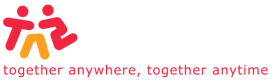
Socially-Aware Multimedia Authoring

Rodrigo Laiola Guimarães

ISBN: 978-94-6259-028-1

Typeset with Microsoft Word
Cover picture: TrendyCovers.com
Printed and bound by Ipskamp Drukkers, Amsterdam, The Netherlands



The work reported in this thesis has been carried out at the Centrum Wiskunde & Informatica (CWI), under the auspices of the group Distributed and Interactive Systems (SEN5). The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. ICT-2007-214793.

To the memory of my beloved mother, Maria Célia,
whose sacrifices, love, and dedication to family taught
me the most important lessons in my life and showed me
how to see the hand of God in all things.

Dedicado à memoria de minha amada mãe, Maria Célia,
cujo sacrifício, amor e dedicação me ensinou as mais
importantes lições da minha vida e me mostrou que
Deus age sobre todas as coisas.

VRIJE UNIVERSITEIT

Socially-Aware Multimedia Authoring

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. F.A. van der Duyn Schouten,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Exacte Wetenschappen
op dinsdag 28 januari 2014 om 13.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Rodrigo Laiola Guimarães

geboren te Vitória, Brazilië

promotor: prof.dr. D.C.A. Bulterman
copromotor: dr. P.S. Cesar

Contents

Acknowledgements

I wish to publicly acknowledge those whom I have been so fortunate to have met or worked with during this long journey.

My Ph.D. thesis is, without a doubt, a team effort. And I owe a very large debt to those who graciously conducted me to make this book a reality. First, I would like to thank my mentor and dear friend Dick Bulterman. It is impossible to describe in a few words my admiration and how much I have learned with him during this period. In stressful times, he would always remind me that "we do what we do (research) simply for the fun of it!". And he was right! Second, I would like to gratefully and respectfully thank my as important advisor, colleague and "big brother" Pablo Cesar. Thank you for doing such a patient and sterling job of conducting me through this process. I guess you should have gotten paid double, not only for shaping me as a researcher, but also as my psychologist ☺. It was an honor and a pleasure to work with both of you guys. Thank you very much once again, and I truly hope we can cross paths and work together in the near future.

I also would like to give a special *thank you* to the members of my reading committee – Anton Eliens, Frank van Harmelen, Janet Murray, Maria da Graça Pimentel and Susanne Boll – for taking time to carefully read my thesis and for all the helpful comments. Your expert endorsement added a great value to this work.

My time at CWI was enjoyable and many people contributed to this. I would like to take a moment to thank some colleagues and friends: Alexandra, Behnaz, Benjamin, Bram, Christoph, David, Emma, Enav, Eleni, Floor, Georgiana, Henke, Irma, Inken, Jannis, José, Karin, Krzysztof, Lara, Léon, Li, Mahdi, Margriet, Milad, Minnie, Natallia (& Mark), Rob, Sara, Shashi, Stephanie, Susanne, Willem, Young-Joo and any other name that might have slipped my mind. I also would like to thank all my colleagues that were part of SEN5 at certain point in time: Bo, Chen, Diogo Martins, Diogo Pedrosa (& Uaiana), Fons, Jack, Ivan, Kees, Ketan, Marcio (& Suzana), Manzato, Marwin, Nelma, Pia, Rufael, Simon and Steven. A big *thank you* to my all time best friends André, Fernando Mario, Ishan and Wagner. We have been through great moments together. Last, but not least, special

thanks to my friend Bikkie, whose smile and cheerful *Bom dia* (good morning) made my days much happier!

During my Ph.D., I had the opportunity to work in the pan-European project Together Anywhere, Together Anytime (TA2). I would like to take this opportunity to thank in particular: Peter Stollenmayer (Eurescom); Ian Kegel, Doug Williams, Tim Stevens, Roland Craigie, Peter Glenn and Amanda Gower (British Telecom); Marian Ursu, Vilmos Zsombori and Michael Frantzis (Goldsmiths College London); Peter Ljungstrand (Interactive Institute); Rene Kaiser (Joanneum Research); Danil Korchagin (IDIAP); Marc Steen and Joke Kort (TNO); Orlando Verde (Alcatel-Lucent Belgium). Special thanks to all the users, families and schools (in the Netherlands and the UK) involved in the evaluation process discussed in this thesis. Due to privacy agreements I will not disclose any particular name.

I also had the pleasure to visit the Georgia Institute of Technology (GA Tech) and intern at Samsung Research America (SRA) during my Ph.D.. I would like to gratefully thank the people whom kindly hosted me in these opportunities: Janet Murray, Sergio Goldenberg (GA Tech); Henry, Ellen and Oma (Bulterman Family); Alan Messer, Alex Bentley, Edwin, George Hsieh, Glenn Algie, Joakim Soderberg, Jun Nishimura, Mina Yoo, Raghuram Reddy, Scott Pan and Shai Kumar (SRA).

While working at CWI, I attended several conferences, workshops and summer schools, where I met great and interesting people. I would like to remind some names: Andrei Bursuc, Cássio Prazeres, Carlos Salles, Cesar Teixeira, Cyril Concolato, David Ayman, David Geerts, Diana Arellano, Ethan Munson, Fabien Cazenave, Florian Stegmaier, Frank Nack, George Lekakos, Hendrik Knoche, Joel dos Santos, Konstantinos Chorianopoulos, Marianna Obrist, Miriam Redi, Mo Rabbath, Pawel Filipczuk, Romualdo Costa, Rudinei Goularte and Sam Davies.

I also would like to take this moment to thank all the inspiring teachers I had along the years, in particular to Flávio Miguel Varejão, Luiz Fernando Gomes Soares, José Gonçalves Pereira Filho, Maria da Graça Pimentel and Celso Saibel.

A tremendous *dank u wel* to my dear friends from A.V.V. Swift – Chudi, Danny, Dennis, Edwin, Fred, Gio, Jesse, Jim, Johan, Kevin, Lodi, Marcel, Marius, Nordin, Patti, Paul, Ramalho, Remi, Romain, Thijs and anyone else who I might have forgotten. Being part of the *Zondag 5* family provided me a regular scape from work and made me feel closer to home. I will miss playing football with you.

I also would like to thank some friends that I had the pleasure to meet in Amsterdam in the most different occasions: Alex, Daniel, Edson, Frans, Hugo,

Jacquelien, Jordan, Malcher, Martinelli, Renato, Ruben, Valéria, Vinicius, Willemijn, Yoko and Zambon. A big *bedankt* to my Dutch friend Michiel Zwerus for the inspiring discussions, pool games and great dinners.

My entire family has been a steady source of support during my entire career. They don't quite get what exactly I have been doing during all this time, but they are still so encouraging. I would like to mention my parents, Antônio e Maria Célia, and my sisters Roberta and Rafaela. Not to forget my parents-in-law, José Braz and Maria do Carmo. I would like to specially thank my uncle João Ademir Laiola (aka João Ameixa), whose encouragement always gave me inspiration to keep going.

Finally, without my girlfriend, partner and best friend, Cristina, I would not have been able to manage throughout this journey. Her loving support, unflagging patience and zealous pulled me together at the most difficult turns along the way. I am so blessed for having you on my side, and thank you very much for everything. I love you, *Namorada*!

1

Introduction

Nearly a decade ago at an ACM SIGMM[1] retreat, one of the grand challenges to the multimedia research community was to develop media authoring tools that would make creating complex media titles as easy as using a WYSIWYG (*What You See Is What You Get*) word processing system [41]. Since that time, a number of consumer-level video editing tools have been developed that would lead a casual observer to believe that multimedia authoring is a solved problem: using tools like *iMovie*[2] or *Windows Movie Maker* (or even more sophisticated tools such as *Adobe Premiere* or *Final Cut Pro*), even relatively novice video editors can match their talents with the likes of Sergei Eisenstein (see Figure 1.1).

The process was further simplified by modern content capture tools, such as smartphones, in which recording, (simple) editing and integrated uploading were combined into a single task. In many ways, video editing has been reduced to transferring content taken from a (personal) camera to a computer, throwing out frames that are unwanted, and uploading the resulting production to a video sharing site. While it is indisputable that media capture and sharing is much easier than at any time in the past, we wonder if the resulting products of such authoring interfaces have provided any significant advances for the viewers of media content. It is even questionable if there have been significant advances for content authors.

[1] SIGMM is the Association for Computing Machinery's Special Interest Group on Multimedia, which specializes in the field of multimedia computing, from underlying technologies to applications, theory to practice, and servers to networks to devices.

[2] The services and technologies mentioned in this thesis, if unknown, could very easily be identified via a simple online search, therefore they will not be Web-referenced.

Recent data suggests not. In spite of the ubiquity of video cameras and the growth in video viewing on social networking sites, about 82% of Internet users have never uploaded even a single video [45]. Although most *YouTube* uploads are amateur content, professional videos are preferred to amateur productions online [38]. From the perspective of personal videos, the problem of creating and sharing content has several dimensions. At a lower level of abstraction, video is not semantically linked; therefore, searching and selecting the desired piece of content to share can become too laborious. At a high-level of abstraction, creating compelling videos – videos that meet the needs and desires of the viewer, not only the producer – is a complex task [69]. Viewers generally expect professionally produced content (in terms of shot selection, story pacing and logical narrative), which most amateur users cannot provide.

Although a number of research efforts have addressed content creation from different perspectives [6][19][26][61], based on user studies [57][63] we observed that traditional authoring tools and current social media services fail to address the interpersonal relationships for sharing media that is personal and important to families and small social groups. Our assumption is validated by other studies [24], which concluded that social media applications like *Facebook* do not take into account the interpersonal tie strength of the users. Thus, we can conclude that the current media landscape demands a revision of traditional research on multimedia authoring to empower users in recalling and sharing personal media experiences with friends and family. This discussion leads to the following question:

Main Question *Is a new multimedia authoring paradigm required to enable end-users[3] to share more personal media within their social circle?*

During the past years our research work has focused on the study of *socially-aware multimedia authoring*. Working with a group of users at local high schools in two different countries (UK and the Netherlands), the process involved research on different facets related to the creation and sharing of multimedia artifacts composed of personal videos. Apart from the underling mechanisms for navigating and reusing personal content, this thesis work argues that a new paradigm, *socially-aware multimedia authoring*, is necessary to better fit end-users' needs. One important aspect of our work is that we decided to follow an interdisciplinary

[3] The terms 'end-user' and 'user' will be used interchangeably in this thesis to describe regular people who operate computer software with minimal technical expertise or previous training.

Figure 1.1. From the authoring perspective, the main challenge has been to make content creation a manageable process.

approach in which both technology and social issues were addressed. As illustrated in Figure 1.2, the core of our methodology integrates knowledge from user-centered design (e.g., need assessment, iterative prototyping and user evaluation) and document engineering.

The remaining of this chapter is organized as follows. In Section 1.1 the main question is split into a number of supportive research questions, which form the main focus of this thesis. Then, the contribution of the thesis is detailed in Section 1.2. Section 1.3 presents the thesis outline and summary of each chapter contribution with respective supportive material. Finally, Section 1.4 overviews the related work, contextualizing the research problem.

Figure 1.2. Timeline of the implementations and evaluations of our system.

1.1 Research Questions

In this thesis, we consider a socially-aware multimedia authoring framework for personalizing video stories from a collection of community assets. The high-level architecture of our framework is sketched in Figure 1.3. The input material includes the video clips that parents agreed to upload, together with a master track recorded by the school. By contributing assets in a shared video repository, each participant gives permission to reuse their own contributions within the community. It is assumed that each participant has the rights to contribute their own material. Privacy and a protected scope for sharing is a key component of our framework. Each media item is automatically associated with the person who uploaded it, and there are mechanisms for participants to restrict sharing of certain clips. Participants can use their credentials for navigating the repository – those parts allowed to them – and for creating and sharing different video compilations intended for different people.

At capturing time, there are no specific filming requirements for users. They can record what they wish using their own camera equipment. The goal is to recreate a realistic situation, in which friends and families are recording at a school concert. This flexibility comes at a cost, however, since most existing solutions that work well in analyzing audiovisual material are not that useful for our use case. As indicated in [59], handling user-generated content is challenging, since it is recorded using a variety of devices (e.g., mobile phones), the quality and lightning are not optimal, and the length of the clips is not standard.

Figure 1.4 enlists four main stages (with the respective application services) that compose the socially-aware multimedia authoring workflow proposed in this thesis: *Capture and Processing*, *Access and Navigation*, *Creation and Production*, and *Content Enrichment*. We should not forget the importance of looking at the social aspects around personal media in this workflow. The key research questions emerging from each of the four stages are presented below. But first, we take a look into the social requirements.

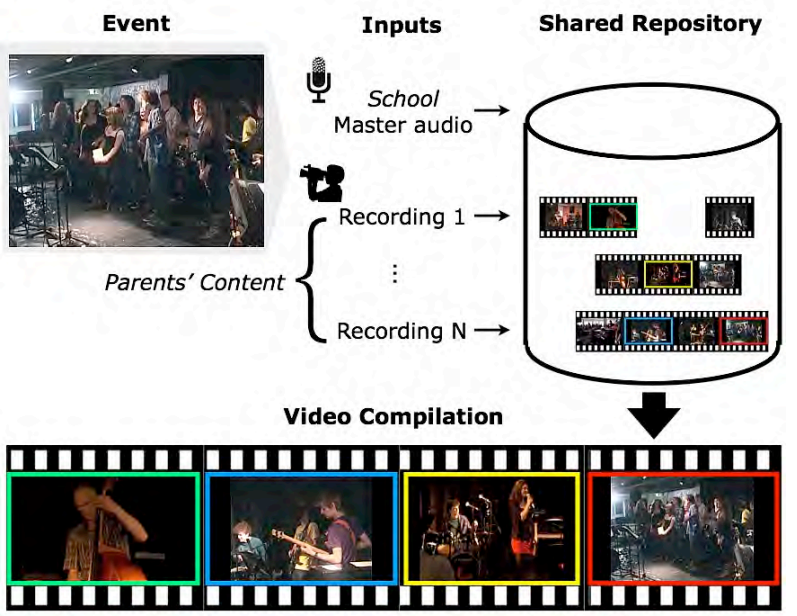


Figure 1.3. High-level architecture of socially-aware multimedia authoring systems.

1.1.1 Social Aspects

Recognizing the importance of looking at personal media as a cornerstone for sharing family experiences [72], the intention of our research is to understand the intersection of social multimedia and social interactions in an asynchronous communication context [37]. We are interested in personal media, and how these can become memory artifacts: the content around which conversation happens. We aim to help small groups of people (such as a family, a school class or a sporting club) viewing, creating, and sharing personalized multimedia. From the technical perspective, a system should combine the benefits of personal focus – knowing whom you are talking with – within the context of temporally asynchronous and spatially separated social meeting spaces.

Sociological theories [24] and user-centered approaches [3][25] have tackled different aspects of the multimedia workflow. For instance, human-centered efforts

explore video-mediated communication to share watching videos together over the distance [4]. Similar to us, other studies investigate what people do with media in an asynchronous context, balancing the preponderance of techno-centric work with appropriate user-centric insight [19]. In our work we pay special attention to social theories and human-centered methodologies. Our work, which is predicated on the intrinsic desire to strengthen existing strong ties among people, tackles different aspects of the socially-aware multimedia workflow. That said, the following research question arises:

Question 1.1 *Can a socially-aware multimedia authoring system be defined in terms of existing social science theories and human-centered processes, and if so, which?*

1.1.2 Capture and Processing

The research question introduced above puts the accent on the social aspects around personal media. While knowledge from online social networks could be mined to determine the strength of ties among people [24], user interest and sentiment analysis also could be used to facilitate media annotation and content

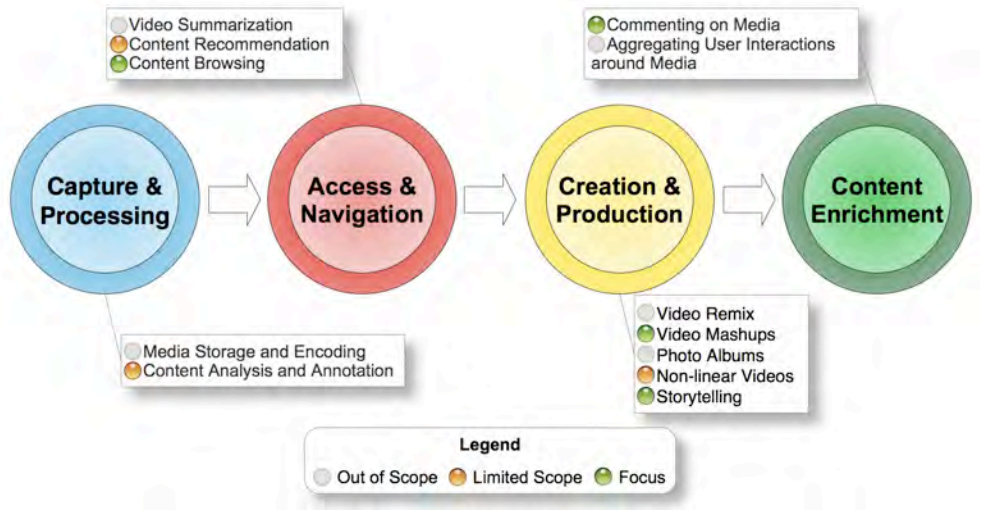


Figure 1.4. Multimedia authoring workflow and application services.

selection [3]. Given the characteristics of end-user content, in this thesis we have chosen to focus on studying the implicit social practices during video capture within groups of people with strong ties. Our assumption, which could be compared to research in the domain of user modeling [29], is that users' recording behavior can provide useful insights to better understand the interpersonal relationships. The attempt to 'understand' end-users and their social practices is just the first step in the socially-aware multimedia authoring workflow. More important are the indications that this new paradigm brings an improvement over the state of the art in multimedia authoring. In this direction, the research question we have is the following:

Question 1.2 *Does the functionality provided by a socially-aware multimedia authoring system provide an identifiable improvement over traditional authoring and sharing solutions? If so, how can these improvements be validated?*

1.1.3 Access and Navigation

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

While in terms of user experience, user interfaces are important; even more is the underlying interaction design and recommendation mechanisms. In particular, we are interested in technological solutions that can help users accessing and navigating media content with which they have social affinity. Our work acknowledges previous research efforts in video abstraction/summarization [7], content recommendation [40], synchronization and organization of user-generated content from popular music events [44] and home video management and navigation [28]. However, we go a step further by integrating knowledge of the social relationships to improve content searching and selection by individual users of a shared media repository. With this in mind, we ask the research question:

Question 1.3 *Does a socially-aware video exploration system provide an identifiable improvement over current approaches for accessing and navigating a repository of shared media?*

1.1.4 Creation and Production

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

A number of research efforts have addressed this problem by focusing on community video remix [14], automatic generation of video mashups from YouTube content [59], social creation of photo albums [58] and configurable and interactive storytelling [49][52]. The main difference of our work lays on the fact that we do not aim at providing a complete description of an event based on the characteristics of individual media fragments, but personalized video stories (narratives) based on the social bonds between people. Convenience and personal effort are also important factors to consider when generating such narratives. In this context, the research question we have is:

Question 1.4 *Where is the balance between automatic and manual processes when authoring personalized narratives users care about?*

1.1.5 Content Enrichment

One of the foundations of socially-aware multimedia is that media can take on new meaning based on the insights of downstream viewers. As an example, consider end-user generated comments. They have the potential to enrich and transform the media viewing experience by allowing users to express themselves and interact with others. Currently, media commenting is supported on an overly coarse level. Still, the lack of a embedded 'media message' in most personal media content actually presents the viewer of such media with a golden opportunity to superimpose his/her own meaning on top (physically or logically) of the content provided by the media object. We believe that a socially-aware system should enable content enrichment beyond 'likes' and out-of-band text comments.

The analysis of user-generated comments around media has resulted in innovative work on the semantic and temporal structure of media events [13], user commenting patterns in video on demand [25] and in live video streaming platforms [34]. Not to forget studies on the aggregated behavior of people and

social media. Examples include motivations behind tagging in *Flickr* [48] and location-aware photo sharing systems [18]. This thesis acknowledges all these efforts for better understanding users and media. But instead of focusing on the aggregation of user interactions around media, we investigate solutions that allow any user – not necessarily the author – to incrementally add personalized comments within multimedia artifacts. By personalized we mean comments that could be used to highlight interesting things for other viewers, e.g., to make a point about a particular event within a video. This discussion leads to the following question:

Question 1.5 *Does the support for timed end-user commenting within pre-authored narratives provide an identifiable improvement over current media commenting approaches?*

1.2 Our Aim

This thesis investigates mechanisms and principles for togetherness and social connectivity around personal media. The main contribution lays on a new paradigm, *socially-aware multimedia authoring*, which empowers users in telling stories and commenting on personal media artifacts. The work has been evaluated through prototype tools that allow users to explore, create, enrich and share rich multimedia artifacts. Results from our evaluation process provide useful insights into how a socially-aware multimedia authoring and sharing system should be designed and architected, for helping users in recalling personal memories and in nurturing their close circle relationships. Our experimental methodology aims at meeting the requirements needed for social communities that are not addressed by traditional authoring and sharing applications. During this process the intention was not to focus on a specific piece of software, but to take a broader look at the process and its implications. The final goal is to reformulate the research problem of multimedia authoring by emphasizing the importance of the social relationships among casual media authors, featured subjects and recipients of the media.

