local media server and three hand-held control devices, one for each participant. Content was delivered on a shared 50″ 16:9 display screen. We encouraged users to explore the system during a colocated test situation and to comment on the features or on any implementation issues encountered at any time. A photograph of one of the evaluation groups is shown in Figure 10.

The individual test sessions were followed up by a group discussion phase in which we asked questions such as:

—What did you like about the system?

—What other things would you like to be able to do?

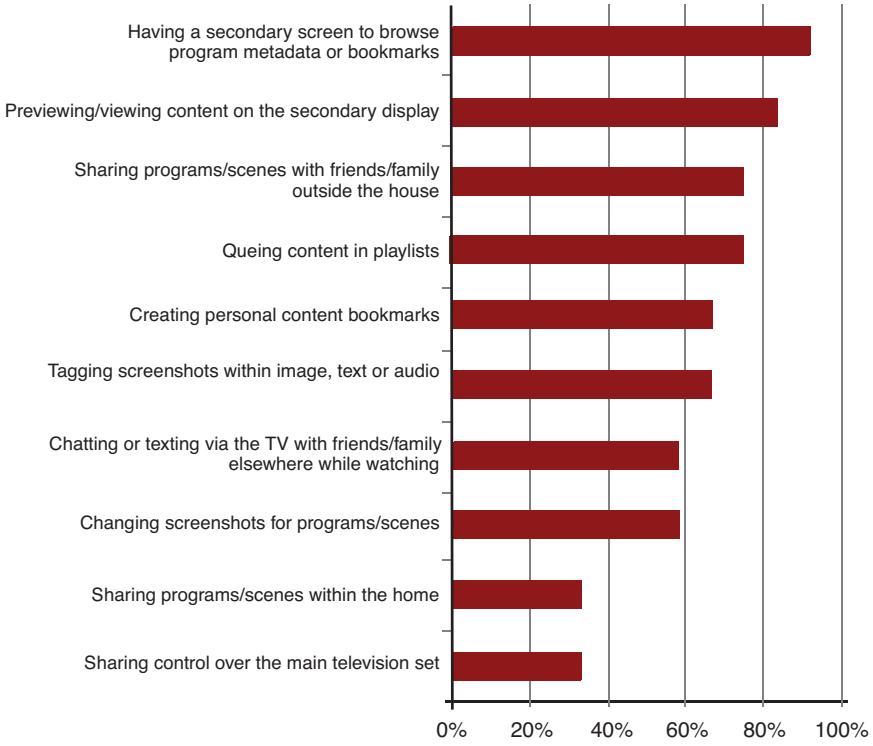

Fig. 11. Evaluations of expected end-user functionality from the UK tests.

—What didn't you like about the system?
—What were the most annoying things that should be changed?
—Would you be interested in using such a system at home or elsewhere?
—What was your experience in a group with multiple remote controls like?
—If you could annotate content how would you like to do it?
—When did you make your decision on what to watch next?

The user groups provided a rich and detailed set of comments on many aspects of the system. The results on the participants' preferences for the available and potential features are shown in Figure 11. The conclusions that came out of the user study are detailed in the following subsections. First, we introduce the results regarding the perceived benefits of the proposed functionality, trying to address what the users wanted to do. Then, we discuss to whom they wanted to send the recommendations. Finally, we provide the main reasons why the participants used the system and an analysis on the preferred devices and interaction techniques.

5.1.1 *Fragmenting and Enriching Videos.* Most users liked the idea of enriching content, either to personalize their own stories or to personalize content when sharing video with friends and family. When personalizing their own content, some wanted to change clip titles, add explanatory notes, or change posters so that later they could remember what a particular clip contained. Also, several participants wanted to manage the structure of their personal content by creating new folders, changing folder titles, and organizing folder hierarchies.