1.3 Thesis Outline and Summary of the Contributions

We summarize below the content and main contributions of each chapter.

Chapter 2 sets the stage by presenting a community video use case in which the social relationships between the people involved plays an essential role. Then, we detail our user-centered methodology, which involved requirements gathering, concert recordings, iterative prototyping and user evaluation. Motivated by social theories, preliminary interviews/focus groups and a survey research about social practices around personal videos, we identify key requirements and specify guidelines for realizing socially-aware multimedia authoring systems. Finally, we report on a long-term evaluation process that validates our approach and shows that socially-aware multimedia authoring is a valid alternative for social interactions when apart. The contributions of this chapter, which directly respond research *Question 1.1* and research *Question 1.2*, include:

- Introduction of a community video use case and motivation of socially-aware multimedia authoring;
- Description of the user-centered methodology followed in this thesis;
- Identification of requirements and specification of general guidelines for realizing socially-aware multimedia authoring systems; and
- Discussion about a 4-year evaluation process that includes the validation of the proposed socially-aware multimedia authoring framework.

This chapter is based on the following papers:

R.L. Guimarães, P. Cesar, D.C.A. Bulterman, V. Zsombori, and I. Kegel. 2011. Creating personalized memories from social events: community-based support for multi-camera recordings of school concerts. In Proceedings of the 19th ACM international conference on Multimedia (MM '11). ACM, New York, NY, USA,303-312.DOI=10.1145/2072298.2072339 http://doi.acm.org/10.1145/2072298.2072339. (17% acceptance rate)

R.L. Guimarães, P. Cesar, D.C.A. Bulterman, I. Kegel, and P. Ljungstrand. 2011. Social Practices around Personal Videos using the Web. In Proceedings of the ACM Web Science Conference (WebSci '11). Available at http://journal.webscience.org/437/ (15% acceptance rate)

Chapter 3 considers the development of innovative mechanisms to enable users to browse and navigate a repository of shared media. Context-aware user interfaces and filtering mechanisms are proposed by taking into account

relationships between users of the system and subjects featured in the videos. This chapter also discusses the importance of semantic annotations to describe personal media. Our approach is then compared to traditional (and less individual) media exploration tools. The contributions of this chapter, which directly address research *Question 1.3*, can be summarized as follows:

- Design and evaluation of an initial interface to facilitate the personalized exploration of a repository of shared media;
- Design and implementation of a new browsing interface based on the requirements elicited in the initial evaluation process; and
- User evaluation of the new interface, demonstrating that, when compared to traditional approaches, we have improved the ability to explore videos users care about, among a pool containing parent-made recordings.

This chapter is based on the following paper:

> *D.C. Pedrosa, R.L. Guimarães, P. Cesar and D.C.A. Bulterman. 2013. Designing Socially-Aware Video Exploration: A Case Study Using School Concert Assets. In Proceedings of the 17th International Academic MindTrek Conference: Making Sense of Converging Media (MindTrek '13).*

Chapter 4 compares automatic approaches for generating video stories (or media artifacts) from user-generated content with more manual mechanisms to reflect personal effort and intimacy. Our findings, which directly relates to research *Question 1.4*, indicate that the balanced combination of manual and automatic processes will be the basis for authoring tools that better fit end-users' needs. The contributions of this chapter are summarized below:

- Two-phased design, implementation and user evaluation of an authoring system to create personalized video stories from community assets; and
- Discussion about the benefits of a compromise between automatic and manual processes when creating personalized video artifacts.

This chapter contains extracts from the following papers:

R.L. Guimarães, P. Cesar and D.C.A. Bulterman. 2013. Personalized Presentations from Community Assets. In Proceedings of the 19th Brazilian Symposium on Multimedia and the Web (WebMedia '13). ACM, New York, NY, USA, 257-264. DOI=10.1145/2526188.2526208 http://doi.acm.org/10.1145/2526188.2526208 (33% acceptance rate) [Won, best multimedia paper]

R.L. Guimarães. Automatic and manual processes in end-user multimedia authoring tools: where is the balance?. In Proceedings of the international conference on Multimedia (MM '10). ACM, New York, NY, USA, 1699-1700. DOI=10.1145/1873951.1874327 http://doi.acm.org/10.1145/1873951.1874327

V. Zsombori, M. Frantzis, R.L. Guimarães, M.F. Ursu, P. Cesar, I. Kegel, R. Craigie, and D.C.A. Bulterman. 2011. Automatic generation of video narratives from shared UGC. In Proceedings of the 22nd ACM conference on Hypertext and hypermedia (HT '11). ACM, New York, NY, USA, 325-334. DOI=10.1145/1995966.1996009 http://doi.acm.org/10.1145/1995966.1996009. (34% acceptance rate) [Nominated, best paper/best newcomer]

Chapter 5 presents mechanisms to support end-user commenting and enrichment of pre-authored video stories. This approach is used as a way to communicate the viewer's personal view by highlighting a particular event that might be interesting to his/her social circle. This chapter, which directly responds research *Question 1.5*, brings the following contributions:

- Motivation based on a survey research about media consumption and commenting habits of a group of Internet users;
- Specification and description of temporal document transformations that allow end-users to create and share personalized timed text comments within third-party online videos;
- Design and implementation of a video commenting tool that realizes such document transformations; and
- User evaluation showing that users appreciated the functionalities of our system and would use it to communicate.

This chapter is based on the following papers:

R.L. Guimarães, P. Cesar, and D.C.A. Bulterman. 2012. "Let me comment on your video": supporting personalized end-user comments within third-party online videos. In Proceedings of the 18th Brazilian Symposium on Multimedia and the Web (WebMedia '12). ACM, New York, NY, USA, 253-260. DOI=10.1145/2382636.2382690 http://doi.acm.org/10.1145/2382636.2382690 (30% acceptance rate)

R.L. Guimarães, P. Cesar, and D.C.A. Bulterman. 2010. Creating and sharing personalized time-based annotations of videos on the web. In Proceedings of the 10th ACM symposium on Document engineering (DocEng '10). ACM, New York, NY, USA, 27-36. DOI=10.1145/1860559.1860567 http://doi.acm.org/10.1145/1860559.1860567 (31% acceptance rate)

Chapter 6 dedicates to open-ended questions and concluding remarks.

This chapter contains extracts from the following article:

D.C.A. Bulterman, P. Cesar and R.L. Guimarães. 2013. Socially-Aware Multimedia Authoring: Past, Present and Future. ACM Transactions on Multimedia Computing, Communications and Applications (TOMCCAP), Volume 9, Issue 1s, Article 35 (October 2013), 23 pages. DOI=10.1145/2491893 http://doi.acm.org/10.1145/2491893

1.4 Related Work

We frame this work within a historical perspective on three areas: conventional authoring systems, interactive storytelling and video mashups/content repurposing.

1.4.1 Conventional Authoring Systems

At the 1993 ACM Multimedia conference, Hardman et al. [43] presented a paper on structured multimedia authoring. Just over a decade later, this study was revised for the initial issue for ACM TOMCCAP [16]. At that time, multimedia authoring was seen by many as a seminal topic within the research community. As described in these publications, several paradigms existed for compositing (or binding) media objects, including:

- *Structure-based composition:* composition where the (often hierarchical) logical structure of the components server as the basis for generating a particular presentation instance timeline;
- *Timeline-based composition:* composition in which a particular presentation instance determines the content relationships among objects;
- *Graph-based composition:* composition in which the relationships among objects have cause/effect relationships, but limited logical structure; and
- *Script-based composition:* composition where the inherent logical structure of elements is hidden as side effects of a procedural execution model.

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

a) The content alternatives available;
b) The interests (and attention spans) of the intended audience; and
c) The formal or informal narrative and cinematographic principles required to build a compelling story.

While significant steps have been made in better understanding the encoding of narrative structures [32], the management of content and the management of viewer-driven interests provide fruitful areas for new work. We argue that there are two primary reasons that personal content viewers are unresponsive to non-professional content. The first reason is that the opportunity to home-editors represented by b) is largely unexploited by formal authoring systems. In many professional editing situations, all three of these aspects have been well understood, albeit for b) at an aggregate level of detail. For more personal content, home editors would seem to have a tremendous advantage: they typically know the person or persons for whom a particular content object is being created. Sometimes the

intended audience is relatively diffuse (such as one's 1,000 closest Facebook friends), but other times it can be highly focused: the grandmother of a young high school musician. The second reason that personal content viewers are unresponsive to non-professional content is that conventional formal authoring systems maintain a push model of content rather than a pull model, in which a content viewer in intimately involved in the process of content selection and personalization. This means that the author/editor determines all of the choices, with little infrastructure support for end-user personalization at the detail level.

1.4.2 Interactive Storytelling

During the past decade, various Artificial Intelligence (AI) approaches have been suggested for the creation of configurable and interactive storytelling [49][54]. A main thread of investigation has so far focused on generated content, often involving intelligent animated characters (e.g., Ibanez et al. [35]). Not to forget the use of interactive video as a basis for scenario-driven interactive tours, with additional mini-games for elaborating on specific topics or tasks that arise during exploration process [2]. Another representative example is Vox Populi [68], in which rhetorical documentaries are created from a pool of media fragments, and the Narrative Structure Language (NSL), a production-independent framework for the authoring and delivery of configurable and interactive video narratives [52]. More recently, a system capable of creating different story variants from a baseline video was presented [5].

In general, these systems generate sequencing video shots, while maintaining local video consistency. In order to support the automated generation of the interactive story, extensive use of metadata annotations on individual media objects is made. These systems have been applied to professionally produced media content, using well-defined (and generic) content and story descriptions. Our view on socially-aware multimedia authoring differs from typical interactive storytelling approaches in two important ways. First, the community content that we consider is not professionally produced and annotated. While we provide a reasonable degree of person and object recognition, the poor lighting and overall moderate quality of the content often requires user intervention to classify and locate content fragments. A second difference is that, although we focus on storytelling, we explore this concept in the context of repositories of UGC (*User-Generated Content*). There is still a structured representation of an overall interactive story space, but there is no control over the way the content is captured. The content

structures that can be made and exploited are only those emerging from the structure of the covered event itself.

1.4.3 Community Video Mashups and Content Repurposing

A second thread of more general story development is represented by work on video mashups and content repurposing. In this respect, it is interesting to note the current shift from local-based home videos management systems [28][65] to global-based video sharing Internet services.

Recent works [39][44] describe frameworks to synchronize and organize user-contributed content from live music events, creating an improved representation of the event that builds on the automatic content match. Shrestha et al. [59] report on an application for creating mashup videos from YouTube recordings of concerts. They present a number of content management mechanisms (e.g., temporal alignment and content quality assessment) that then are used for creating a multi-camera mashup. Saini et al. [53] go a step further by incorporating history-based diversity measurement, state-based video editing rules, and view quality in automated video mashup generations. Naci and Hanjalic [71] report on a video interaction environment for browsing records in music concerts, in which the underlying automatic analyzer extracts the instrumental solos and applause sections in the concert videos, and also the level of excitement along the performances. Lately, crowdsourcing has been proven to be a good ally for content analysis. For example, fans of a band can be useful for improving content retrieval mechanisms, where a video search engine allows for user-provided feedback to improve, extend, and share, automatically detected results in concepts from video footage recorded during a rock n' roll festival [11]. Our work builds on previous findings in event modeling [74] and identification [30][31], and video abstraction/summarization [7]. The main difference lays on the fact that we do not aim at providing a complete description of the shared event, but a better understanding of how community media can serve individual needs. Other interesting works propose a community video remixing tool [14], a video repurposing tool [66] and a video enrichment system that enable reciprocity [56]. In this direction, we should mention current practices around news stories, where users can reuse fragments of video clips for expressing opinions [46].

When compared with all these approaches, socially-aware multimedia authoring intends to help end-users generate stories in which social bonds between people play a major role. The previous approaches did not take into consideration

the case in which video authors and the people depicted in the videos are closely related. Similar to us, recent work has proposed a media sharing application that takes into account the interpersonal ties. This tool is capable of producing audio-visual media shows based on events, people, locations, and time [75]. In comparison to our work, this application does not allow for the creation of a narrative-based story based on multi-camera community recordings.

2

Personalized Memories of Social Events: Studying Asynchronous Togetherness[1]

The place: the Exhibition Hall in Prague. The date: August 23, 2009. *Radiohead* is about to start their concert. The band invites fans to capture personal videos, distributing 50 Flip cameras. After the concert the cameras are then collected, and the videos are post-processed along with Radiohead's audio masters. The resulting DVD[2] captures the concert from the viewpoint of the fans, making it more immersive and proximal than typical concert productions.

The concert of Radiohead typifies a shift in the way music concerts – and other social events – are being captured, edited, and remembered. In the past, professionals created a full-featured video, often structured according to a generic and anonymous narrative. Today, advances in non-professional devices are making each attendee a potential cameraperson who can easily upload personalized

[1] This chapter is based on the following papers:

R.L. Guimarães, P. Cesar, D.C.A. Bulterman, V. Zsombori, and I. Kegel. 2011. Creating personalized memories from social events: community-based support for multi-camera recordings of school concerts. In Proceedings of the 19th ACM international conference on Multimedia (MM '11). ACM, New York, NY, USA,303-312.DOI=10.1145/2072298.2072339 http://doi.acm.org/10.1145/2072298.2072339. (17% acceptance rate)

R.L. Guimarães, P. Cesar, D.C.A. Bulterman, I. Kegel, and P. Ljungstrand. 2011. Social Practices around Personal Videos using the Web. In Proceedings of the ACM Web Science Conference (WebSci '11). Available at http://journal.webscience.org/437/ (15% acceptance rate)

[2] Available at http://radiohead-prague.nataly.fr. Last access on May 15th 2013.

material to the Web, mostly as collections of raw-cut or semi-edited fragments. From the multimedia research perspective, this shift makes us reflect and reconsider traditional models for content analysis, authoring, and sharing.

This thesis considers the case in which performers and the audience belong to the same social circle (e.g., parents, siblings and classmates at a typical school concert). Each participating member of the audience records content for personal use, but they also capture content of potential group interest. This content may be interesting to the group for several reasons: it may break the monotony of a single camera viewpoint, it may provide alternative (and better) content for events of interest during the concert (solos, introductions, bloopers), or it may provide additional views of events that were not captured by a person's own camera. It is important to understand that the decision to use substitute or additional content will be made in the particular context of each user separately: the father of the trombone player is not necessarily interested in the content made by the mother of the bass player *unless* that content is directly relevant for the father's needs. Put another way, by integrating knowledge of the structure of the social relationships within the group, content classification can be improved and content searching and selection by individual users can be made more effective.

In order to understand the role of the social network among group members in a multi-camera setting, consider the comparison presented in Table 2.1. This table compares the use of multi-camera content in three situations: by a (professional) video crew creating an archival production, by a collection of anonymous users contributing to a conventional user-generated content mashup, and finally within a defined social circle as input for differentiated personal videos. (Semi-) Professional DVD-style productions often follow a well-defined narrative model implemented by a human director, and are created to capture the essence of the event. Anonymous user-generated content mashups are created from ad-hoc content collections, often based on the content classification methods [44][59]. In socially-aware communities, friends and family members capture, edit and share videos of small-scale social events with the main purpose of creating personal (and not group) memories[3].

In particular, this chapter considers the following two research questions in the context of a multimedia authoring system from community assets:

---

[3] Interested readers can find a video picturing the general concept of personalized community videos at http://www.youtube.com/user/TA2Project#p/u/6/re-uEyHszgM. And an example of personal video at http://www.youtube.com/user/TA2Project#p/u/4/Ho1p_zcipyA. Last access on May 15th 2013.

Table 2.1. Handling Multi-Camera Recordings of Concerts.

	Preparation, Capturing	*Relationship between Performers and People Recording*	*Intelligent Processes*	*Purpose*
Professional DVD	Scripted	Professional	Human director (planning)	Complete coverage
Anonymous UGC Mashup	Ad-Hoc	Similar likings, idols	Video search, video analysis	Complete coverage
Socially-Aware Community	Ad-Hoc	Family & friends	Video analysis, video authoring	Memories, bonds

Question 1.1 *Can a socially-aware multimedia authoring system be defined in terms of existing social science theories and human-centered processes, and if so, which?*

Question 1.2 *Does the functionality provided by a socially-aware multimedia authoring system provide an identifiable improvement over traditional authoring and sharing solutions? If so, how can these improvements be validated?*

Our work focuses parents, family members and friends of students participating in a high school concert. In this scenario, parents capture recordings of their children for later viewing and possible sharing with friends and relatives. Working with a test group at local high schools in two different countries (UK and the Netherlands), we investigate how focused content can be extracted from a shared repository, and how content can be enhanced and tailored to form the basis of a personalized multimedia artifact, that can be eventually transferred and shared with family and friends (each with different degrees of connectedness and tie strength with the performer and his/her parents). Results from a four-year

evaluation process provide useful insights into how a socially-aware multimedia authoring and sharing system should be designed and architected, for helping users in recalling personal memories and in nurturing their close circle relationships.

The remaining of this chapter is structured as follows. Section 2.1 discusses the user-centered methodology followed in this thesis, in which both technology and social issues were addressed. Then, motivated by social theories and interviews/focus groups with potential users, Section 2.2 identifies key requirements for socially-aware multimedia authoring and sharing systems. This section addresses the first research question, by providing guidelines to realize systems that meet those requirements. Section 2.3 reports on results and findings regarding the utility and usefulness of the proposed framework, thus directly responding the second research question. Lastly, Section 2.4 concludes the chapter.

2.1 Methodology

This thesis is part of an extended study to better understand the role that multimedia authoring tools can play in improving social communications between friends and families living apart. In particular, we are interested in understanding how individual users can personalize the use of community assets to make unique video stories that can be shared within a closed social circle (see Figure 2.1).

This work has been realized in the context of the pan-European project Together Anywhere, Together Anytime[4] (TA2). The goal of this project was to understand how technology can improve relationships between groups of people separated in space and time. We focused on an asynchronous authoring and sharing framework in which highly personalized music videos are constructed from a collection of independent parent-made recordings. For that, a system called *MyVideos* was developed as a collection of configurable processes, each of which allowed us to study one or more aspects of the development of socially-aware multimedia authoring systems.

We have been actively investigating this problem for several years. The methodology reported in this section (and complemented in the next chapters) integrates knowledge from human factors (e.g., focus groups/interviews for need assessment, iterative prototyping and user evaluation) and document engineering. Potential users have been involved in the design and evaluation process since the

[4] http://www.ta2-project.eu/

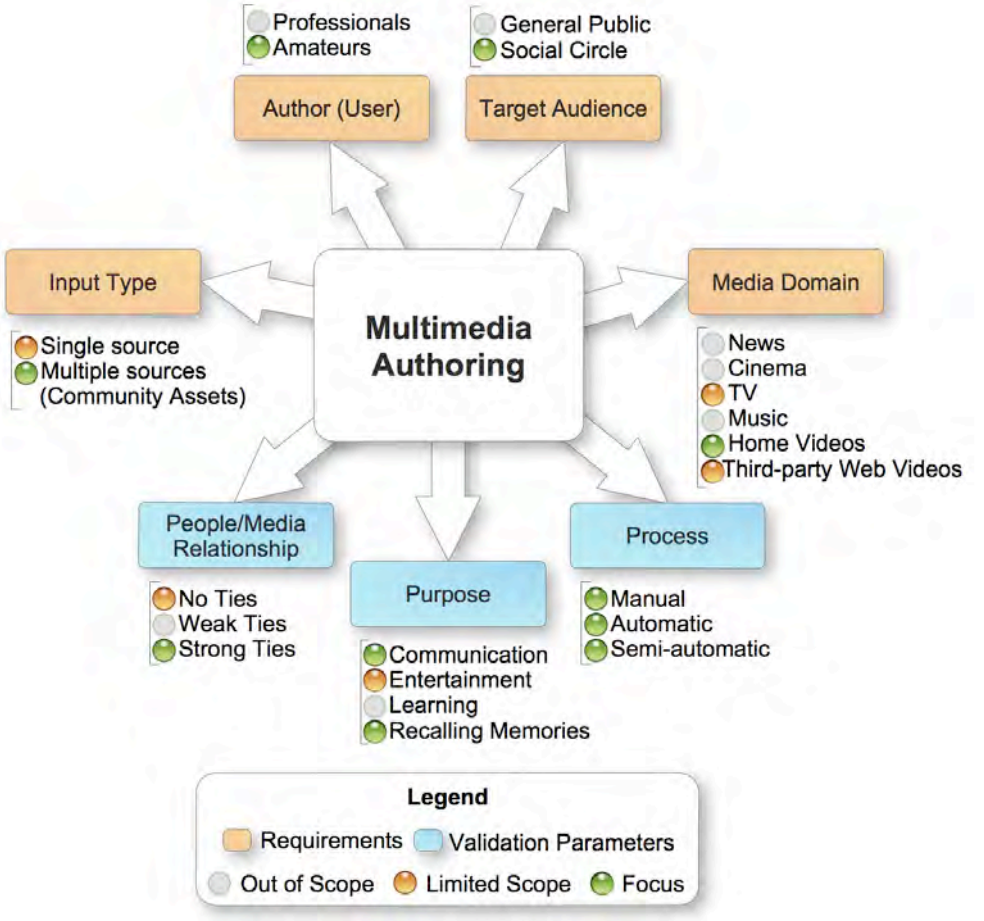


Figure 2.1. Overview of the requirements and validation parameters for socially-aware multimedia authoring systems.

beginning of the project, starting with interviews and focus groups, leading up to the evaluation of a two-phased prototype system.

A set of parents from local high schools has actively collaborated with this research. Starting in December 2009, the parents were invited to a focus group that took place in Amsterdam; in April 2010 they recorded (together with some researchers) a concert of their children. From Jul-Sep 2010, these parents used our

prototype application with the video material recorded in that concert. Based on the feedback and results, the software was re-designed in a second phase. This second time, we involved a high school in Woodbridge (UK), where a concert was recorded in November 2011. Subsequently, the parents that participated in that concert evaluated our second prototype implementation. During these years, we have systematically investigated mechanisms for helping users explore assets from a community collection of videos and to automatically generate 'stories' from these assets based on a narrative model.

2.1.1 Content Recording and Preparation

MyVideos has been tested and evaluated using data recorded in 4 different concerts as summarized in Table 2.2: a school rehearsal in Woodbridge[5] in the UK, a jazz concert by an Amsterdam local band called the Jazz Warriors[6], a school concert at the St. Ignatius Gymnasium[7], and finally another school concert in Woodbridge.

In December 2008 in the Woodbridge School concert (UK), a total of five cameras were used to capture the rehearsal. The master camera was placed in a fixed location, front and center to the stage, set to capture the entire scene (a 'wide' shot), with no camera movement and an external stereo microphone in a static location physically near to the rehearsal performance.

In the end of November 2009, a jazz concert was recorded as part of an asset collection process for the MyVideos phase 1. The goal of the capture session was to gain experience with a user setup that would be similar to that expected for the first trial. The concert took place on November 27th, 2009 at the Kompaszaal[8], a public restaurant and performance location in Amsterdam. The Jazz Warriors is a traditional big band with approximately 20 members. In total 8 cameras were used to capture the concert, where two cameras were considered as 'masters' and were placed at fixed locations at stage left and stage right. In total, about 220 video clips and approximately 80 images were collected at the event. The longest video clip was 50 minutes, the shortest 5 seconds.

These first two concerts were primarily experimental. They were very useful for testing the automatic processes for analyzing and annotating video clips: a

[5] http://www.woodbridge.suffolk.sch.uk
[6] http://jazzwarriors.nl
[7] http://www.ig.nl
[8] http://www.kompaszaal.nl

Table 2.2. Data gathering events.

Concert	*Date*	*Event Duration (approx.)*	*Musicians*	*Cameras (incl. master)*	*Videos Recorded*	*Media Objects*
Woodbridge School (UK)	Dec/2008	50min	25	5 (1)	100	100
Jazz Warriors (NL)	Nov/2009	50min	20	8 (2)	220	220
St. Ignatius Gymnasium (NL)	Apr/2010	1h35min	20	12 (2)	197	197
Woodbridge School (UK)	Nov/2011	1h20min	18	12 (1)	331	668

temporal alignment algorithm and a Semantic Video Annotation Suite. The temporal alignment tool is used to align all of the individual video clips to a common time base. The core of the temporal alignment algorithm is based on perceptual time-frequency analysis with a precision of 10ms. Figure 2.2 sketches the temporal alignment of a recorded dataset (more information on the datasets will be provided below). The level of accuracy of our tool is of around 99%, improving state-of-the-art solutions [44][59]. Since the focus of this thesis is not on content analysis, we will not further detail this part of the system. The interested reader can find the algorithm and its evaluation elsewhere [20]. The Semantic Video Annotation Suite [64] provides basic analysis functions, similar to the ones reported in [59]. The tool is capable of automatically detecting potential shot boundaries, of fragmenting the clips into coherent units, and of annotating the resulting video sub-clips.

In the next sections, we discuss the media gathering and annotation processes that preceded the user evaluations of MyVideos phase 1 and phase 2 prototype implementations.

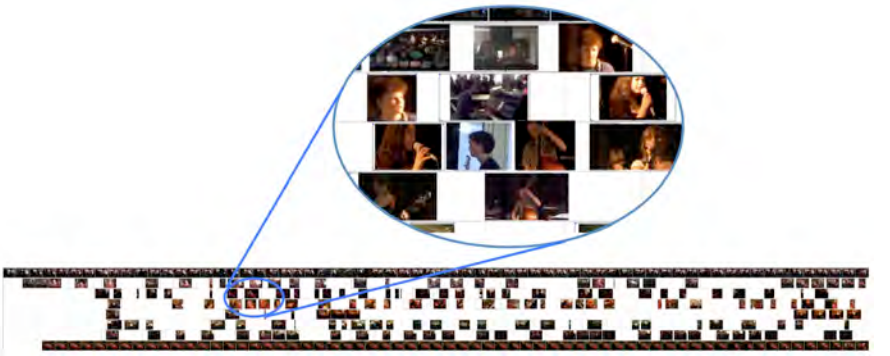


Figure 2.2. Temporal alignment of a real life data set from a concert, where a community of users recorded video clips.

2.1.1.1 Data Gathering for Phase 1

On April 16th, 2010 the concert from the Big Band – school band of the St. Ignatius Gymnasium – was recorded. In this case a core group of parents took part in the recordings and provided the research team with all the material. In total around 197 media objects were collected for a concert lasting about 1 hour and 35 minutes. Twelve (12) cameras were used; two of them used as the master cameras.

Once the footage was captured, the process to tag people, instruments and songs was realized in two stages. The first one was carried out manually. This task was performed looking through the videos and marking a line in a spreadsheet for each event (effectively it was almost always multiple lines to account for the multiple people/instruments). There were 7 kits in this process; each kit included 10 video files, ranging in length from about 5 seconds to 5 minutes. The quickest person took about 1 hour to complete while the longest kit took about 6 hours. The total time spent annotating 'manual' kits was approximately 16 hours. Later, a second approach was implemented by using a pre-populated data spreadsheet and an annotation sheet that used drop-down boxes taking data from the datasheet. This approach was more effective and the total time spent annotating 8 kits was approximately 12 hours. Yet computing the time spent to annotate the master track a rough approximation of total time spent annotating the concert was of about 40 hours. After the annotation phase, the initial prototype was ready to be evaluated.

2.1.1.2 Data Gathering for Phase 2

For the evaluation of the second prototype implementation, new recordings took place again in the Woodbridge high school (UK) in November 2011. The concert lasted around 1 hour and 20 minutes, in which 18 students performed in 14 songs. A total of twelve cameras were used to capture the concert. The master camera was placed in a fixed location, front and sideway to the stage. Eight cameras were distributed among parents, relatives, and friends of performers. Members of the research team used the other 3 cameras. In total about 331 raw video clips were captured, some of which were recorded before or after the event.

For this dataset, a hired group of people manually sub-clipped and annotated songs and performers. The total amount of time spent examining, sub-clipping and preparing the footage was around 156 hours. This includes a number of tasks apart from annotating clips, such as importing and transcoding all the videos to the same format, sub-clipping the footage, assigning annotations, transferring the annotations to machine readable CSV (*Comma-Separated Values*) files via OCR (*Optical Character Recognition*) and error checking. The outcome of this process was the creation of 668 sub-clips – or media objects out of the 331 original videos (see Table 2.2) – used in the evaluation of MyVideos phase 2.

2.1.2 MyVideos Implementation

The MyVideos application has been implemented as a Web-based application, targeting users with little technical background. From the user viewpoint this means that they only need access to the public Internet and everything runs within a JavaScript-enabled Web browser on their device. The server components are hosted on a dedicated testbed with a high bandwidth symmetrical Internet connection and virtualized processor clusters dedicated to hosting Web applications and serving video. In our architecture, each school would rent space and functionality on the testbed, in order to make systems like ours available to their community.

The server-side of our system includes a Mongrel Web application server (implemented in Ruby and Rails), a narrative engine (implemented in Java) that creates personalized narratives, a MySQL database that stores all the relational data concerning the media assets, and a media server that stores the recorded video clips and delivers them through HTTP (*Hypertext Transfer Protocol*) video streaming. The communication between the Web application and the narrative engine uses

JavaScript Object Notation (JSON). Only the application server and the video server are directly accessible through the Internet, while the remaining components are hidden to the outside world.

The client side only requires a Web browser and the *Ambulant Player*[9], for playing the video compilations in SMIL (*Synchronized Multimedia Integration Language*) [17]. The application on the client's devices was implemented using JavaScript and AJAX (*Asynchronous JavaScript and XML*). Additional JavaScript libraries have been used for simplifying the development of the client-side software. In particular, *YUI 2* and *jQuery* have been useful for event handling and AJAX interactions. For playback of individual video clips, two different solutions have been used. When supported by the browser, HTML5 video elements have been used (e.g., for an *iPad* implementation). Otherwise, we used an embedded *Flash* player (JW player).

2.1.3 Participants

The number of participants in both phases was kept small so that we could establish directed and long-term relationships. The qualitative nature of our interactions provided us with a deep understanding of the ways in which people currently share experiences to foster strong ties. The participants involved in both phases represent a realistic sample for the intended use case: parents, relatives, and/or friends of the kids going to the same high school; all of them tend to record the kids; some of them have some experience with multimedia editing tools. We believe that this sample of users provides us a relevant picture of the ways people currently record videos of other people they care about, and how they use such footage to share experiences within their (probably restricted) social group.

Since our main focus is to better understand small groups of people with strong interpersonal ties, the evaluation of MyVideos was realized with a fixed selection of users. It would have been impossible to do crowdsource testing, since we wanted to explore the fact that people had a social connection with the recorded footage. This section describes the subjects and methodology applied in each evaluation phase.

[9] http://www.ambulantplayer.org

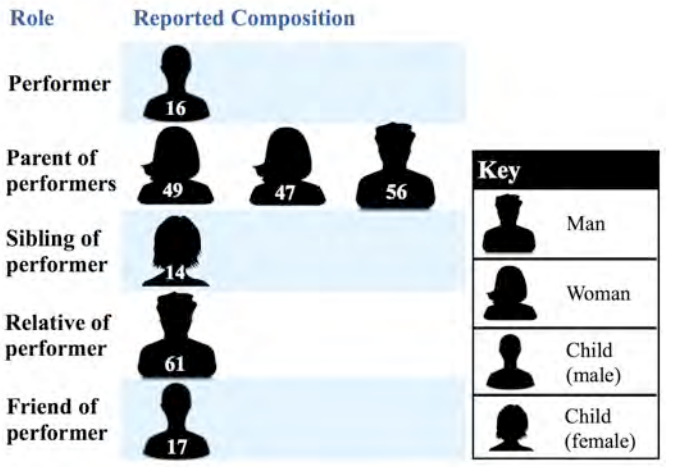


Figure 2.3. The makeup, age and gender of participants in phase 1 evaluation.

2.1.3.1 Phase 1 Setting

As illustrated in Figure 2.3, 7 people, among relatives and friends of the performers that attended the school concert in Amsterdam, were recruited. The rationale used for selecting the participants was diversity. We wanted to gather as many roles as possible for better understanding the social needs of our potential users. The participants were three high school students, a social scientist, a software engineer, an art designer and a visual artist, resulting in a variety of needs that may influence the video capturing, editing and sharing behaviors. All participants were Dutch. The average age of the participants was 37.1 years (SD = 20.6 years); 3 participants (42.8%) were female. Among the participants, 3 had children (ranging from 14 to 17 years old). All participants were currently living in the Netherlands, but the uncle of a performer that lived in the US. He was recruited to serve as an external participant (the only one that was not present in the concert). The prototype evaluation was conducted over a two-month span in the summer of 2010 (Jul-Sep).

More interested in subjective results that in statistical data, our approach was largely exploratory and interactive. The evaluation process consisted of 2 sessions. The initial one was used to collect background information about video recording habits, e.g., participants' intentions and the social relations around media. We also

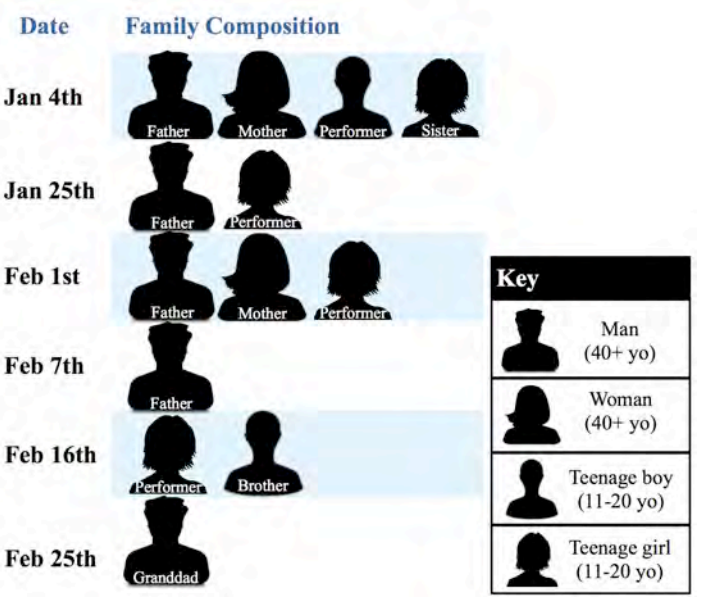


Figure 2.4. The makeup, age and gender of participants in phase 2 evaluation.

used this session as an opportunity to understand how participants conceptualized the concert. The second (in-depth) session was dedicated to capture video editing practices and media sharing routines of the participants, based on their interactions with the system. We used the footage they had recorded during the high school concert in the spring of 2010 to evaluate our initial prototype system. Both sessions were started with an ice-breaking activity on the whiteboard, followed by discussions around the research questions.

2.1.3.2 Phase 2 Setting

Thirteen (13) people (from 6 families) participated in the evaluation of our second prototype implementation. Participants consisted of performers, parents and other relatives of the teenagers that performed in the Woodbridge school concert, as illustrated in Figure 2.4. All participants were English speakers and were currently living in the UK. Seven of them (~54%) were 40+ years old; the other 6 people were in the 11-20-age range, 4 of which performed in the concert. Six (6)

participants were female. Participants kindly volunteered themselves for their participation, and the experiments were conducted over a two-month span in the beginning of 2012 (Jan-Feb).

We used a semi-structured approach for data collection. We started the individual interviews by explaining the high-level goals of our system and by asking participants about their video recording and sharing practices. Then, the participants were instructed to interact with the prototype system and to answer the evaluation questionnaires. Nine (9) out of the 13 participants committed to fill in the questionnaires discussed in Section 2.3.

2.2 Generic Architecture for Socially-Aware Authoring Systems

The motivation of our work is rooted in the inherent necessity of people for socializing and for nurturing relationships. As discussed in the previous section, we followed an interdisciplinary approach in which both technology and social issues were addressed. At the core of this approach was the establishment of a long-term relationship with a group of parents within local high schools (in the UK and in the Netherlands) as a basis for gathering requirements, evaluating prototype implementations and validating the socially-aware authoring concept proposed in this thesis work.

Motivated by social theories and focus groups/interviews with potential users, in this section we formalize the general guidelines for realizing socially-aware multimedia authoring and sharing systems. In Section 2.3 and in the next chapters, we discuss the evaluation of MyVideos, a system that realizes and validates such guidelines. The design and architecture of our socially-aware multimedia authoring framework are direct results from the long-term process reported in this thesis.

2.2.1 Social Science Principles

The experimental methodology presented in this thesis is based on two social science theories: *social connectedness* and *strength of the interpersonal ties*.

Social connectedness theory helps us to understand how social bonds are developed over time, and how existing relationships are maintained. Social connectedness happens when one person is aware of another person, and confirms his/her awareness [67]. Such awareness between people can be natural and intense

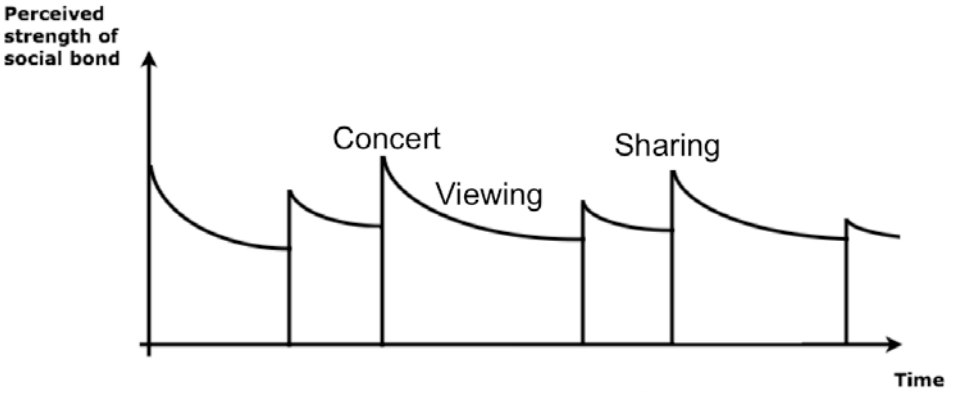


Figure 2.5. A schematic view of the perceived strength of social bond over time in relation to our scenario.

or lightweight. As reported elsewhere, even photos [72] and sounds [42] can be good vehicles for creating the feeling of connectedness. Figure 2.5 illustrates a schematic view of the perceived strength of a social bond over time, showing reoccurring shared events ('interaction rituals' in the Durkheim sense [23]), with a fading strength of the social bond in between. The peaks in the figure correspond to intense and natural shared moments, when people participate in a joint activity (e.g., a music concert) re-affirming their relationships and extending their common pool of shared memories. The smaller peaks correspond to social connectedness actions, such as sending a text message or sharing a personalized video of the shared event, that build on such shared memories. If we were to follow the social connectedness theory, we would design a system that mediates the smaller peaks and thus helps in fostering relationships over time.

Granovetter [55] defines interpersonal ties as:

> *"… a combination of the amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services which characterize the tie."*

If we were to design a video sharing system intended for family and friends, we would exploit the social bonds between people by taking into account their personal relationships (*intimacy*). The system would provide mechanisms for

personalizing the resulting videos (*adding personal intensity*) with some effort (*amount of time*), and would allow the recipient to acknowledge and reply to the creator (*reciprocity*).

2.2.2 Family Interviews and Focus Groups

In order to better understand the problem space, we involved a representative group of users at the beginning of the evaluation process. The first evaluation, in 2008, consisted of interviews with sixteen families across four countries (UK, Sweden, Netherlands, and Germany). The second evaluation, in-depth focus groups – with three parents each – was run in the summer of 2009 in the UK and in December 2009 in the Netherlands.

As social connectedness theory suggests, many participants engaged in varied forms of media sharing. Participants felt that reliving memories and sharing experiences helped them (and other households) feeling closer. Parents e-mailed pictures of the kids playing football to the grandparents, shared holiday pictures via *Picasa*, or on disk, or using Facebook, enabling friends and families to stay in touch with each other's lives. Nevertheless, the interviewed people said that if they shared media, they would do so via communication methods they perceived as private and then only to trusted contacts. There was a general reticence from the parents towards existing social networking sites. In the UK, the parents stressed that they would not share the videos with 'the world', but would share it with other family members for fun. For example, when asked about YouTube one parent said:

> *"I haven't... my wife's side of the family... they're always putting clips of video on YouTube and all these sorts of things... that makes me cringe a bit... I think… well, why would I want to do that? Do I think that's interesting to anybody?"*

A number of parents reported photography as a hobby and would routinely edit their shared images. Their children, on the other hand, even if interested in photography, seemed less keen to manually edit pictures, and declared a strong preference for automatic edits or relied on their parents. The participants would then discuss the incidents relating to the pictures later on with friends and family, on the phone or at the next reunion. Home videos tended to be watched far less frequently, although the young pre-teen participants appreciated them and were

described by their parents as having *"worn the tape[s] down"* from constant viewing when much younger.

Based on the interviews, we concluded that current social media approaches are not adequate for a family or a small social group for storing and sharing collections of media that is personal and important to them [63]. Much richer systems are needed and will become an essential part of life for family relationships. In general the participants' responses converged to:

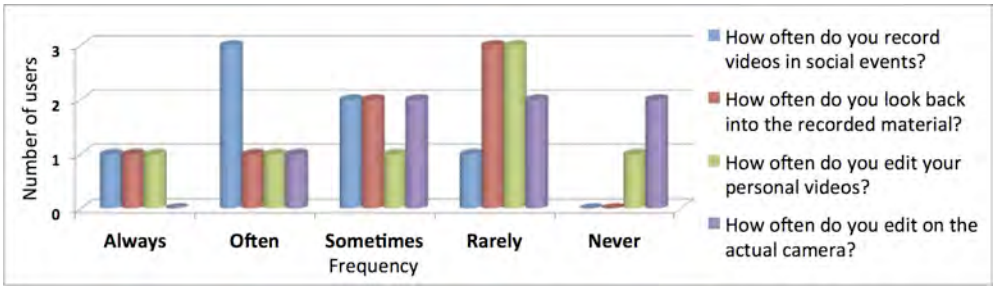
- A willingness to engage in diversified forms of recollections through recorded videos;
- A clear requirement for systems that could be trusted as ensuring privacy;
- A positive reaction to the suggestion of automatic and intelligent mechanisms for managing home videos.