Many users wanted to create their own scenes, either through fragmenting existing scenes or chapters or by defining a scene by specifying a start and end point. Some also wanted to define single points or bookmarks to allow quick access to certain points in the video. When making changes to the content by enriching or fragmenting it some participants were worried about two things: destroying the original structure and affecting the state of the content on the shared screen. Several participants wanted new scenes to be stored in separate areas so that existing chapter/scene structures were maintained. One participant suggested that his personalization should be kept in a different area such that it did not "confuse the original chapter structure."

Many participants were surprised to learn that not everybody had direct access to the content located on their personal devices. Because in the first version of the system the content sharing delivery functionality was not fully available, we introduced this situation to create a realistic condition and to further evaluate sharing between participants. It showed that many people approached the system with an assumption that they shared a common pool of content. In many cases, they shared their own screen first to point out content items to their friends which might be new to them and then use the recommendation functionality to share the item with that friend.

5.1.2 *Sending and Receiving Fragments of Videos.* Participants were generally very enthusiastic about the idea of sharing content. For the most part, they viewed this as something they would like to do with friends or family who did not live with them, although they could also see the value of sending content to people living in the same household if they were in another room or otherwise not concurrently present. Nevertheless, a system like this could profit from a simplified sharing mechanism for people that are colocated. Several participants mentioned that they wanted to recommend clips while out with a friend or at a friend's house so the friend could play the clip on their TV. Similar results have been reported elsewhere, when studying video consumption in mobile phones [O'Hara et al. 2007].

Most of the participants appreciated having a secondary display that allowed for browsing and annotating content, as well as sending and receiving fragments of videos. A scenario in which content was shared among users within one home did not have a large appeal to the participants in our study, but this might if the subjects were from families living in shared households.

5.1.3 *Rationale for Sharing.* When sharing videos with friends and family, many participants wanted to add written or audio comments to a recommended content item, either to provide more context to the recommendation (e.g., to explain why they were sending it or to ask the recipient to note something in the video). "I think people would really get into that 'check out this trailer—we've got to go and see this movie'". Some participants imagined using the annotation functionality to point to a particular character or object when recommending content. For example, one suggested that he might use it to circle an offside soccer player. In the current implementation, recommendations are not presented as incoming messages but as annotations to the existing content items. Many participants did not want these recommendations to just show up in their selection but they were interested in being alerted to the fact that a recommendation had arrived. A recommendation was seen as a message: knowing the sender was a key piece of information that helped gauge the value of the recommendation itself. In addition, we can see the desire to reciprocate recommendations as a potentially strong driver for the adoption of impromptu recommendations from the couch. As observed in other studies the social norms around gift-giving include the demand for reciprocity [Taylor and Harper 2002].

5.1.4 *Benefits of a Secondary Personal Display.* For many participants, a key advantage of the system was the fact that they were able to browse content and choose programs without interrupting

what was playing on the main screen. Thus, in a group viewing situation, everyone could browse for themselves without interrupting others. However, this aspect did not appeal to everyone. A number of participants did not want to browse while something was already showing.

In terms of the hand-held control device, our tests used UMPC-style tablet PCs. A number of people preferred a lighter device that supported one-handed operation. Some people expressed concern that the stylus would be easily lost. Both the preference for one-handed operation and the stylus concern suggest a preference for finger-based interaction. We expect that these problems would be eliminated with the introduction of iPhone-style control devices.

5.2 Belgian Study

A second study was carried out in Belgium approximately 10 months later, with a system containing extra functionality identified during the UK study, such as improved delivery across heterogeneous mobile telephone devices.

In the Belgian study, 12 groups of 2 to 5 people were recruited for the test. Each test session involved 1 group, lasted for 2 hours, and was video recorded. Each group consisted of participants that knew each other well, either as friends or as family members. In total 36 participants took part in the test, with ages ranging from 14 to 72. As for gender, 13 participants were male and 23 participants were female. The test sessions took place in a simulated livingroom and consisted of four main parts: an explanation of the system, a colocated test situation, a remote test situation, and a group interview. After the second and third part, questionnaires were filled in by each participant.