In each case, creating personalized video stories (tailored for family use) remained a core issue.

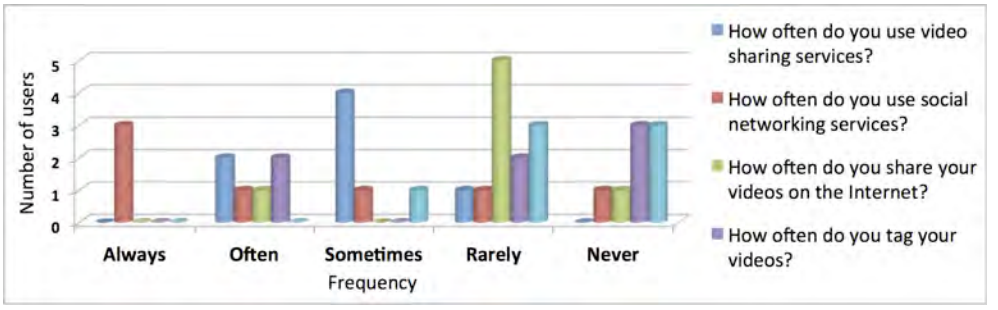
2.2.3 Requirements Gathering

Figure 2.6 a) shows the answers of the participants in Amsterdam to the questionnaires about video recording and editing practices during phase 1 evaluation. Most participants said they *often* record videos in social events (e.g., family gatherings, vacation trips and/or school concerts). However, validating previous studies [19], they *rarely* look at the recorded material afterwards. According to the participants, one problem is the relatively high number of media assets captured during an event – for instance, around 200 media assets from 12 cameras for a concert lasting 1h35min. Another problem is that the footage, as captured, cannot be easily explored.

For most of them, video editing was considered time consuming and way too complicated. Therefore, they *rarely* edit their videos. Most users said that they had an editing suite at home. PC users were familiar with Windows Movie Maker, while Mac users with iMovie. Some participants described how they would create a movie about the high school concert using their preferred editing tool. They would choose some clips and drag them to the timeline. Then, they would use visual effects, transitions and sounds that are usually provided with the video editing software. In general, they indicated that they would tell the story of the concert using their personal videos. Some participants mentioned that video editing



a) Video recording and editing practices.



b) Media sharing habits and social relations.

Figure 2.6. Results of the questionnaires about social practices around personal videos (phase 1 evaluation).

also could demand high processing power, which would slow down the computer. As a workaround, they occasionally (between *sometimes* and *never*) would perform minor editing operations (e.g., clipping) on their own video camera.

Figure 2.6 b) presents the results of the questionnaires about media sharing habits and social relations around the media. Participants said they were used to watch videos on YouTube *sometimes*, and many of them used Facebook quite frequently (*always*). However, they were not used (between *never* or *rarely*) to tag videos and/or photos. When prompted whether and how they shared their videos, they repeatedly said that in general they *rarely* posted personal videos on the Web. While the youngest participants argued their personal videos were not interesting enough, for our older respondents privacy was the main concern not to share personal videos on the Web.

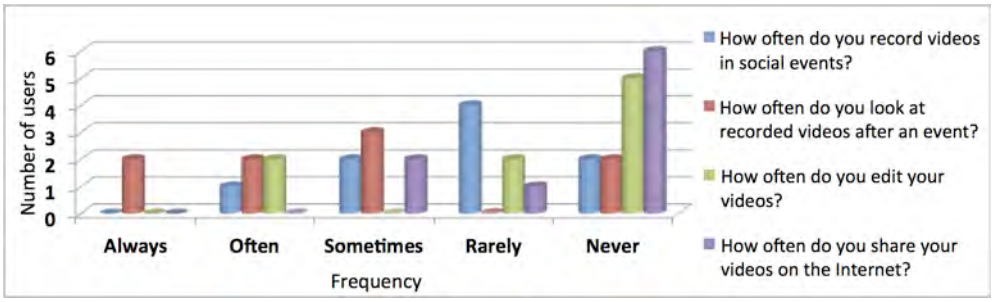


Figure 2.7. Social media habits (phase 2 evaluation).

> *"It is personal… if I make a personal shot, a close-up of my daughter, for example, and I do this for personal reasons, I never do this for the others." (Mother of a performer)*

Figure 2.7 shows some responses of the British participants to the background questions related to media capturing, editing, and sharing. Most subjects said they *rarely* record videos in social events (less frequently than the group in Amsterdam). Although, they declared to *sometimes* look at the videos they recorded after the event has taken place. Five (5) out 9 participants said they were unfamiliar with video editing tools, and therefore, they *never* edit their videos. The vast majority said they were quite concerned about sharing personal videos on the Internet, and they were not used to do so (6 participants said *never*, while 1 *rarely*).

Based on these general user needs, social theories and initial interviews with focus groups, we defined a number of requirements for socially-aware multimedia authoring and sharing systems, as follows:

i. *Support social connectedness:* it should provide tools and mechanisms for maintaining relationships over time. The goal is not so much on supporting high intensity moments – the event – but for the small peaks of awareness (recollection of the event);
ii. *Support privacy and control:* most parents in the interviews and the focus groups expressed that current video sharing models do not fit the needs of family and friends due to privacy issues. Thus, new systems should address the parents' concerns, and provide adequate privacy mechanisms;

iii. *Support effortless interaction:* people are reluctant to invest time in processes they consider that could be done automatically. Future systems should include automatic processes for analyzing, annotating, and creating videos of the shared event; and
iv. *Support personal effort, intimacy and reciprocity:* while such automatic processes lower the burden for the user, they do not conform to existing social theories. Since we do not want to limit the joy of handcrafting videos for others, systems should offer optional manual interfaces for personalization purposes.

We used these requirements as the basis for specifying the guidelines discussed in the next section.

2.2.4 Guidelines

In order to support the social theories described in Section 2.2.1 and the requirements identified in Section 2.2.3, our socially-aware multimedia authoring framework considers a number of automatic, semi-automatic and manual processes that assist in the media exploration and creation of personal memories of an event. These processes balance convenience and personal effort when making targeted, personalized videos. Emotional intensity is provided by a recommendation algorithm that searches for people and moments that might bring memories to the user. For mediating intimacy, our framework proposes means to enrich videos for others by including highly personalized comments. With these features we intend to increase the feeling of connectedness, particularly among family members and friends who could not attend the social event.

2.2.4.1 Supporting Emotional Intensity

An assumption leading to the design of our socially-aware framework was that in a community setting, users are particularly interested in looking for video clips in which people close to them are featured (social-based searches). Such assumption is validated in Chapter 3, which presents our efforts in designing and implementing an interface for browsing multi-camera recordings. The core of the navigation interface is a recommender algorithm that takes into account not only the filters selected by the user and the content quality assessment, but also the recording behavior of each user individually. This feature considers the semantic annotations

associated to the user's media and on the subjects that more frequently appear on his/her recordings.

For example, a father can make a request for his daughter playing 'Cry Me River', since he remembers this was an emotive moment of the concert. Given an example query:

SelectedPersons = [Julia];
SelectedSong = [Cry Me River].

The result will be:

QueryPersons(Julia) ∩ *QueryEvents(Cry Me a River)*

The query algorithm works as follows:

1. Select fragments of the video clips matching the query; in case of complex queries, select intersecting sets;
2. If the result consists of one fragment, return it;
3. If the result consists of more than one fragment, order the resulting list based on the following criteria:
    - The requested person;
    - The video clips uploaded by the logged user;
    - The subjects that appear more frequently in the video clips uploaded by the logged user (affection parameter);
    - The content quality assessment (e.g., shot type, resolution, duration).

In addition to the query interface that allows users to find moments that they particularly remember, a socially-aware multimedia authoring framework should offer optional manual interfaces for improving semantic annotations. When users are searching for specific memories, it might happen that results are not accurate due to errors in the annotations. Our approach considers that users could correct such annotations while previewing individual clips. For example, they can change/add/remove the name of the performer and the title of the song.

2.2.4.2 Reflecting Personal Effort

One of the major differentiators of our work is that its primary purpose is not the creation of an appealing video summary version of the event or the creation of a collective collaborative community work. Instead, our approach intends to facilitate the reuse of collective contents for individual needs. Rather than using personal fragments to strengthen a common group video, our work takes groups fragments to increase the value of a personal video. Each of the videos created by a socially-aware multimedia authoring system should be tailored to the needs of particular members of the group – the video created for the father of the trombone player will be different from the one for the mother of the bass player, even though they may share some common content.

Users should be able to automatically assemble a story based on a number of parameters such as people to be featured, songs to be included, and duration of the compilation. Such selection triggers a narrative engine that creates an initial video using multi-camera recordings. The narrative engine selects the most appropriate fragments of videos from the repository, based on the user preferences, and assembles them following basic narrative constructs.

Given an example query:

SelectedPersons = [Julia];
SelectedSong = [Cry Me River];
SelectedDuration = [3minutes].

The algorithm extracts the chosen song from the master audio track, and uses its structure as backbone for the narration. It then selects all the video content aligned with the selected audio fragment; the master video track provides a good foundation and possible fall back fragments that are not well covered by individual recordings. The *audio* object is the leading layer and, in turn, it is made of *AudioClips*. This structure generates a sequence of all the songs that relate to the query. As soon as the length of the song sequence extends beyond the *SelectedDuration*, the compilation is terminated. The *video* object has the role of selecting appropriate video content in sync with the audio. An example of the selection criteria is the following:

1. Select video clip that is in sync with the audio;
2. Ensure time continuity of the video;

3. If there are more potential clips that ensure continuity, select those with *Person* annotations matching the user choices stored in *SelectedPersons*;
4. If the result consists of more clips, select those which *Instruments* annotation match the instruments that are active in the audio layer;
5. If the result consists of more clips, select those which *Person* annotation matches the persons currently playing.

Once the automatic authoring process is complete, a new video compilation is created in which the selected song and people are featured. As reported elsewhere [76], such narrative constructs have been developed and tested together with professional video editors. Our assumption, based on the social theories, was that automatic methods – while useful – were not sufficient for creating personal memories of an event. Such assumption is validated in Chapter 4. Figure 2.8 shows a comparison between automatic and manual generation of mashups. Automatic techniques are better suited for group needs such as a complete coverage or a summary of the event, but are not capable of capturing subtle personal and affective bonds. We argue instead for hybrid solutions, in which manual processes allow users to add their personal view to automatically assembled videos.

A socially-aware multimedia authoring system should provide such interfaces for manually fine-tuning video compilations. Users can improve and personalize existing productions by including other video clips from the shared repository. For example, a parent can add more clips in which his daughter is featured for sharing with grandma, or he can instead add a particularly funny moment from the event when creating a version for his brother. As we will discuss in Chapter 4, participants liked such functionality, which automatic processes are not able to provide.

2.2.4.3 Supporting Intimacy and Enabling Reciprocity

Apart from allowing fine-tuning of assembled video stories, a socially-aware multimedia authoring system should enable users to perform enrichments. Users can record an introductory audio or video, leading to more personalized stories. As we will see in Chapter 4, this functionality (we call it ‘capture me’) was appreciated by most of our participants.

Our framework also addresses reciprocity by enabling life-long editing and enriching of compiled videos. As indicated before, videos created using our framework can be manually improved and enriched using other assets from the

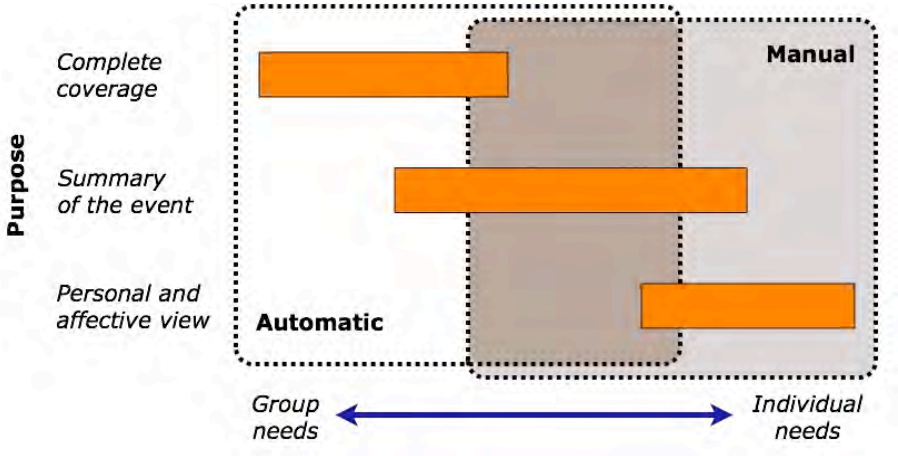


Figure 2.8. Comparison between automatic and manual generation of video compilations. Automatic methods are not sufficient for creating personal and intimate memories.

repository, and adding personal video and audio recordings. In Chapter 5, we go a step further, discussing the possibility of the recipients adding comments synchronized to specific moments within the video productions. Thus, users receiving an assembled video story can easily include further timed comments as a reciprocity action intended to the original sender. For example, a grandmother, who receives a video story from her son, might add a *"Isn't my granddaughter cute?!"* reply as a reciprocal message within the video. The main benefit is that this functionality enables people to comment and enrich existing video stories.

2.2.4.4 Guidelines relative to Requirements

In addition to supporting *emotional intensity* (requirement i), reflecting *personal effort*, supporting *intimacy* and enabling *reciprocity* (requirement iv), our socially-aware multimedia authoring framework also meets the other requirements identified in Section 2.2.3, as discussed below.

Using a trusted storage media server (provided, for instance, by the school) we address the *privacy* issue (requirement ii). Parents can upload the material from

the concerts to a common media repository. The repository is a controlled environment, since it is provided and maintained by the school, instead of being an external resource controlled by a third-party company. Moreover, all the media material is tagged and associated with the parent who uploaded it, and there are mechanisms so parents can decide not to share certain clips in which their children appear. Users can use their credentials for navigating the repository – those parts allowed to them – and for creating different stories for different people.

The requirement on *effortless interaction* (requirement iii) is met by the provision of a number of automatic processes that analyze and annotate the videos, and that help users to navigate media assets and to create memories. As introduced in the previous subsections, users can navigate the video repository using a recommender algorithm, and they can automatically generate video compilations from the multi-camera recordings.

2.3 Evaluation

In this section we report on evaluation of the utility and usefulness of our socially-aware multimedia authoring framework. In particular, our results address the requirements on *social connectedness*, and *privacy and control* (requirements i and ii, respectively). As described above, the evaluations of the prototype system have taken place in two different countries (UK and the Netherlands) since 2008, when we started exploring this novel area of research. Our results have been obtained via questionnaires, user testing and observations.

During phase 1 evaluation, users were instructed to interact with the MyVideos prototype system after responding the background survey presented in Section 2.2.3. Figure 2.9 presents the answers regarding the overall assessment after users interacted with the system. In general participants liked MyVideos and considered its functionality useful (Q1.1). Based on the received feedback, we can conclude that participants appreciated the benefits of our system and considered it a valuable vehicle for remembering events, thus improving social connectedness (requirement i). In particular, participants largely agreed that MyVideos would help them in recalling memories of social events (Q1.2). They also indicated that by using MyVideos they would share more videos with others (Q1.3). As shown in Figure 2.5, this feedback is aligned with the small peaks of awareness we intended to mediate with socially-aware multimedia authoring tools.

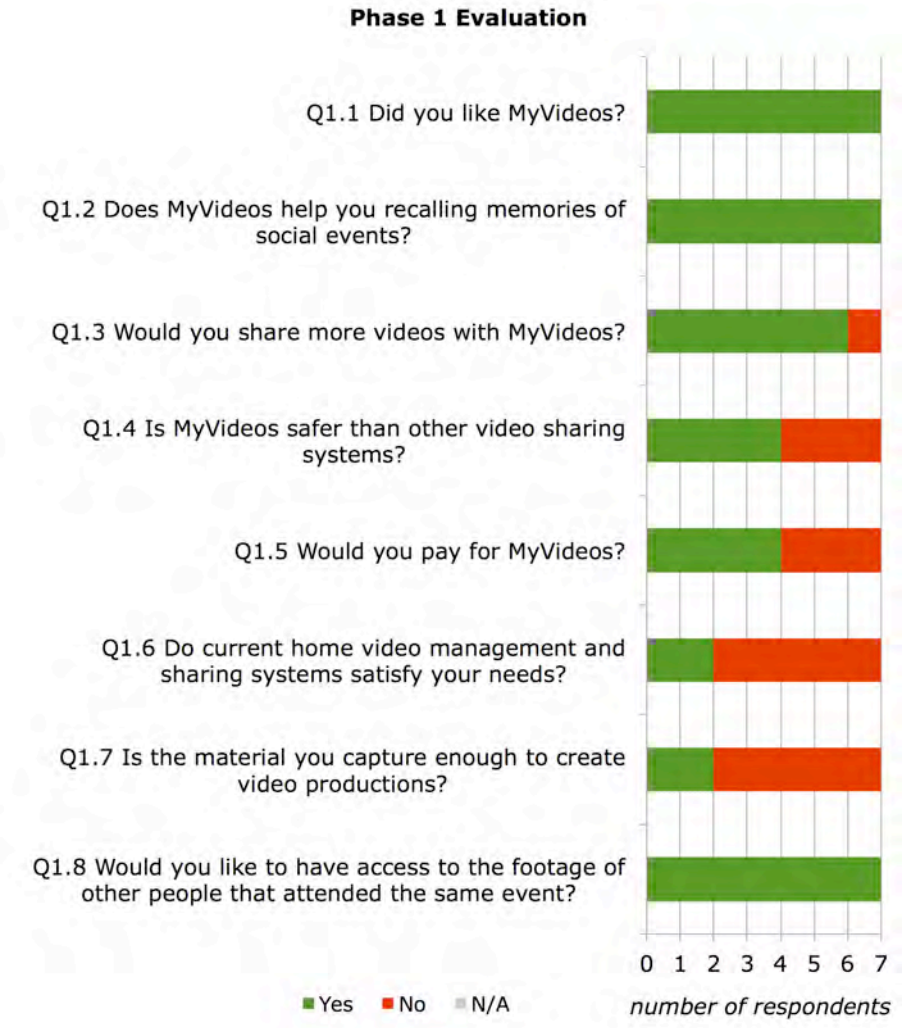


Figure 2.9. Utility and usefulness of MyVideos. Results of the questionnaire from phase 1 evaluation (Amsterdam/NL).

It might be surprising that although participants liked the system, some of them said that they did not find it much 'safer' than other video sharing services (Q1.4), or that they would not pay for it (Q1.5). As discussed earlier, the first issue has been motivated by privacy concerns (requirement ii). Most senior users were reluctant to uploading material outside their reach, hard drive, (even though it was a controlled environment). For the latter issue, we present more insights in the second evaluation process. Lastly, most of our subjects said that current home video management and sharing systems do not satisfy their needs (Q1.6). When questioned whether their video material would be enough to create a compelling video, they mainly answer negatively (Q1.7). They agreed that content captured by other people that participated at the same event could be interesting for others (Q1.8). However, most of the users asserted that current tools do not allow for easy watching and repurposing other parents' footage.

Figure 2.10 presents the answers to the questions related to the utility and usefulness of the second prototype system, including comparisons to other existing solutions. Overall, participants were enthusiastic about MyVideos (Q2.1). As in phase 1 evaluation, all participants declared that our socially-aware multimedia authoring framework helped them to recall memories of social events (Q2.2), and it made them feel more connected with their loved ones (Q2.3). These results directly meet requirement i.

> *"Overall, I had great fun. It was more than just getting into that concert again. It was doing something completely different. Almost like another activity. Which could almost have been anything. But the fact it was this concert, with my daughter in it, made it extra special." (Father of a performer)*

> *"I was especially keen to use this to create a video of my son playing cello to share with my father who lives in Wales… I actually don't have any videos of him playing cello as it is often not the done thing to video concerts…" (Mother of a performer)*

Similarly to the result obtained in phase 1 evaluation, participants indicated they would share more videos if they had a tool like ours at hand (Q2.4). However, only 4 (out of 9) considered the system 'safer' than current video sharing services (Q2.5), while 5 said they would spend money on it (Q2.6). A user argued about the cost-benefit of having a system like MyVideos.

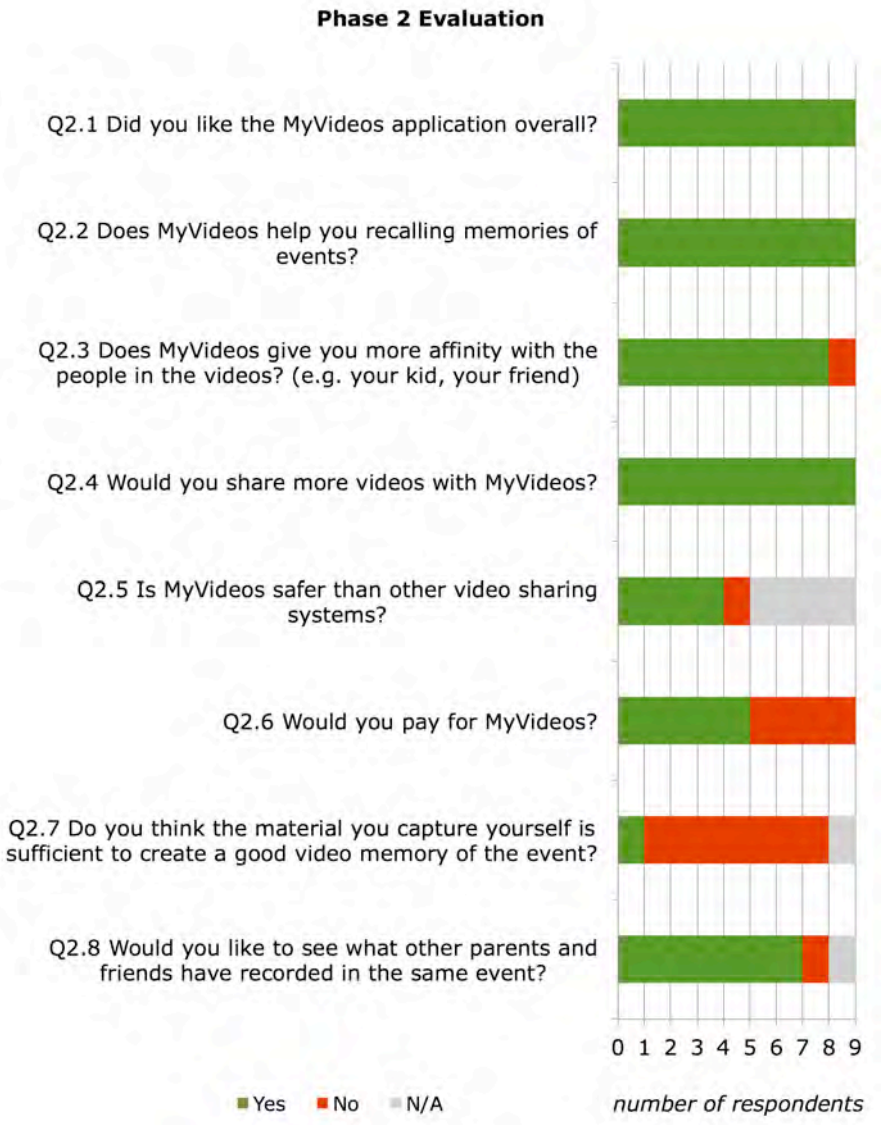


Figure 2.10. Utility and usefulness of MyVideos. Results of the questionnaire from phase 2 evaluation (Woodbridge/UK).

"Maybe I would pay for it, but it depends on cost and how much it would be used." (Mother of a performer)

On the other hand, a teenager justified his opinion, which is common among his age group.

"I tend not to bother with paid services; I just do without the service." (Brother of a performer)

It is important to highlight that most participants agreed that the material they usually capture is not sufficient to create a good video memory of an event (Q2.7). Therefore, it would be useful to have access to the content recorded by other parents' (Q2.8). Based on the participants' comments and answers, we get a strong sense that current tools are not enough to attend their needs. Current video sharing platforms on the Web do not allow for a collection of families that may have limited interactions to be brought together by contributing media assets for common use.

2.4 Discussion

In this chapter we reformulated the research problem of multimedia authoring, by investigating mechanisms and principles for togetherness and social connectivity around media. During 4 years, our user-centered methodology involved interviews/focus groups with users, prototype implementations and user evaluation. Motivated by general user needs, social theories and initial interviews, we specified a set of guidelines for the design and implementation of socially-aware multimedia authoring and sharing tools. We aim at nurturing strong ties and improving social connectedness by supporting *emotional intensity*, *personal effort* and *intimacy*, and by enabling *reciprocity*. As shown in this chapter, our approach is aligned with the requirements needed for social communities that are not addressed by existing social media Web applications. These guidelines characterize the first contribution of this chapter, and directly answer the first research question.

The overall evaluation process of a system that realizes such guidelines represents the second contribution of this chapter. It contemplated a long-term process in the Netherlands and in the UK, in which people actively participated and recorded concerts of their relatives/friends. Results from the evaluation process

show that the functionality provided by our socially-aware multimedia authoring system meets our requirements and brings an identifiable improvement over traditional approaches. These results, which are complemented by other findings in the next chapters, directly answer our second research question, and show that a system like ours is a valid alternative for social interactions when apart.

In the next chapters, we look into detail at each step that composes the socially-aware multimedia authoring workflow discussed in Chapter 1. First, in Chapter 3 we present our efforts in enabling community-based users to explore and navigate a large content space based on their personal interests. While following the *emotional intensity* guideline, our design meets requirement i (social connectedness). Then, in Chapter 4 we discuss the balance between convenience and personal effort when generating highly personalized video compilations of targeted interest within a social circle. This chapter addresses the *personal effort* guideline, and the evaluation results show that we meet requirements iii and iv (effortless interaction and personal effort/intimacy, respectively). Finally, while following the *intimacy* and *reciprocity* guidelines, Chapter 5 turns its attention to supporting the recipient in commenting within a video story (requirement iv).

3

Designing Socially-Aware Video Exploration from Community Assets[1]

The previous chapter provided the basis of socially-aware multimedia authoring. Our results validated the main assumptions, showing that users appreciate the importance of video sharing for building common experiences and for increasing the feeling of togetherness with others. Our results also indicated that current video sharing services fail to meet users' needs, because they miss useful mechanisms for navigating media and do not take into account emotional intensity and intimacy. In this chapter we argue that there is a need for useful mechanisms for navigating and sharing media, and socially-aware video management systems should provide efficient automatic processes to manage personal interests.

The wide availability of video recording devices in mobile telephones and pocket cameras has made documenting shared events easy (see Figure 3.1). The collected set of videos provides a rich archive from which users can enjoy content that matches their personal interest. Unfortunately, current browsing tools, including social networks, are not geared to supporting this form of selective consumption; these tools are geared towards throwing away unwanted content from a single collection, and not for browsing a broader community collection of temporally aligned alternatives. Current video tools often support only a high-level

[1] This chapter is based on the following paper:

D.C. Pedrosa, R.L. Guimarães, P. Cesar and D.C.A. Bulterman. 2013. Designing Socially-Aware Video Exploration: A Case Study Using School Concert Assets. In Proceedings of the 17th International Academic MindTrek Conference: Making Sense of Converging Media (MindTrek '13).

abstraction of objects and events, and do not help users to explore community videos that portray people within their social circle. Even though social networks archive media based on higher-order social relationships, they do not provide support for searching and navigating media content that was captured at a particular event by different camera people.

Most social events have an inherent structure that can be used to aid searching for content. We can take advantage of this structure for the development of socially-aware video exploration interfaces. Most participants at an event will attach different levels of importance to any given sub-event, based on their personal/social preferences. If we consider a high school concert, it has a structure (the order of the songs), a sub-structure (individual songs) and multiple levels of sub-sub-structure: solos, duets, vocal announcements and other often-unpredictable happenings. As discussed in the previous chapter, not everyone at the concert (or viewing it) will be equally interested in all parts. Parents will focus on their own children, students on their friends, and invited guests *on the clock*.

This chapter focuses on our efforts in designing and implementing an interface for browsing community assets, in which the relationships between users of the system and performers, featured in the videos, play an essential role in content selection. Our work, which follows the *emotional intensity* guideline (see Chapter 2), includes our findings and key results from the two-phased series of evaluations. In the next chapter we will show that such social bonds are key not only for navigating a shared media space, but also for authoring personalized stories users care about. Here, we focus on the importance of providing a rich representation of an event (in this case, a high school concert) in a way that helps users to navigate and explore a community repository based on their social/personal interests. The research question we address is:

Question 1.3 *Does a socially-aware video exploration system provide an identifiable improvement over current approaches for accessing and navigating a repository of shared media?*

To answer this research question we first present a browsing interface, and the underlying system infrastructure, that allow for socially-aware exploration of a collection of media assets captured in an event. Users can explore and navigate (fragments of) video clips recorded by several people based on their own personal/social interests. The design, deployment and evaluation of the system resulted in the identification of key requirements for this novel type of browsing

Figure 3.1. Typical interface for watching videos on the Web. It does not take into account the social affinity between viewers and subjects featured in the video.

interfaces. In particular, our approach 1) supports exploration based on the inherent event structure; 2) it makes use of contextual information to help in the navigation process; 3) it allows for flexible searches based on combination of filters; and finally, 4) it provides a way to switch between cameras angles that might have captured different aspects of the event.

The structure of this chapter is as follows. First, Section 3.1 provides an overview of the design and evaluation of an initial prototype system for socially-aware video exploration. Based on the users' feedback, a set of functional requirements was gathered. Then, in Section 3.2 we describe the design and implementation of the second version of the browsing interface that addresses these requirements. Next, Section 3.3 reports the evaluation of such system, analyzing the results. Finally, Section 3.4 provides a reflection on how our findings fit in the context of this thesis.

3.1 Community-based Browsing

The family interviews and focus groups in the beginning of this journey (see Chapter 2) provided us valuable data for identifying a series of requirements. The conclusion was that current social media sharing interfaces are not adequate for satisfying the expectations of strong ties. In this chapter, we focus on innovative interfaces that help users to explore a shared media repository they have social affinity with (*emotional intensity* guideline defined in Chapter 2). The final goal is to provide interfaces that can help shaping and sharing memories of important events with family members and friends.

The starting point of our investigation was traditional video browsing interfaces, such as YouTube (see Figure 3.1). Nevertheless, early in this process,

Figure 3.2. Initial prototype implementation for browsing videos (thumbnail view).

we realized that this kind of service does not provide social filters (e.g., to select videos by a particular performer) for concert videos, and it does not take advantage of the temporal relationships between videos belonging to the same event.

To address these limitations of current video sharing services, our initial video browsing interface offered two views for exploring community contributed video clips. The *thumbnail view* (Figure 3.2) displayed media assets in a paginated grid, while the *timeline view* (Figure 3.3) showed how recorded videos temporally fitted the event timeline. In both views a user could apply six different filters to refine a query. These filters were: *all media*, *my media*, *cameras*, *people*, *instruments* and *events*. 'All media' referred to all videos uploaded to the system. 'My media' restricted navigation to only the videos uploaded by the current user. 'Cameras', 'people', 'instruments', and 'events' filters would display the respective annotated video clips based on the filter selection (e.g., 'Julia' or 'Drums').

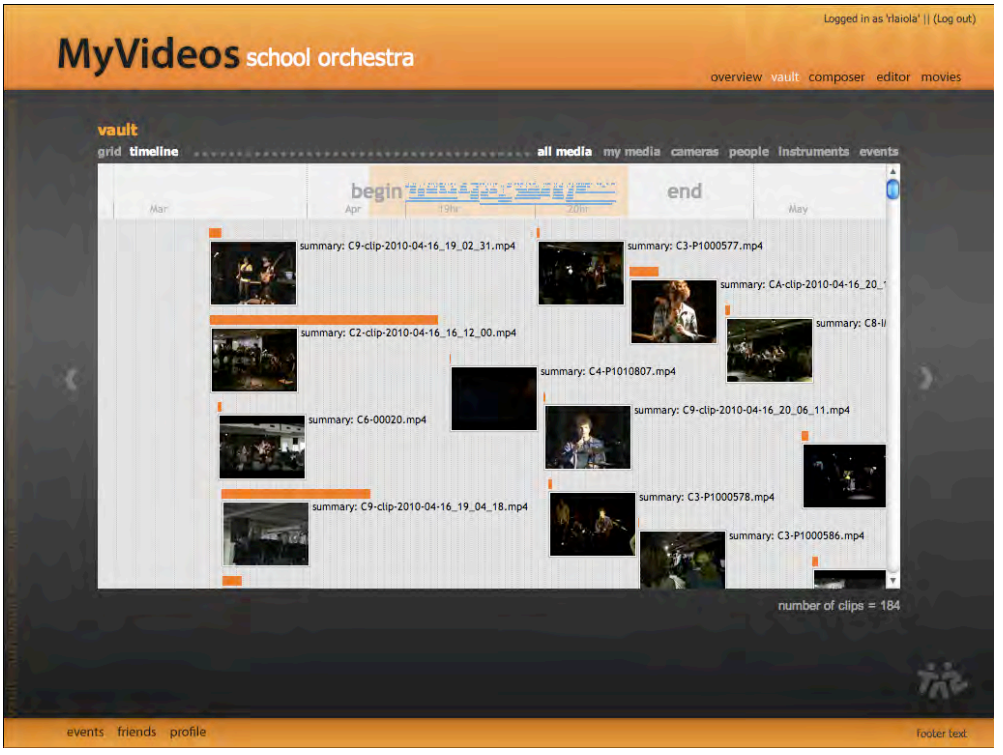


Figure 3.3. Initial prototype implementation for browsing videos (timeline view).

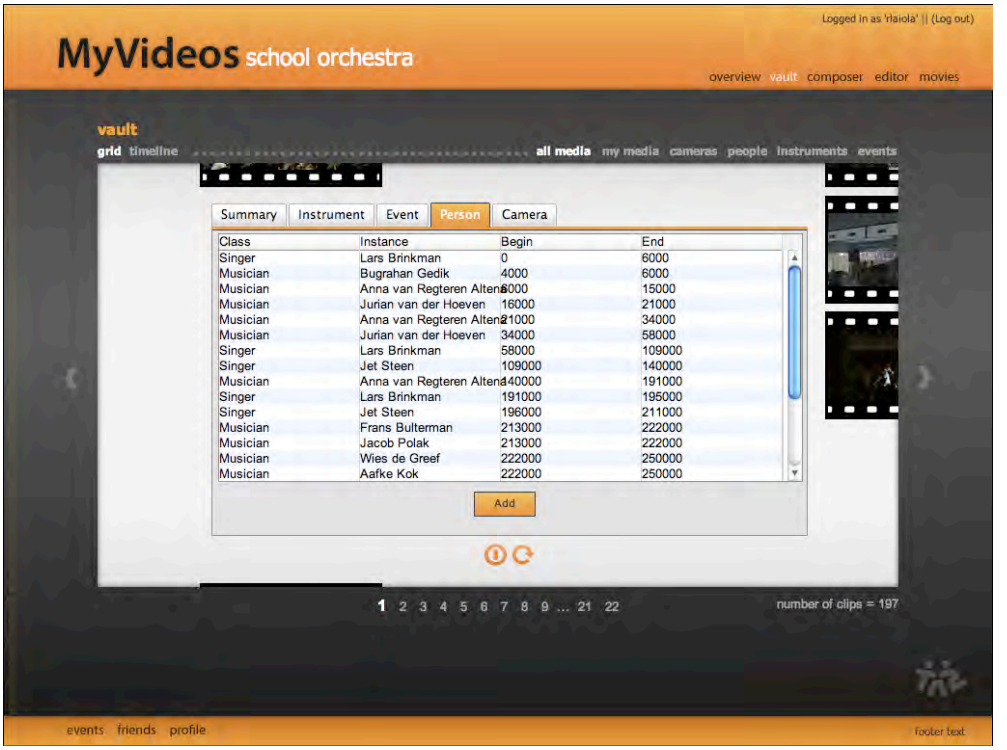


Figure 3.4. Initial interface for adding and correcting annotations.

Besides allowing for navigating the video clips, our initial interface also enabled users to annotate media assets and to correct existing annotations that could be wrong. When showing a video clip, a user could 'flip it' over and access all annotations related to that clip (as shown in Figure 3.4).

Using the footage recorded during the Big Band concert in Amsterdam, potential users were invited to evaluate our initial system. Details about the methodology and user assessment can be found in Chapter 2. In the remaining of this section, we discuss the results regarding media exploration obtained in the first evaluation phase. From these results a set of new requirements were elicited, and used for the design of the second phase.

3.1.1 Phase 1 Evaluation

In general, participants' feedback for the first version of our system was positive (see Figure 3.5). Four (4) out of 7 participants said it was better than traditional tools to find people they cared about (Q1.1). We received slightly better feedback when we asked whether our system was better to browse videos recorded by other parents (Q1.2). These results are directly aligned with the requirements of emotional intensity and easiness of use.

During the evaluation session, participants were actively looking for video clips of their close friends and relatives. In particular, some participants wanted to immediately share video clips with members of their close circle. *"Can I send it now?"* was a common reaction after seeing a video clip they especially liked. When asked how they would share the videos, teenagers expressed they would rather download the video files to their local computers, send a link of a particular video by email or share on YouTube and/or Facebook. Parents, on the other hand, indicated that a 'Burn to DVD' functionality of the selected videos also would be convenient given that grandparents usually do not have Internet access at home.

When prompted about what they remembered of the concert, most participants that attended it said that they recalled superficially the spatial arrangement of the stage (see Figure 3.6). At this point, some participants

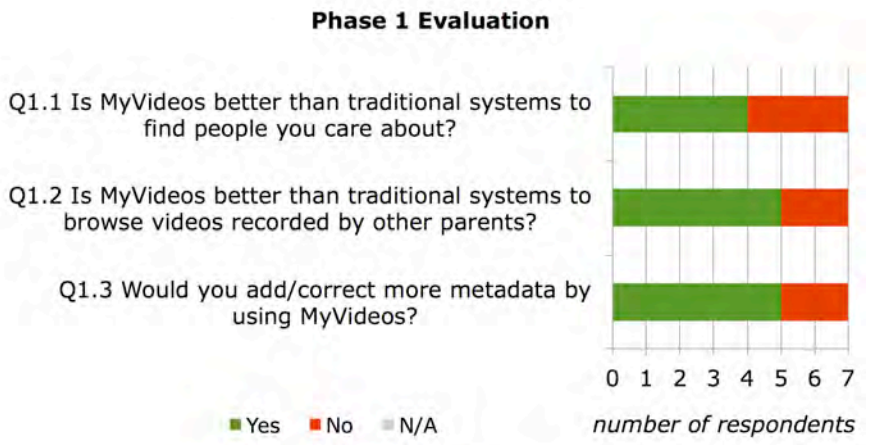


Figure 3.5. Results of the questionnaires from phase 1 evaluation.

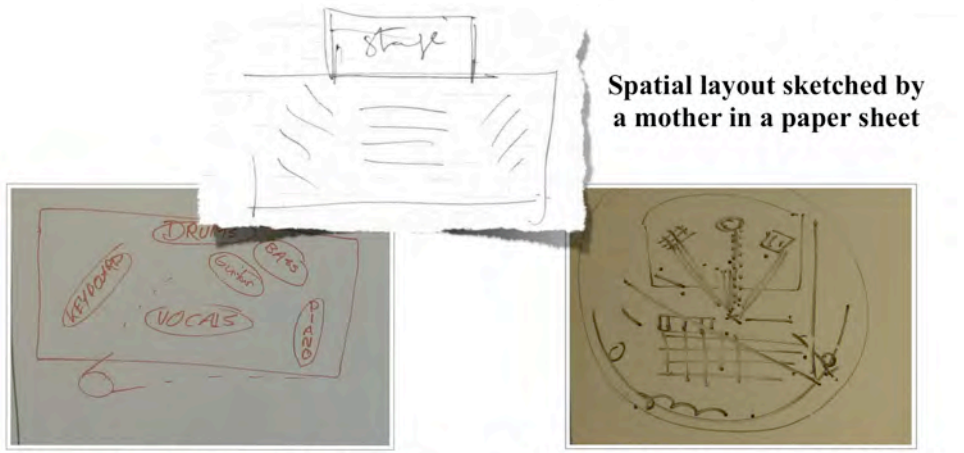


Figure 3.6. Sketches from participants illustrating the concert setup.

mentioned that it would be interesting to have a spatial representation of the concert venue to help browsing the event footage. When inquired about particular events they remembered, participants reported on solos performed by different musicians. Among the youngest participants, an event in particular was pointed out as the most memorable of the concert.

> *"I think that the jamming at the end I liked the most... I found that the most memorable of the whole evening..." (Friend of some performers)*

In some cases participants complained – and were desperate – when the quality of the video was not good enough or when the metadata was wrong (see Figure 3.7). Most participants expressed they would add/correct metadata with our system (Q1.3). However, they were quite resistant about the amount of time they would spend on this process, arguing that it demanded a lot of effort.

> *"It is not my problem (correct the wrong metadata)… people don't have time to play with the system." (Uncle of a performer)*

Figure 3.7. Participants' reactions during the evaluation process.

When questioned about the filter functionality, participants appreciated such feature because it would allow them to retrieve only the videos related to their interest. Nevertheless, almost all participants manifested interest in using a combination of filters, when searching for videos (e.g., show all videos of the trombone player in the 3rd song). Despite being feasible using the recommendation algorithm presented in Chapter 2, such functionality was not contemplated in the first version of our user interface. At last, some participants also mentioned that a person or instrument should be considered featured in a video only if this was a prominent shot, e.g., close-up or solo. They would not be interested in a video clip in which the subject of interest barely appears.

> *"If he (my nephew) is in the background but he is on the shadow it is OK but I would like to see a video in which he really shows up… My mother (performer's grandma) would not enjoy seeing this video of him because there is not much to see." (Uncle of a performer)*

3.1.2 Lessons Learned

In the first evaluation phase we followed an interactive approach, where a number of new requirements were defined. The most relevant observation was the necessity of providing contextual information for browsing, searching and watching community assets.