The first part of the test was used as an extensive training session for the system; the goal of this session was not to evaluate usability, but to ensure that all participants had the skills required to evaluate key application functionality. During the second part of the test, all members of the group stayed in the same room, and were asked to perform basic system-related functions. This included browsing through the available content, selecting items they wanted to share with someone, creating content clips, and optionally annotating these clips with text, audio, or graphic comments. Finally, they were asked to send the selected recommendations to other users. In the third part of the test, the group was split into two subgroups. One subgroup (sometimes a single person) stayed in the simulated livingroom, whereas the other subgroup (or single person) was led to a separate room. For this part of the test, the participants in the livingroom were asked to edit and annotate clips to send to the participants in the other location. This way, the participants knew their edited clips were actually received by someone else. Finally, the fourth part of the test consisted of a group interview lasting about 20 minutes, which covered topics such as the reasons for sending and annotating clips, the use of separate screens, or desires for extra system functionality.

When analyzing the results, the video observations and participant's answers were coded and clustered. Only those concepts that were observed in or mentioned by at least two groups were taken into account. Some issues were repeated by almost all groups, and were treated as more important when interpreting the results. In the discussion that follows, quotes from selected individuals are used to illustrate a greater concept mentioned by several participants.

Figure 12 shows the results of the answers to questions about how good or bad the participants rated annotating clips and having a secondary screen. The results show that most participants preferred annotating clips as well as having a secondary screen. The conclusions relevant to the main topics of this article are discussed in the following sections.

Similarly to the UK study, we address key issues regarding the system such as what do the users think about the provided functionality, the reasons why they recommended content, and their preferences regarding devices.

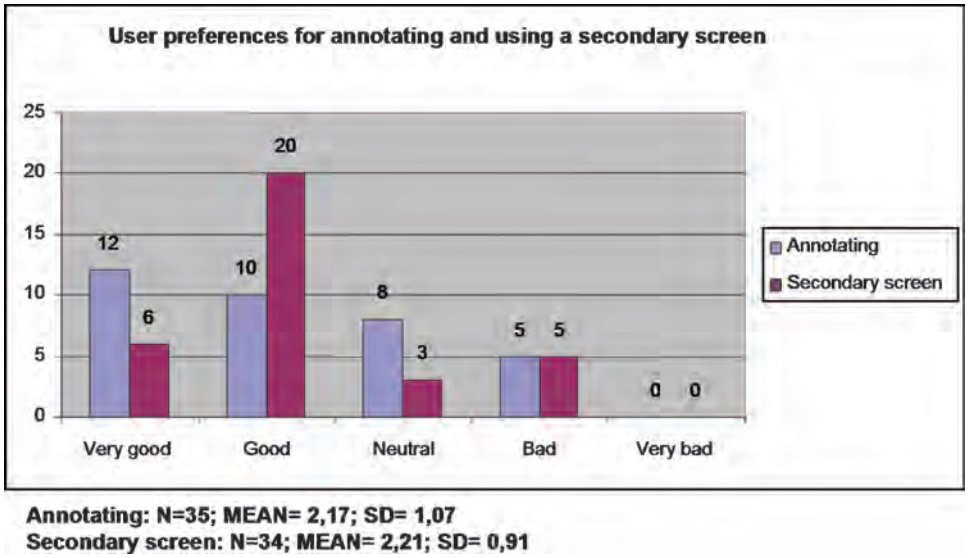

Fig. 12. User preferences for annotating and using a secondary screen.

5.2.1 *Annotating Clips.* The possibility of making annotations was positively perceived by most participants. Several people explicitly said they liked the personal aspect of making annotations. They said it allowed them to give private commentary or navigation suggestions. One person said it was the same as talking with one another while watching, even when not being at the same place or at the same time. Some participants said annotation would allow them to add a rationale for sending something, as sending the clip alone might be unclear. On the one hand, some participants said they wanted to add comments attached to the media, but not overlaid on the image. Several other people said they wanted to add audio comments at times when they had lots to say, since writing or texting these comments requires too much effort. Other reasons for using audio included giving a spoken introduction, laughing or singing along, creating voice-overs or telling the recipients when to pay attention.