On the one hand, the thumbnail view did not show the temporal relationships between the video clips. On the other hand, the interface of the timeline view was considered complex. Participants were looking for a more intuitive and simple visualization model. We observed during the evaluation process that they tended to remember the inherent structure of the event (e.g., the concert program or spatial arrangement). Rather than treating each media asset as a discrete entity, archival theory and practice suggests that digital videos should be managed, preserved and presented to users in a way that reflects the social and documentary context in which they were originally embedded [8]. This argument leaded us to the specification of following requirement:

i. *Support inherent event structure:* users indicated the need for a more intuitive metaphor to organize or cluster community assets. Such approach would help them in exploring and searching for people or events of interest;

Although the interface allowed users to add/remove and correct existing annotations, these were not directly accessible. In order to see and change any information regarding a video clip (e.g., associated performers, songs or instruments), users had to click on a button to show the annotation interface (see Figure 3.4). When playing a video, the same problem was evident: annotations were again 'hidden' behind the media. In some situations, users would just click and watch a video in order to know more about its content. This was a time consuming process that led to frustration of the users. Based on these issues, we introduce our second requirement:

ii. *Make contextual information explicit:* feedback from users suggested that by clearly showing associated annotations, it would facilitate the browsing experience. It would also minimize the chance of 'blind' navigation or of getting 'lost' in the media space;

In the previous section we said that both thumbnail and timeline views offered a number of different filters for content selection. Despite appreciating this functionality, participants manifested interest in using more than one filter at the same time when searching for people or events of interest. The use of individual filters did not fulfill their needs. Based on this we present our next requirement:

iii. *Allow combination of filters:* users should be able to combine filters to compose robust queries. Such functionality would allow them to find videos of interest more effectively and faster;

Users feedback also suggested that they would like to have a spatial representation of the videos, in which content recorded from different angles could be activated in parallel. The work of Kennedy and Naaman [44] indicates that in a music scenario, like the one addressed in this thesis, alternative camera views could significantly reduce the required time to scan or to watch the content, while still providing a complete overview. In these lines, we introduce our last requirement:

iv. *Allow multi-camera navigation:* when watching a particular event (e.g., a solo), users should be able to switch between different camera angles (if there is any other available). Such functionality would enrich the browsing experience by providing spatial context.

In this section we introduced a set of functional requirements based on user feedback and results from the first evaluation phase. These new functional requirements motivated the design and evaluation of a second prototype system. In the next section we discuss our efforts for providing more effective socially-aware visualization mechanisms and innovative navigation paradigms.

3.2 Socially-Aware Media Browsing

The first prototype was helpful for better understanding user requirements for socially-aware video exploration of community assets. The evaluation results suggested we were in the right direction and helped in identifying a number of requirements for improving the user experience. With such requirements in mind, we started a new design from scratch. The browsing component discussed in this

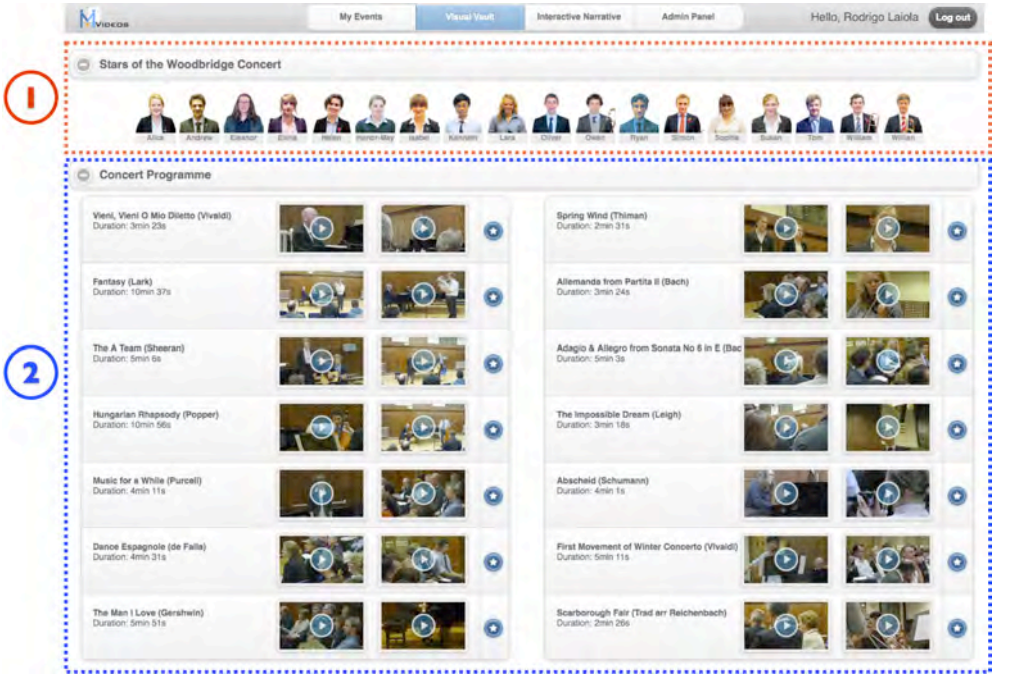


Figure 3.8. Browsing interface based on the concert program.

section intends to simplify the exploration of media assets, without compromising the flexibility of query specification [47].

To address our first functional user requirement, we designed an interface based on the concert program (Figure 3.8 (2)). This digital version of the original paper-based program handed out at the day of the event (Figure 3.9) clusters songs in two columns. Rather trivial in concept, it provides a general overview of the event schedule. In this interface, performers have a prominent position at the top (Figure 3.8 (1)). After all, these 'raising stars' are the main reason for users (friends and family) to use the system.

For each song in the concert program, a few video clips are recommended. This design choice provides contextual information without having to select a specific song. We also implemented a clip hovering functionality that shows a key frame animation on mouse over. It aims at providing a summary of the video

Figure 3.9. Paper-based concert program handed out at the day of the event.

without the need to watch it. These design decisions are directly aligned with our second requirement.

Hovering the mouse over interface elements (i.e., performers, songs and clips) also provides efficient and informative feedback. For instance, when a user hovers the mouse cursor over a performer thumbnail, the associated songs and media clips containing that person are highlighted in the user interface. This functionality, which has been designed to react in rapid response time, reduces the short-term memory load [47] and makes clear the relationship between performers, songs and clips.

Another functionality supported in the new prototype is the specification of queries based on performers. When the user clicks on a particular performer, the selection is sent to the server, which recalculates the recommendations considering the selection. Our design also allows for more complex query specifications such as the combination of two or more performers. In this case, a conjunction operator is used to connect the selection of performers. Thus, only songs (and the respective video clips) in which there is an intersection among the selected performers will be

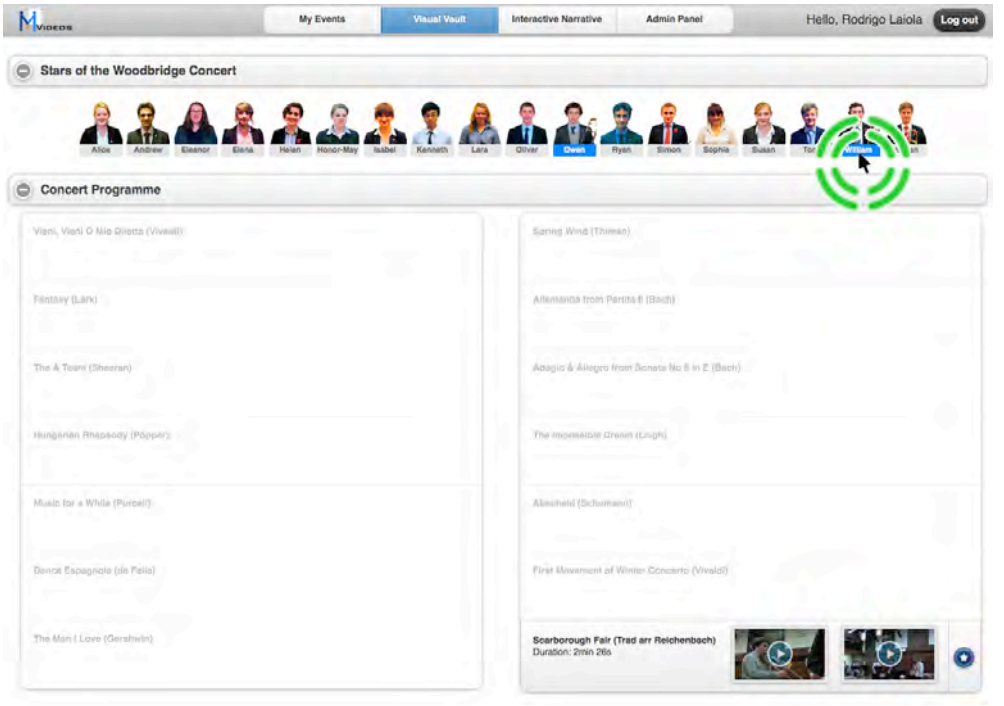


Figure 3.10. Supporting combination of filters. In this example, one performer is selected and the mouse cursor over another performer highlights the songs and video clips in which both performers played together.

highlighted in the interface (see Figure 3.10). This functionality addresses our third requirement by allowing participants to search for videos using combined filters.

Next we present our efforts on integrating context information, video playback, and supporting multi-camera navigation. As aforementioned, some video clips are listed in each song of the concert program. These videos are the entry points for media playback and multi-camera navigation. The video clip recommendations are based on the selected search terms and on the user profile (as we will see in the next chapter, the user profile is computed automatically considering user recording behavior).

When the user clicks on one of the recommended video clips, the playback interface is launched as illustrated in Figure 3.11. This interface is divided in three

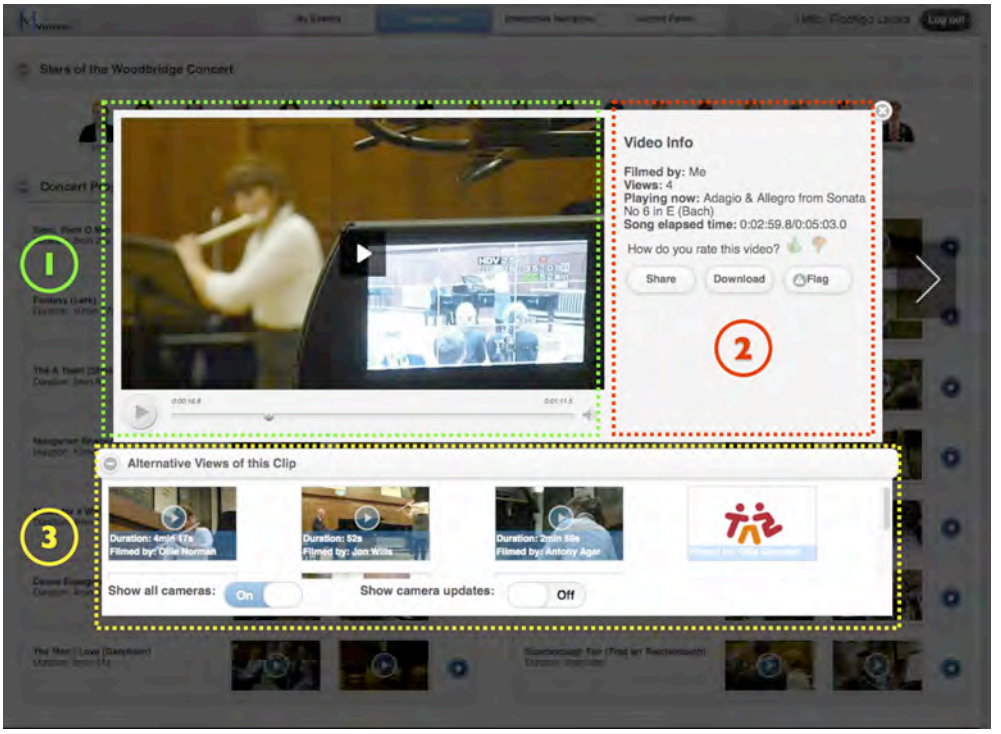


Figure 3.11. Interface for watching video clips.

main areas: media player (1), video clip information panel (2), and alternative views of a video clip (3). The information panel shows metadata associated to the video clip (e.g., who has recorded it, the number of views). It also provides information that is constantly updated based on the video playback (e.g., the song elapsed time). This panel also offers users a way to share a link of the current clip with someone by e-mail, to download the current clip, or to inform the system administrator that the clip has inappropriate content. The rationale behind this last functionality is to cope with the privacy concerns discussed in Chapter 2.

The area of alternative views of a video clip (Figure 3.11 (3)) presents other camera angles that happened at the same time of the main video. In other words, this area shows concurrent video clips recorded by other people during the event. By design choice, only a limited number of alternative views is presented (or

recommended) to the user. It is possible that more cameras were active at that point in time. The position of each camera is set when the player interface is launched and remains unchanged during playback.

When the user clicks on an alternative camera view, this will take the place of the main video, and the playback will continue from the same point in time as if the user had changed his position or angle. Such interface provides support for watching and navigating the media space, which directly addresses our fourth and last requirement about multi-camera navigation. Due to performance limitations in a Web browser environment, alternative videos are not played at the same time as the main clip. As an elegant workaround, our design provides a camera update functionality that – during the playback of the main clip, – periodically changes the key frame of each alternative. This approach aims to minimize the blind camera navigation problem discussed in Section 3.1.

3.3 Evaluation

Using the footage recorded during the Woodbridge high school concert (UK) in the beginning of November 2011, 13 people from 6 families participated in the evaluation of the new prototype system. While this section reports on the observations from the interactions of all the 13 participants, the quantitative data shows the answers from 9 people (the others did not fill in the evaluation questionnaires). More information about the methodology and participants' profiles can be found in Chapter 2. Next, we analyze the user responses and discuss the findings regarding socially-aware video exploration. Our results are based on a qualitative analysis of the interviews and observation of the system usage.

3.3.1 Results and Findings

Figure 3.12 and Figure 3.13 present the results of the questionnaires. Overall, participants appreciated the browsing interface (Q2.1). They indicated that it is useful for finding videos of performers and that it is better than traditional tools to explore the event media space (Q2.2 and Q2.3, respectively). Therefore, users would find videos more efficiently using our system (Q2.4 and Q2.5). If we compare with the results obtained in the initial evaluation (see Figure 3.5), there was a clear improvement, even though these were two distinct experiments.

> *"It (the browsing interface) has everything in one place and you can access other (people's) videos without having to import / open them." (Brother of a performer)*

The concert program metaphor was well assessed by our participants. In general, they expressed that this *inherent event structure* provides a simple and intuitive overview of what happened during the concert. Performers' thumbnails displayed at the top of the user interface were also appreciated. Participants said this was a good way to quickly look for videos they were interested in.

> *"Very easy to use! Performers at top is a good idea, and the concert programme is very clear!" (Father of a performer)*

When asked how much they liked the mouse over functionality in the concert program, 8 out of 9 participants said *a lot*, while the other participant said *some* (Q2.7). This was by far the most appreciated functionality of our prototype system. Participants enjoyed the rapid *contextual information* feedback when they hovered any of the interface elements (e.g., performers, songs or videos).

> *"I really liked the mouse over feature in the concert programme!" (Mother of a performer)*

> *"This is really good!" (Performer about filters and mouse over functionality)*

One of the participants mentioned that this mechanism was a bit slow though. Rapid response time is critical to support effective feedback. Providing highly responsive interactive results is important for dynamic browsing interfaces like ours, and fast response time for query reformulation allows the user to try multiple queries rapidly [47].

One aspect that needs further investigation is how to present recommendations for each song. Some users indicated that more recommendations could be showed: they assumed that there were more videos available. Apart from that, they seemed to like the video recommendations (see Q2.8). As mentioned earlier in this chapter, our video recommender takes into account the social bonds between users and performers. In the next chapter we detail the profiling approach used by our video recommender.

> *"I'm wondering why it (the browsing interface) particularly picked those 2 videos." (Father of a performer)*

While exploring videos displayed in the concert program, users expressed that they were having an engaging experience, but they did not have an option to play a song from begin to end. This feedback suggests the need for supporting more complex narrative alternatives that not only take into account the temporal alignment between videos, but also the preferences and social relationships of each individual user. This subject is the focus of the next chapter, which discusses the

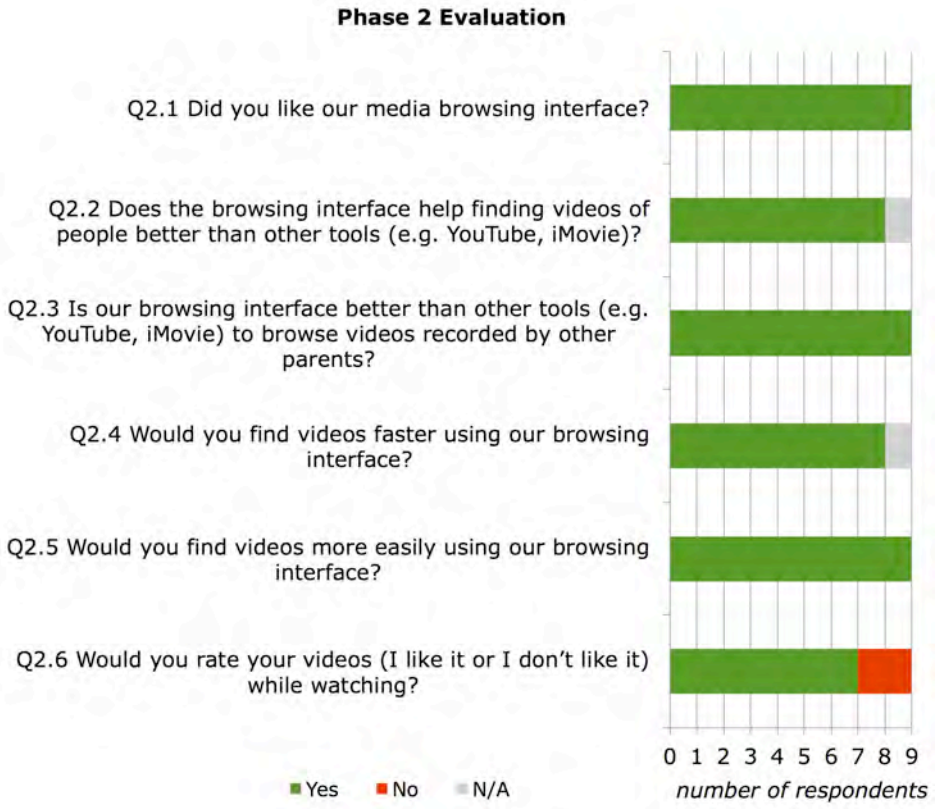


Figure 3.12. Results of the questionnaires from phase 2 evaluation.

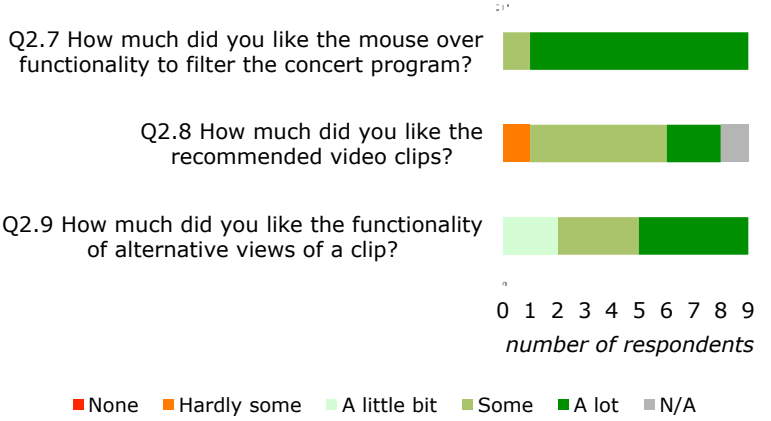


Figure 3.13. More results of questionnaires from phase 2 evaluation.

balance between automatic generation of video narratives and use of manual processes to reflect personal imprint.

> *"I think when you got individual songs, or individual pieces, [...] you might well want to, say, see the whole three minutes or something from the beginning." (Father of a performer)*

Our participants appreciated the *multi-camera navigation* support (Q2.9). This functionality raised the demand of having all the alternative videos playing at the same time and for seamless transition between camera angles. However, users also were aware of the browser and bandwidth limitations in our scenario. In some sessions there were some technical problems when switching from one camera view to another. Instead of starting the new clip from the current time, the playback would start a video from the beginning. This problem clearly led to frustration.

> *"I liked having a lot of different camera angles, which is something you don't get with anything else." (Performer)*

> *"Found the alternative views slightly complicated as regards case of use – couldn't always tell whereabouts in the performance we were, seemed a bit jumpy. Probably just an issue of getting to grips with the programme though!" (Performer)*

Still regarding the multi-camera navigation interface, some participants suggested that it would be nice to have a visual representation of the duration of each clip within the song, as it would help them to situate themselves temporally. This goes in the same direction of the work proposed by Yu et al. [77].

Regarding video annotation, 7 participants declared they would explicitly rate videos while watching (Q2.6).

> *"I would rate a clip while watching to tell it does not belong to a song or it has poor quality… just to make sure it would not be recommended again!" (Mother of a performer)*

> *"I would tag videos (thumbs up/down) as not being of good quality or in poor position e.g., performers face not visible as obscured by music stand." (Father of a performer)*

A few participants mentioned they normally do not use to rate videos at all.