While several participants mentioned that they would be willing to annotate while watching if the system were easy and natural to use, other participants preferred to take screenshots while watching and then annotate the clip afterwards.

5.2.2 *Rationale for Sharing.* Many participants were excited about the idea that you can send something you see immediately to other people. One person who was very skeptical during the initial briefing admitted changing his mind while using the system and found it fun because he could not only send indirect comments about a (part of a) program, but could also send excerpts from the program itself to others.

Several reasons for sending clips were mentioned by the participants, with similar categories as the ones identified for sharing photos made on a mobile telephone [Van House et al. 2005]. One reason can be classified as informative, such as sending something they knew to be of interest for someone. Sometimes, the clips were shared to highlight specific social activities: one participant wanted to send a recipe to his wife so that she could cook it for him. Another participant said she would send a clip from a home improvement program to her friend because that person was building a house. Several participants related using the system to their work situation, allowing them to send clips to colleagues for several practical reasons. For example, a teacher wanted content that could be used in class. Another

reason mentioned by several participants was for sending clips that other people might have missed or could not see because they were on holiday. They added that this could be done spontaneously, knowing that someone is abroad and might have missed it, or on request, people asking in advance to record a clip for them. Finally, another major reason why to share video clips was for maintaining relationships, when a clip reminded them of a person or a memory together, or as a birthday gift.

5.2.3 *Devices / Context.* The fact that a separate screen was used for creating, annotating, and sharing clips was considered beneficial because other viewers could continue to watch content on the primary screen without interruption. Two participants referred to how changing the settings of current video or DVD players disturbs the viewing experience, because a menu comes up on the screen, overlaying or even replacing the television program. During the observations, it was also clear that the private screen stimulated interaction between the participants. When one person was creating clips or annotating, the person sitting next to her looked over her shoulder to see what she was doing, comment on what she should do next, or even to point at the screen to indicate certain actions to be taken. The private device was often passed on to or claimed by the other person, so he could take over control.

Regarding the preferred device to receive the video clips, participants mentioned several factors such as clip length, content quality, and immediacy. A mobile phone was usually preferred for short clips with low content quality that can be watched in private, and that have an urgent character (high immediacy). Email received on a PC was preferred for medium-length clips of medium content quality, which may or may not have an urgent character but which also would be watched in private. Finally, television was seen as a device on which people would want to receive long, high-quality clips without an urgent character that can be watched in company. Some participants noted they wanted the mobile device to be used only for receiving notifications of new content, with the actual content being forwarded to more appropriate viewing devices such as their television or their PC depending on the situation. This result clearly aligns with the UK tests, in which a good mechanism for notifying recommendations is a must before deployment of this kind of products. For a better understanding of these results, it has to be noted that during the tests the receiving device, based on the link to the enriched video clip, launched the video player after reception.

5.2.4 *Privacy.* Privacy and presence control were two major concerns raised by the participants, as identified in previous research of social media systems [Huang et al. 2009]. One participant wanted to put a signature on the his clip, to make sure the comment was authentic and not spam. Others were more concerned about privacy in the context of delivery. For example, some were afraid that their recommendations, if going directly to the television, would be perceived as a disturbance. Several participants indicated they found it problematic if multiple people would be able to watch a clip intended for a specific person. One person said she was afraid that when sending a clip to her friend, her husband would watch it as well, while another person was unsure of delivering clips to her girlfriend if the potential existed that her family was sitting in the livingroom. As expected, people found email more private than the television set, so the large majority preferred to send the annotated video clips to either an email account or to a mobile phone, which has some implications in the overall architecture of the system.

6. CONCLUSIONS

The personal content management application described in this article is one example of how passive users can be allowed to interface more directly with the content they watch in informal settings. The hypotheses we wanted to validate with this work were: (1) that the end-user could become an active node in the multimedia delivery chain and (2) that the lifecycle of content does not have to end at consumption time, but that the moment of consumption can start a new interaction within the lifecycle.

These two hypotheses essentially extend the definition of *personal media* from a sense of directed selection to a sense of active involvement. We argue that this personalized media, in addition to being a personal view on a particular piece of content, is also a window on personal intentions within specific social communities. We feel that, by extending popular video sharing systems with video manipulation features such as video fragmentation, annotation, and enrichment, users can be empowered to interact with content beyond simple consumption.

Using qualitative evaluation methods, we gained several perspectives on end-user acceptance of our infrastructure and their preferences on potential system configurations. One clear result is that, if available, direct sharing of video clip fragments can be expected to be widely used. At the same time, sharing is not only about sharing a video fragment, but also about adding textual or audio commentaries, or highlighting objects or people. While several video sharing systems (such as YouTube) allow users to overlay message on video clips, these solutions typically do not provide the target personalization that we feel will encourage more intimate social interactions to be built around content. They also do not provide the content copyright protection that ensures that the baseline video is not changed or copied without the owner's permission.

A trend in online sharing has been to move Web content off the PC and on to the mobile telephone or the home video screen. In many ways, this was a seminal component of the move to user-contributed content. At the same time, however, we feel that current delivery systems do not integrate well within the home environment: They bring the Web's content model to the TV, but they are restricted by a classical single-point control model. This makes it difficult to synchronize media consumption across devices at home. Our system attempts a seamless integration in the home environment, providing the possibility of using primary and secondary screens for video consumption/control and video manipulation activities. As shown in the results from the user evaluations, users have a clear preference for systems that take into consideration their contextual situation, since sometimes they might be on the move or simply want to immerse themselves in a nice movie evening without wanting to be disturbed.

There are a number of lessons we have learned from the evaluation and implementation of our prototype which we feel will have implications for future implementation efforts of other researchers. From an architectural perspective, it is essential that a tight integration with existing external social networks be achieved. Even though most current social networks do not provide the desirable level of interoperability, they can provide identity management and presence awareness information that are essential for personalized sharing. In the near future, the generated recommendation can be sent to the virtual identity of the recipient and the social network infrastructure can be responsible then for routing the message to the most suitable device depending on the contextual situation of the user. This mechanism provides a solution for key concerns of the users regarding notification of the recommendation and privacy.

Even though our implementation considers the authoring process as an incidental activity done while watching, some of the participants preferred to simply bookmarking the selected timing information of the enrichment for later creation. This result has implications on the user interfaces needed for the kind of applications introduced in this article. They need to be adaptable and to dynamically accommodate to different user needs and requirements. One easy starting point is to take into consideration the narrative structure of the watched content, normally responsible for the user immersion, when displaying the user interface. We can imagine that complex narrative content requires a simple user action for bookmarking a timestamp, while less complex narratives can offer the option of enriching while watching. In the future, user level of attention gathered by sensors in the livingroom can be used for predicting the user behavior.

Personal archival is another issue that needs further study. Personal recommendations are inherently persistent, even though expiring timing mechanisms can be incorporated. Since they will be stored for

later retrieval, significant questions on preservation and consistency among different versions impose a number of challenges. Even though at a basic level such issues have been considered when modeling the recommendation message, investigation on the topics is required, possibly as an extension of management tools for emails with attachments.

More resources should be dedicated for a truly interoperable Web of media, in which original material can be universally located. Work in this direction has been initiated by the TV-Anytime consortium, but still popular content providers tend to restrict the consumption of their content to their own infrastructure, thus naturally limiting their potential viewers and possibly their revenues. The approach proposed in this article does not modify the original content, but only generates an overlaying and independent layer that it is later shared with others. Then, the recipient of the message has to access and pay for the underlying content so he can watch enriched material. In our opinion, initiatives like this constitute a window of opportunity for content providers, minimizing the risk of a Web populated with freely available and modified copies of their content.