> *"I never really use the rating features of YouTube." (Brother of a performer)*

3.4 Discussion

In this chapter we presented our efforts in designing and implementing an interface for browsing a collection of user-generated videos from a shared event. The interface aimed at helping users to easily access contents based on their social interests. This chapter described a two-phased development and experimentation.

First, we discussed the design and development of our initial prototype system. The evaluation of this tool allowed us to identify a number of functional user requirements for interacting with a set of videos from the same concert. These findings guided the development and evaluation of a new video browsing interface. Results from the experiments show that our new prototype satisfied the requirements and led to a clear improvement when compared to the initial system. Using a concert program metaphor (requirement i), participants could search for videos using combined filters (requirement iii) and experience moments of interest from different camera angles (requirement iv). Not to forget that our system provides efficient and informative feedback to help in this process (requirement ii).

Overall, our design decisions have improved the ability to explore videos users care about, among a pool containing the recordings of different parents. Our results clearly indicate that a socially-aware video exploration system like ours (which fulfills the *emotional intensity* guideline and social connectedness requirement introduced in Chapter 2) provides an improvement over current tools for accessing and navigating a repository of shared media assets. These results directly answer the research question asked in the beginning of this chapter.

Enabling users to explore an event and search for video clips they, and other participants, have recorded is an important step towards making personal media more accessible. But it is just the beginning. Individual video assets most of the times do not provide rewarding narrative experiences that help users remember important events. In the next chapter we discuss the balance between automatic and manual processes for creating personalized stories from community assets.

4

Automatic and Manual Processes for Creating Personalized Stories from Community Assets: Where is the Balance?[1]

In the previous chapter we have seen that the ability to search and browse content based on the social bonds is very important for making personal media more accessible. Nevertheless, it is too often the case that personal recordings are abandoned on memory cards or as downloaded files on hard drives never to be accessed again [19]. The main reason for this is that, as captured, video is not ready for being looked at. Video, as a time-based medium, necessarily requires processing after capture. Editing, for instance, can be performed on a handheld

[1] This chapter contains extracts from the following papers:

R.L. Guimarães, P. Cesar and D.C.A. Bulterman. 2013. Personalized Presentations from Community Assets. In Proceedings of the 19th Brazilian Symposium on Multimedia and the Web (WebMedia '13). ACM, New York, NY, USA, 257-264. DOI=10.1145/2526188.2526208 http://doi.acm.org/10.1145/2526188.2526208 (33% acceptance rate) [Won, best multimedia paper]

R.L. Guimarães. Automatic and manual processes in end-user multimedia authoring tools: where is the balance?. In Proceedings of the international conference on Multimedia (MM '10). ACM, New York, NY, USA, 1699-1700. DOI=10.1145/1873951.1874327 http://doi.acm.org/10.1145/1873951.1874327

V. Zsombori, M. Frantzis, R.L. Guimarães, M.F. Ursu, P. Cesar, I. Kegel, R. Craigie, and D.C.A. Bulterman. 2011. Automatic generation of video narratives from shared UGC. In Proceedings of the 22nd ACM conference on Hypertext and hypermedia (HT '11). ACM, New York, NY, USA, 325-334. DOI=10.1145/1995966.1996009 http://doi.acm.org/10.1145/1995966.1996009. (34% acceptance rate) [Nominated, best paper/best newcomer]

smartphone to trim out poor and redundant content that is always captured alongside quality material. This is because video carries complex information and quality judgments cannot always be made on the spot, while filming. Editing is also required to create attractive artifacts for what we believe could later become valuable memories we want to watch and share with friends and family members. A simple juxtaposition of recorded fragments does not necessarily result in attractive mementos. But editing is not a simple process and people often do not want to engage with it, vide the results provided by our participants in Chapter 2. This is true for personal content from one source, but it is especially true when considering mixing content recorded at a single event from many sources: the *community* video problem.

This chapter provides an analysis of our efforts on multimedia authoring using community assets. As with browsing and navigation, we have developed a first version of an authoring system, subjected it to an extensive long-term user testing, and then developed an improved version that follows the guidelines of socially-aware multimedia authoring. As described in Chapter 2, our initial work was subjected to a 10-month evaluation process, enabling end-users to create stories reusing collective content for individual needs. Our initial results showed a general enthusiasm from participants, which were validated in the first evaluation phase. The initial implementation, which was aligned with the *personal effort* guideline, made use of a narrative engine to automatically compile personalized stories based on the community media assets [76]. While the video compilations produced by the initial system were considered visually compelling, end-users missed the capability of personalizing those by adding their own 'imprint'. The complexity of authoring *personalized* stories from community assets have led to the consideration of the following research question:

Question 1.4 *Where is the balance between automatic and manual processes when authoring personalized narratives users care about?*

We have approached this research question from three more concrete and strongly interlinked perspectives. In particular, this chapter investigates:

1. The degree to which media authoring can be simplified by the use of a narrative engine to produce a 'rough cut' (an initial video story) automatically;

2. The degree to which this rough cut can be automatically tailored based on the relationships within an end-user's social network; and
3. The degree to which automatically generated video stories can be easily refined and further personalized using intuitive manual extensions with minimal extra effort.

The primary contribution of this chapter is a hybrid authoring system that allows users to create and share personalized media with others. This chapter is structured as follows. Section 4.1 motivates the problem of creating personalized stories from community assets, and discuss the evaluation we have carried out during the first phase. Section 4.2 describes the design and implementation of a new hybrid (or semi-automatic) authoring system that meets the functional user requirements elicited in phase 1. Section 4.3 reports on the results from the user evaluation of our prototype, demonstrating the benefits of our hybrid authoring approach. Finally, Section 4.4 concludes the chapter offering a discussion about the lessons learned.

4.1 Community-based Authoring

Creating compelling multimedia productions is a non-trivial problem. The problem is compounded when authors want to integrate community media assets: media fragments donated from a potentially wide and anonymous recording community. The purpose of this section is to describe our initial efforts to facilitate the creation of personalized stories from community assets.

Our initial approach provided users both independent manual and automatic authoring threads (called *Editor* and *Composer*, respectively). The intention was to compare the quality of easy-to-create fully automated compilations with the amount of effort required to manually creating personalized video stories.

Figure 4.1 shows Composer, the thread for automatically assembling video compilations in our initial prototype system. Users only had to explicitly select the subject matter (people, songs, instruments) and two other parameters (style and duration). Then, by pressing the 'GO' button, a narrative engine would be triggered, and in less than three minutes a video using the assets captured by different cameras at the concert would be created. The narrative engine would select the most appropriate fragments of videos from the repository, based on the declared user parameters, and assemble them following narrative constructs.

Figure 4.1. Initial prototype implementation for automatic video editing.

As mentioned in Chapter 2, the automatic authoring capabilities of the system were also assessed using expert input. Three video professionals with between 5 and 20 years experience were interviewed. All three agreed with the basic footage preparation and narrative structures that were used to build the video compilations. They were especially keen with the approach of using an audio track as a master timeline to drive the story development. They also concurred with our approach of automatically selecting alternative shots from cameras available in parallel tracks and using rules that selected clips based on shot types [76].

Our initial prototype system also provided an interface for manually creating video compilations (Editor). To find videos of interest, users could use the same set of filters and views available in the video exploration tool (refer to Chapter 3). Using the Editor, users could just drag and drop recommended video clips from the shared repository to the storyboard (see Figure 4.2). For example, a parent could add more clips in which his daughter was featured for sharing with grandma, or he could instead add a particularly 'funny' moment from the event when creating a version for his brother.

Figure 4.2. Initial prototype implementation for manually editing videos.

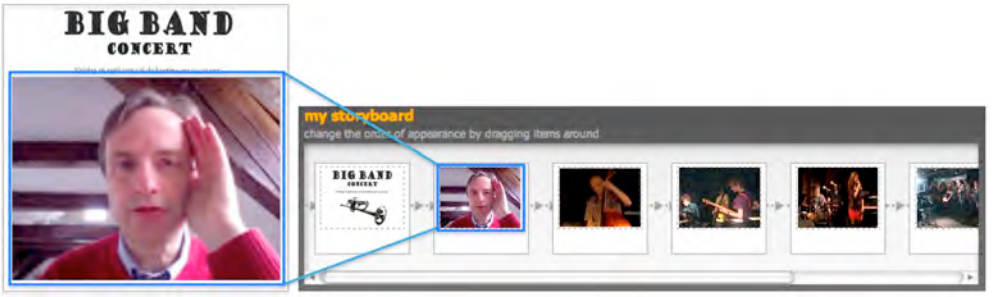


Figure 4.3. Elements of an authored video composition: the parent has included an introductory image and a video for making it more personal and intimate.

Apart from allowing fine-tuning of productions, the Editor also enabled users to perform enrichments. It provided mechanisms for including personal audio, video, and textual commentaries. For example, these could be subtitles aligned with the video clips commenting the event for others. Users could as well record an introductory audio or video, leading to more personalized stories. Figure 4.3 illustrates some elements of an authored video, where a parent has created his own version of a concert. He has also added some personal assets, such as an introductory image and a video recording of himself that acts as the message envelope. This functionality was called 'capture me'.

As reported in Chapter 2, the evaluation of the initial system was preceded by 3 social events. While the first two recording experiments mainly focused on the evaluation of the annotation processes and narrative structures, the third one, a school concert in Amsterdam, allowed us to engage a group of parents, relatives and friends of performers for evaluating the initial version of our system. In the remaining of this section we discuss the lessons learned about the authoring threads during the evaluation of the first phase.

4.1.1 Phase 1 Evaluation

In this study, all participants first interacted with the community-based browsing interface (see Chapter 3), and then they were introduced to the authoring threads. In general, they appreciated both approaches to create personalized video compilations and considered the functionalities useful. Using our authoring tool they felt they could create more stories faster and easier (if compared to traditional systems – Q1.1-Q1.3 in Figure 4.4 from the evaluation of the first phase). Overall, the automatic assembled videos were considered visually compelling (see reactions in Figure 4.5). Although participants also indicated that they would like to have more manual processes available to further personalize and fine-tune the video compilations (Q1.4-Q1.6).

> *"I want more portraits of my daughter (in this automatic generated compilation)... is it possible to edit an existing movie (in the Editor)?" (Father of a performer)*

In the manual authoring thread participants could find and select their favorite video clips. However, a complain was that they had to choose each and every clip for the compilation. Regarding optional processes, for most of our

participants the 'capture me' function – for including personal assets in a video compilation – was seen as a way to personalize videos for a target audience. As shown in the results (Q1.5), such functionality was mostly appreciated. Participants indicated they would use it, for instance, when creating a birthday present video.

In the initial version of our prototype system users could either generate video compilations automatically (not being able to change these later on) or edit manually (having total control but starting from scratch). While automatic compilations were quite appreciated because of shot selection and camera diversity, users provided important evidences that manual processes were indispensable to reflect intimacy and effort.

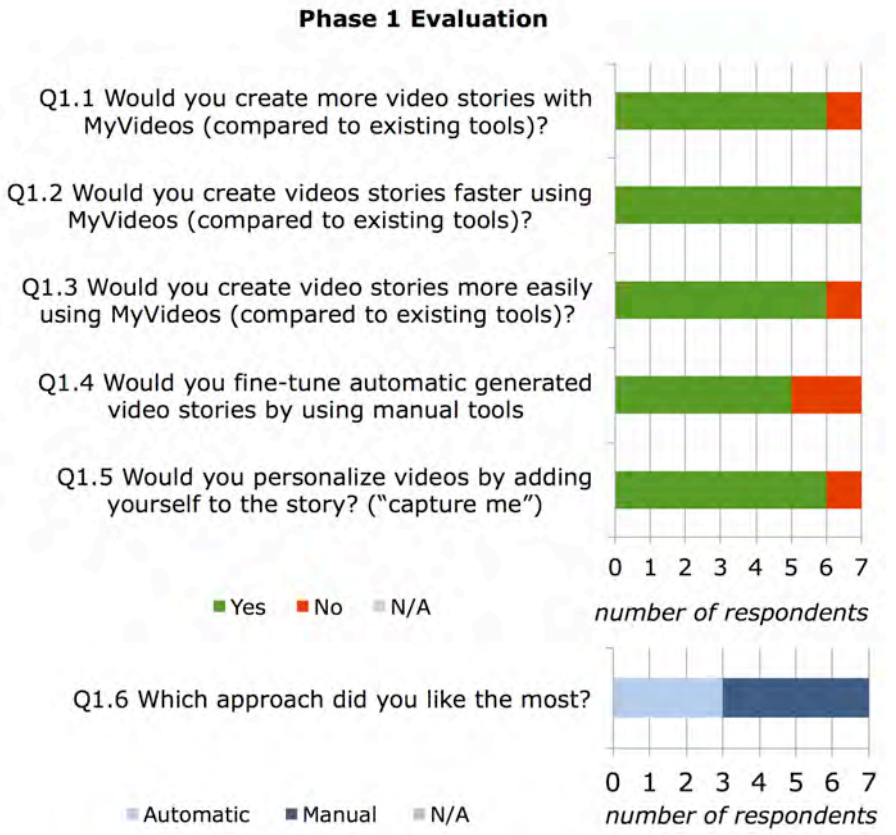


Figure 4.4. Results of the questionnaires from phase 1 evaluation.

Figure 4.5. Quotes of a participant using the automatic authoring thread.

Based on participants' comments, reactions, and answers to the questionnaires, we can conclude that they appreciated the benefits of our authoring system and considered it a valuable vehicle for creating enjoyable memories. While these results were highly relevant, we were aware that they were not complete. More importantly was the indication that instead of the automatic or the manual authoring thread, a hybrid solution would better fit the participants' needs (Q1.4). In the next section, we discuss the functional requirements that motivated the design of a new version of our socially-aware multimedia authoring system.

4.1.2 Requirements Gathering

Regarding manual authoring, participants identified a number of issues that could improve the creation process. Even though some participants were familiar with end-user video editing tools, for most of them this process was time consuming and complicated. Even though they appreciated the filtering functionalities included in the Editor, they indicated that they would not like to start the process from scratch.

Given the difficulties inherent in video editing, they would rather first use an automatic system that provided them with an already compiled story. Based on this feedback we introduce our first functional requirement:

i. *Not start from scratch:* users indicated their preference for an authoring paradigm, in which an initial narrative compilation would be created on their behalf. Such approach would simplify the authoring/editing task and increase their productivity;

Regarding automatic authoring, participants generally appreciated the easiness of use. The interface for automatically generating stories only required users to select a number of parameters such as duration, people, instruments, and songs to be shown in the compilation (see Figure 4.1). After a few minutes, users could watch a static narrative story based on their preferences. Even though they generally enjoyed the final results, they would have preferred that the system selected some of the parameters. In particular, they requested for automatic methods capable of identifying the interpersonal relationships with the performers of the concert. This discussion leads to our next requirement:

ii. *Consider implicit interpersonal relationships:* participants assumed that the system could automatically identify and process their interpersonal relationships with performers when creating video stories;

A common frustration with automatically generated videos in the initial prototype was that the automated process created a video story that could not be modified. Participants indicated that they would like to fine-tune (or personalize) automatic generated stories by using manual tools. They felt that the final result could potentially be more personal by adding assets and personal comments that more closely reflected their view of the event. This was of particular importance in video sharing situations, in which some participants wanted to send stories of the event to particular people within their social circle, such as an uncle or the grandmother of a performer. This result is consistent with our hypothesis that emotional intensity and intimacy should play a key role in socially-aware multimedia authoring systems (see Chapter 2). Geared by this discussion on personal effort we present our last requirement:

iii. *Allow for personal imprint:* participants suggested that automatically generated compilations could be modified. They wanted to remain in control over the final production, being able to make small changes. This approach would allow them to create more personalized stories.

Based on these requirements, we concluded that a new version of the authoring system was needed. The new approach would allow users to request a first compilation based on their implicit preferences and interpersonal relationships with performers. The system would then present an initial narrative, which could be edited and personalized on a per-clip basis. This hybrid authoring system ambitiously brings together both automatic and manual processes, so that narrative segments can be compiled, adjusted and edited successively. In the next section we discuss our efforts in designing and implementing such new authoring paradigm.

4.2 Hybrid Multimedia Authoring

The high-level workflow of our new authoring tool is detailed in Figure 4.6. Since we intend to improve the creation of video compilations based on multi-camera recordings, the input material still includes the school master track and the actual video clips that users agreed to upload. As shown in Chapter 3, all video clips are stored in a shared video repository that also serves as a media clip browser in which parents, students, and authorized family members can explore (and selectively annotate) the videos.

In the new design, the event exploration is the starting point for the authoring process. With the goal of creating a personalized video compilation based on a song, the user simply clicks on one of the songs in the concert program interface for triggering the narrative engine (Figure 4.7). The engine is in charge of creating a first montage from the video assets (and from video fragments) based on narrative structures and on interpersonal relationships (dependent on the identity of the user that is logged in). Such compilation, from now referred as the *Director's Cut*, can be later modified by the end-user for making it more personal.

Next, we discuss how the three requirements identified over our initial prototype have been considered in the design and implementation of the new authoring system.

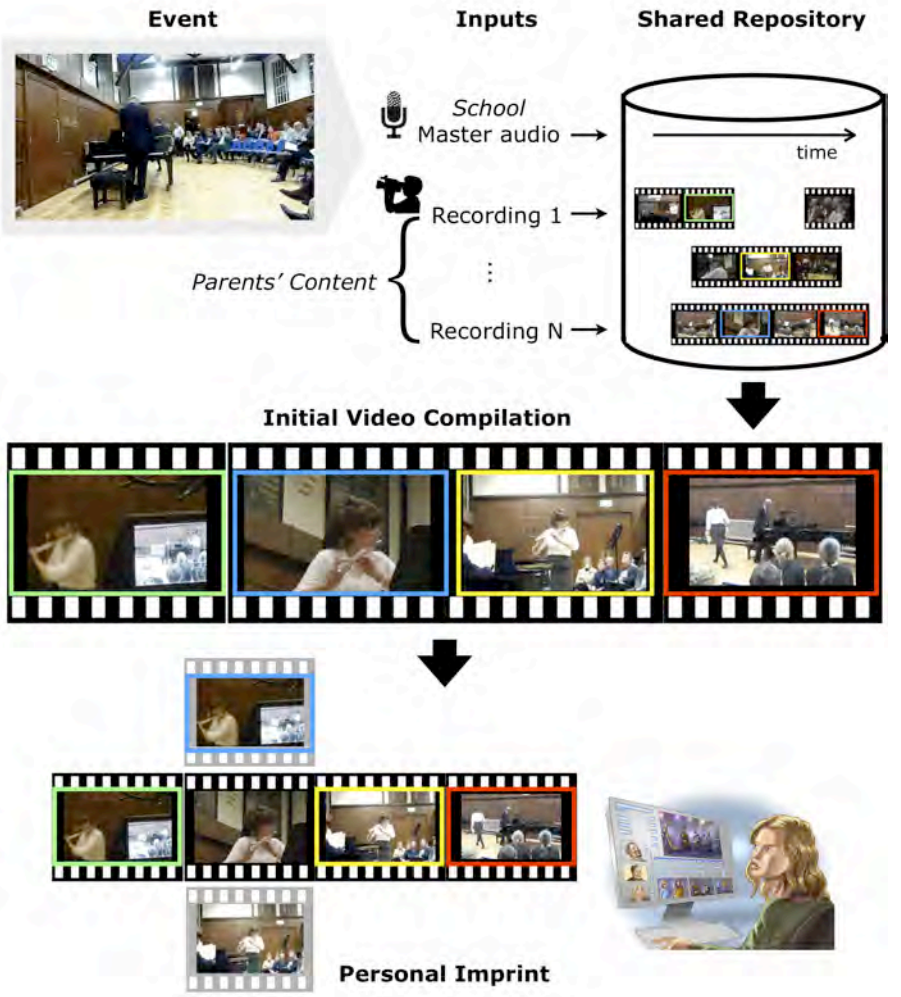


Figure 4.6. High-level workflow of our hybrid authoring tool.

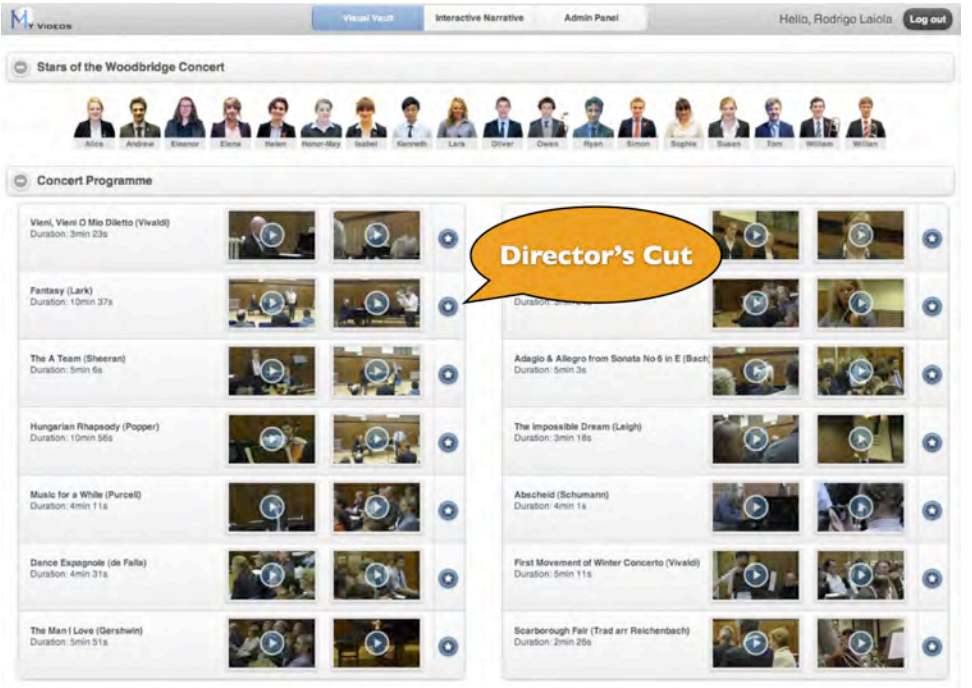


Figure 4.7. Triggering the Director's Cut from the browsing interface.