The final intention of this article is to demonstrate that incidental authoring of media content is an important functionality that users will expect in a future media landscape. In order to provide an adequate infrastructure that respects the rights of all the players, including the users, and that can provide new revenue flows for the content providers, a set of architectural decisions is mandatory. The provision of better integration between devices, the support for structured multimedia containers that can encode the enrichments, a better interoperability with social networks, the development of predictive interfaces and context-aware notification mechanisms are some of the key lessons learned during the process.

In the future, we expect to continue our efforts to gain a deeper understanding of the types of content management and manipulation that can be support in a dynamic, social community. This includes new systems for managing recommendations and migrating content. We also expect an impact on media standards such as SMIL and TV-Anytime for encoding runtime behavior and persistent sharing of recommendations and annotations. The next step is to have a widely deployed version of our system for undergoing further studies that can shed some light on social usages of multimedia content. For such deployment underlying system issues such as diversity of devices support, content delivery and synchronization, and communication mechanisms need to be investigated. As shown in this article, while commenting on media is a daily activity, the research community still needs to solve a number of fundamental problems so it can be adequately mediated by digital devices.

ACKNOWLEDGMENTS

We are grateful for the constructive suggestions given by the anonymous reviewers and the issue editors.

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 SIGMM 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.

Bulterman, D. C. A., and Hardman, L. 2005. Structured multimedia authoring. *ACM Trans. Multimedia Comput. Commun. Appl. 1*, 1, 89–109.

Bulterman, D. C. A. and Rutledge, L. R. 2009. SMIL 3.0: *Interactive Multimedia for the Web, Mobile Devices and Daisy Talking Books*. Springer Verlag, Heidelberg/New York.

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. A., Geerts, D., Jansen, J., Knoche, H., and Seager, W. 2008. Enhancing social sharing of videos: Fragment, annotate, enrich, and share. In *Proceedings of the ACM SIGMM International Conference on Multimedia*, 11–20.

Cha, M., Kwak, H., Rodriguez, P., Ahn, Y.-Y., and Moon, S. 2007. I tube, YouTube, everybody tubes: Analyzing the world's largest user generated content video system. In *Proceedings of the ACM SIGCOMM Conference on Internet Measurement*, 1–14.

Chorianopoulos, K. 2007. Content-enriched communication supporting the social uses of TV. *J. Commun. Net. 6*, 1, 23–30.

Resende Costa, R. M., Ferreira Moreno, M., Ferreira Rodrigues, R., and Gomes Soares, L. F. 2006. Live editing of hypermedia documents. In *Proceedings of the ACM SIGWEB Symposium on Document Engineering*, 165–175.

Dearman, D., and Pierce, J. S. 2008. It's on my other computer!: Computing with multiple devices. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 767–776.

Gill, P., Arlitt, M., Li, Z., and Mahanti, A. 2007. YouTube traffic characterization: A view from the edge. In *Proceedings of the ACM SIGCOMM Conference on Internet Measurement*, 15–28.

Goularte, R., Moreira, E. S., and Pimentel, M. G. C. 2003. Structuring interactive TV documents. In *Proceedings of the ACM SIGWEB Symposium on Document Engineering*, 42–51.

Halvey, M. J., and Keane, M. T. 2007. Exploring social dynamics in online media sharing. In *Proceedings of the International Conference on the World Wide Web*, 1273–1274.

Hesselman, C., Cesar, P., Vaishnavi, I., Boussard, M., Kernchen, R., Meissner, S., Spedalieri, A., Sinfreu, A., and Raeck, C. 2008. Delivering interactive multimedia services in dynamic pervasive computing environments. In *Proceedings of the International Conference on Ambient Media Systems*.

Hua, X. S., and Li, S. 2006. Interactive video authoring and sharing based on two-layer templates. In *Proceedings of the ACM SIGMM International Workshop on Human-Centered Multimedia*, 65–74.

Huang, E.M., Harboe, G., Tullio, J., Novak, A., Massey, N., Metcalf, C.J., and Romano, G. 2009. Of social television comes home: A field study of communication choices and practices in TV-based text and voice chat. In *Proceedings of the ACM SIGCHI International Conference on Human Factors in Computing Systems*, 585–594.

Jokela, T., Lehikoinen, J. T., and Korhonen, H. 2008. Mobile multimedia presentation editor: Enabling creation of audio-visual stories on mobile devices. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 63–72.

Kindberg, T., Spasojevic, M., Fleck, R., and Sellen, A. 2005. I saw this and thought of you: Some social uses of camera phones. *Extended Abstracts on Human Factors in Computing Systems*, 1545–1548.

Kirk, D., Sellen, A., Harper, R., and Wood, K. 2007. Understanding videowork. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 61–70.

Kosch, H., Boszormenyi, L., Doller, M., Libsie, M., Schojer, P., Kofler, A. 2005. The life cycle of multimedia metadata. *IEEE Multimedia 12*, 1, 80–86.

Liu, J., Huang, Y., Li, D., Wu, F., and Li, B. 2007. A Web-based aggregated platform for user-contributed interactive media broadcasting. In *Proceedings of the ACM SIGMM International Conference on Multimedia*, 541–544.

Marlow, C., Naaman, M., Boyd, D., and Davis, M. 2006. HT06, tagging paper, taxonomy, Flickr, academic article, to read. In *Proceedings of the ACM SIGWEB International Conference on Hypertext and Hypermedia*, 31–40.

Metcalf, C., Harboe, G., Tullio, J., Massey, N., Romano, G., Huang, E. M., and Bentley, F. 2008. Examining presence and lightweight messaging in a social television experience. *ACM Trans. Multimedia Comput. Commun. Appl. 5*, 2, article 27.

Nathan, M., Harrison, C., Yarosh, S., Terveen, L., Stead, L., and Amento, B. 2008. CollaboraTV: Making television viewing social again. In *Proceedings of the International Conference on Designing Interactive User Experiences For TV and Video*, 85–94.

O'Hara, K., Mitchell, A. S., and Vorbau, A. 2007. Consuming video on mobile devices. In *Proceedings of the ACM SIGCHI International Conference on Human Factors in Computing*, 857–866.

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.

Pfeiffer, S., Parker, C., and Schremmer, C. 2003. Annodex: A simple architecture to enable hyperlinking, search and retrieval of time–continuous data on the Web. In *Proceedings of the ACM SIGMM International Workshop on Multimedia Information Retrieval*, 87–93.

Rutledge, L. and Schmitz, P. 2001. Improving media fragment integration in emerging Web formats. In *Proceedings of the ACM Multimedia Modeling Conference*, 147–166.

Sgouros, N. M., and Margaritis, A. 2007. Towards open source authoring and presentation of multimedia content. In *Proceedings of the ACM SIGMM Workshop on Human-Centered Multimedia*, 41–46.

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 ACM SIGMM International Workshop on Multimedia Information Retrieval*, 275–284.

Shaw, R., and Schmitz, P. 2006. Community annotation and remix: A research platform and pilot deployment. In *Proceedings of the ACM SIGMM International Workshop on Human-Centered Multimedia*, 89–98.

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.

Stamou, G., van Ossenbruggen, J., Pan, J. Z., Schreiber, G. 2006. Multimedia annotations on the semantic web. *IEEE Multimedia 13*, 1, 86–90.

Taylor, A. S., and Harper, R. 2002. Age-old practices in the 'new world': A study of gift-giving between teenage mobile phone users. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 439–446.

Van House, N., Davis, M., Ames, M., Finn, M., and Viswanathan, V. 2005. The uses of personal networked digital imaging: An empirical study of cameraphone photos and sharing. *Extended Abstracts on Human Factors in Computing Systems*, 1853–1856.

Received June 2009; accepted June 2009