4.2.1 Profiling Users

Profile of users logged in the system can facilitate the automatic creation of personalized video compilations. Traditional ways of user profiling include implicit activity monitoring (log) and explicit insertion of personal data. While these approaches provide relevant results for a statistically significant group of people interacting during a long time span, they are not sufficient for our highly personalized environment. For this reason, we have implemented a mechanism to automatically compute the relationships between users and performers.

Such mechanism follows three steps. First, we fill a database table with the songs each performer participated in. This is done by inspecting the annotations

regarding performers in video clips, and looking for intersections with the songs that compose the event timeline. A key part of this procedure is that a weight is associated to each song/performer row in the database table. Such weight (or ranking) is calculated based on some parameters: the number of annotations each performer has in that particular song, the duration of these annotations (how long a musician is featured within a video clip), the quality of the annotated videos (e.g., high-definition or low-quality), and the shot type annotation (e.g., close-up or wide-shot). Note that the final ranking can be tweaked by giving different weights to the parameters. After all final weights have been calculated, they are normalized per song basis. This means that the performer with the highest weight in a song gets 1, while another that is not featured in the same song (or has the lowest weight) gets 0 (zero). All the other song/performer weights will then fall in the range [0, 1]. The result of this process is a table with normalized weights, which suggest the importance of each performer in each of the songs. The weighted song/performer table is used in the video selection process (compilation generation and alternative clip recommendations).

Second, we make use of the capturing behavior of each recorder individually. By taking into account the same parameters discussed in the first step, we model the behavior of a recorder towards the musicians in each of the songs. For that a similar database table, with an extra column (recorder) is used. By computing a normalized weight for each recorder towards each of the performers in each of the songs, we can derive their affection level, which as assumed, greatly influences the overall time a recorder spends capturing a specific musician. Based on these data, we can model relationships (was the performer his daughter? Was a friend of his daughter?), and thus provide information for the profiling process. Figure 4.8 shows the results of analyzing the metadata associated to the media captured during the high school concert in Amsterdam. In the figure, the recording behavior of a mother towards her kid is compared with the average behavior of the rest of the parents. We can observe that the affection level towards a performer is greatly influenced by the normalized weight of a particular recorder. In other words, the recording habits provide an important cue about the social relationships between recorders and featured performers.

Finally, the profiling process takes into account the user activity when browsing the shared media repository (e.g., videos a user watched, videos a user liked, most watched videos overall, most liked videos overall). This approach provides dynamic information when compared with the previous steps.

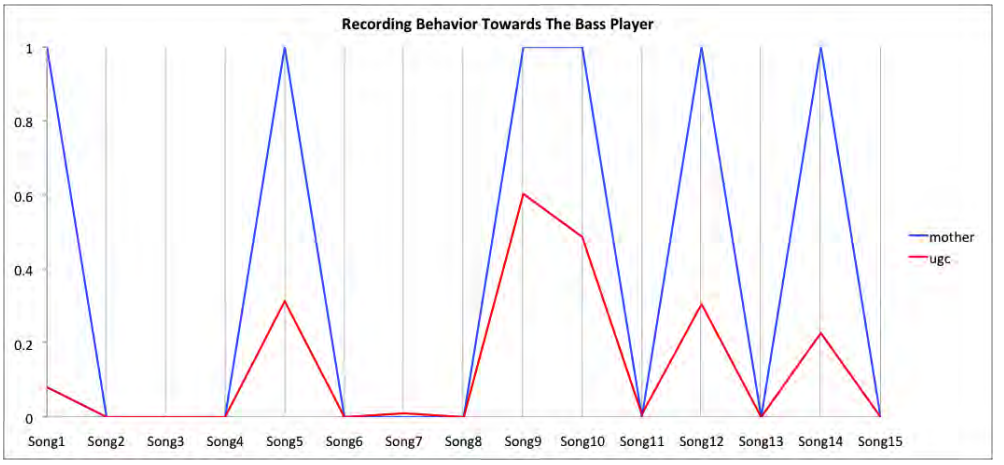


Figure 4.8. Comparison between the recording behavior of a mother towards her kid (in blue) and the average behavior of the rest of the parents (in red).

The information from these three steps is stored globally in the database and it is accessible by different engines. Based on normalized weights, inputs can be provided to the narrative engine, so automatic compilations do not only take into account narrative constructs, but as well *interpersonal relationships* between the users of the system and the people depicted in the video clips. This approach is directly aligned with our second requirement.

4.2.2 Automatic Generation of Stories

The first requirement identified in Section 4.1 was to provide automated authoring functionality, so the author does not have to *start from scratch*. Our system includes a reimplementation of the narrative engine used in the first phase. The new engine provides an initial story, as a playlist of video fragments. By itself, this functionality addresses our first functional requirement.

The narrative server wraps a narrative engine as a Web application, so that engine instances can be launched on the server. The Web application runs inside a generic Java Application Server (Tomcat) and it can handle request from other applications. These requests include the command dispatcher for starting/stopping the engine and the playlist dispatcher for requesting playlists. Further information

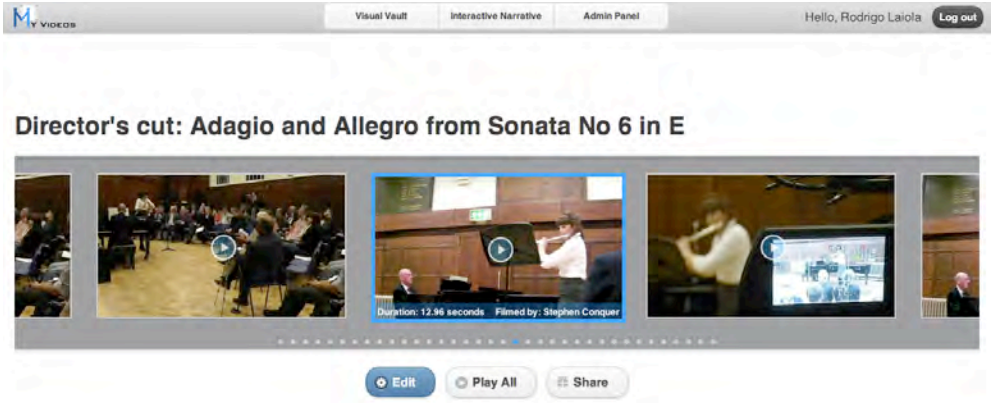


Figure 4.9. Director's Cut: an initial video compilation is created automatically by the system.

about the NSL language can be consulted elsewhere [51]. Figure 4.9 shows a video compilation created out of the 'Adagio and Allegro from Sonata No 6 in E' song.

As we will see below, the implementation of the narrative engine presented in Chapter 2 was modified to provide a set of alternatives (video clips) that can replace specific parts of the initial Director's Cut, while still maintaining the narrative structure and the story line.

4.2.3 End-User Personalization of Stories

The third requirement we identified was the need for fine-tuning and further personalizing the automatically generated productions. To support manual personalization, the narrative engine does not only create a Director's Cut, but it also provides a set of alternative clips that can potentially replace parts of the compilation (see Figure 4.10).

Once an initial compilation is ready, the user can modify it, allowing for *personal imprint* (third requirement). In order to enable such functionality we use a structured playlist format. In our work, we selected W3C's SMIL playlist profile [17]. The benefit of SMIL is that it aims at integrating a set of independent multimedia objects (in our case video fragments) into a synchronized multimedia

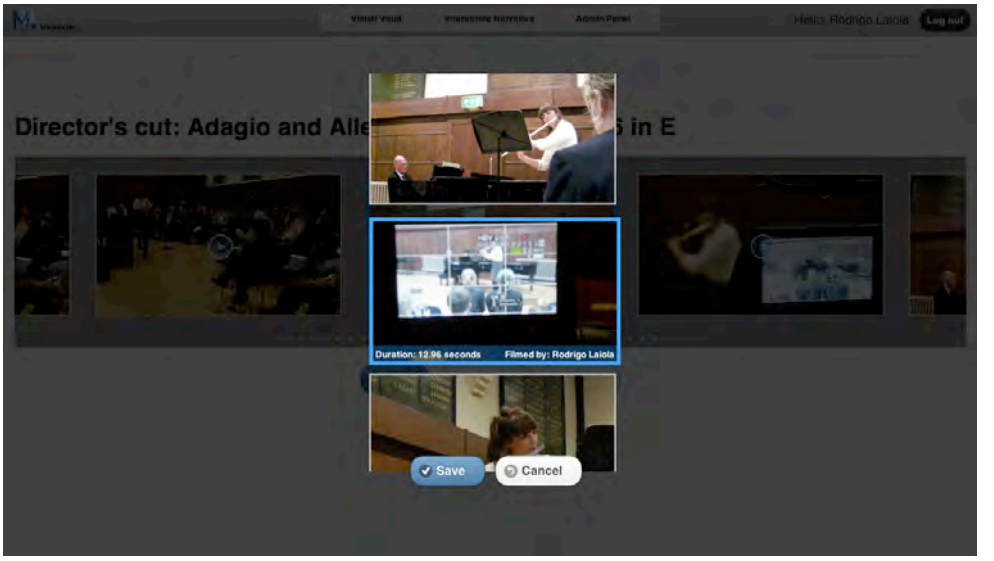


Figure 4.10. Director's Cut: visualizing alternative clips available.

presentation. It contains references to the media items, not the media content itself, and instructions on how those media items should be combined spatially and temporally. Other approaches on video mashups typically provide a final encoded video item, in which it is not possible to modify or enrich individual sequences. In our case, the richness of the SMIL language permits the user to perform dynamic operations on the initial video stories by simply modifying a text document (the SMIL file). The actual process of manipulating the document is hidden from the author, who simply sees an interactive user interface in the browser's Web page.

The video compilation generated by the narrative engine contains a set of references to video fragments (using `clipBegin` and `clipEnd` parameters). In addition, it provides a number of switch containers (`<switch>`) that contain the alternative clips (or set of clips), which can be selected for personalizing the initial story. Such alternative video clips have been selected by the narrative engine, so the narrative intent is not lost. For example, it will offer the option of selecting a different camera angle or of selecting a different point/person of interest. In addition to these features offered by the narrative engine, the end-user can decide to perform more radical modifications by adding other assets from the database or

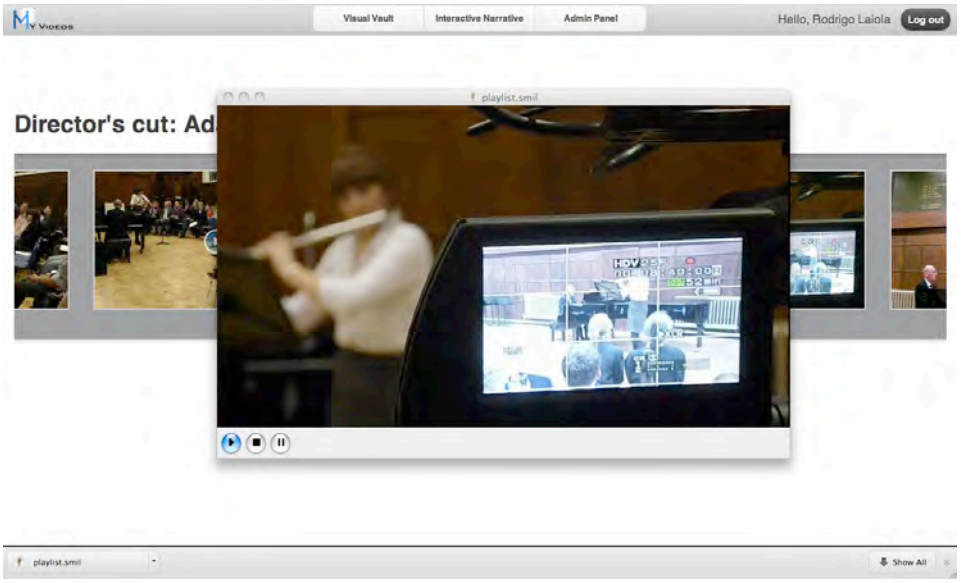


Figure 4.11. Director's Cut: song player.

by enriching the video compilation (e.g., adding comments). All these modifications will be incorporated into the original SMIL file. For viewing purposes we use the Ambulant Player, which provides a full implementation of the SMIL language (see Figure 4.11). The benefit of using SMIL is that the recipient of the video can easily further enrich and modify the video compilation, and send it to others or maybe return it to the original author, enabling reciprocity. In Chapter 5 we will discuss our efforts to support personalized end-user enrichment within third-party content.

The combination of a profiling infrastructure based on interpersonal relationships, a narrative engine capable of creating attractive video compilations, and the use of manual mechanisms for tweaking and personalizing such compilations results in a unique authoring tool. The validation of this authoring tool for creation of highly personalized (but compelling) productions characterizes the major contribution of this chapter, as reported in the next section.

4.3 Evaluation

Nine (9) participants, enrolled in the second phase of the evaluations, filled in the questionnaires about the Director's Cut functionality (for more information about the evaluation process, please refer to Chapter 2). Based on our observations, responses to the questionnaires, and analysis of the collected audio/video material from the interviews, in this section we present the results and discuss the findings from the evaluation process.

4.3.1 Results and Findings

Figure 4.12 shows the answers given by the participants after making use of the Director's Cut functionality. In general, all participants appreciated the new prototype (Q2.1). Six participants said that the Director's Cut offers a better way to edit videos if compared to existing video editing software they know (Q2.2). The other 3 users claimed they were unfamiliar with such tools, and therefore, they were unable to judge.

Again, similarly to the results obtained with the initial system (Q1.1-Q1.3), almost all participants argued that they would create more video stories (Q2.3) and quicker (Q2.4) because the tool was easy to use (Q2.5).

> *"It was very easy to use and it selected which videos I wanted well." (Brother of a performer about the automatic generation component)*

> *"Very easy to use (editing based on alternative clips). I wouldn't want to spend hours looking at a help menu. This was simple enough for me." (Mother of a performer)*

When asked whether they would add themselves to personalize a story (Q2.6), 6 users, mainly youngsters, mentioned this would be a good functionality. However, our senior participants claimed they would not do so. A similar feedback was obtained in the first evaluation process (Q1.6).

> *"It would be interesting to have a functionality to add other videos (that not only the ones suggested)." (Father of a performer)*

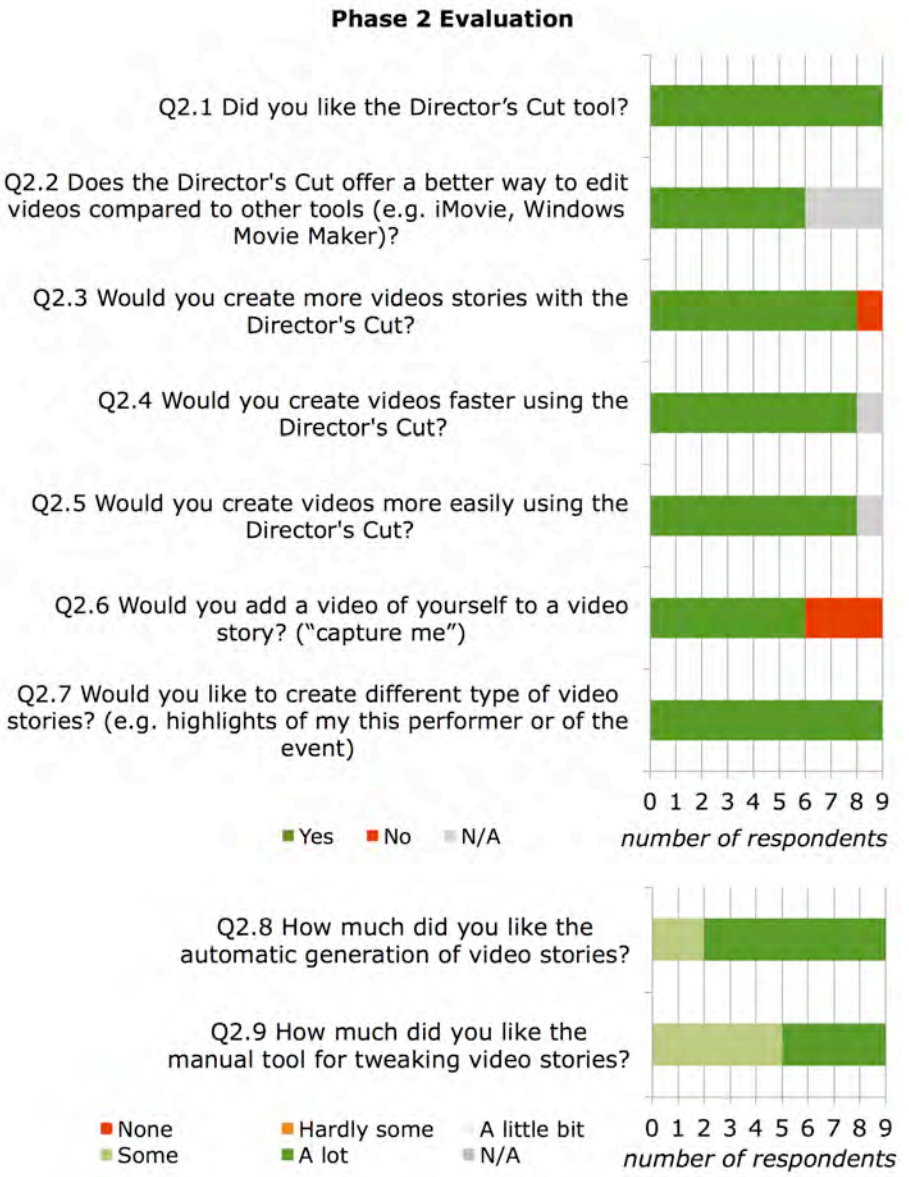


Figure 4.12. Results of the questionnaires in Woodbridge (UK).

Figure 4.13. Participants interacting with the Director's Cut.

All participants indicated they would like to create different types of productions (Q2.7). When questioned about the types of video stories they envisaged, the 'song-based' video came up as first choice among most of them. Some argued that depending on the social situation they would create and share different versions with different audiences.

> *"If my family misses my performance I would send the full performance to them… but if I want to send it to my singing teacher I would share a more focused version." (Performer)*

Figure 4.13 shows some participants during the evaluation process. A one-to-one comparison between the first and the second phases would be unfair (different users, events, tools). What we can say is that in the second evaluation both the automatic generation of initial video stories and the manual tools for tweaking had extremely good scores (Q2.8 and Q2.9 respectively). These results provide strong evidences that a hybrid framework builds on the best of each approach: assisting on complex tasks (*start from scratch*) but still making sure the user plays an active role in the process whenever desired (*personal imprint*).

4.4 Discussion

Creating compelling multimedia presentations remains a complex task. This is true for both professional and personal content. For professional content, extensive production support is typically available during creation. Content assets are well structured, content fragments are professionally produced with high quality, and production assets are often highly annotated (within the scope of the production model). For personal content, nearly none of these conditions exist: content is a collection of assets that are structured only by linear recording time, of mediocre technical quality (on an absolute scale), and with only basic automatic annotations.

The problem is made worse when authors use community assets of an event. In events such as high school concerts, a single concert can generate hundreds of video clips, taken from multiple vantage points, using tens of cameras. With our initial prototype we could generate syntactically correct automated stories that served generic needs (much like a conventional video mashup). Our users found these compilations compelling but not their own: they missed a personal touch.

In this chapter, we reported on a hybrid authoring approach that provides mixed support for automated creation (requirement i) and manual enhancement of personalized video stories (requirement iii). We targeted small-scale events, where lightly annotated assets are provided. Our assumption is that editors at these events will want to highlight personal aspects: a particular instrument, a particular child, a particular solo (or goal). This places demand on a system to help users to select appropriate content of personal interest (requirement ii), and to help build compelling stories with minimum effort (in accordance with the *personal effort* guideline presented in Chapter 2).

We acknowledge there are some limitations regarding the amount of automated personalization that a system can provide. Abstractly, given unlimited personalized annotations and unlimited information on all members of a potential target user community, we suspect that great strides could be made in automated personalization. The reality is, however, that for community assets, personalized annotations are limited, and the target user group is lightly profiled. This requires an interface that allows direct user intervention in creating content.

Providing direct user intervention has tremendous benefits: the user best knows his/her target audience. The differences between uncle Henry's interest and those of Grandma are often clear in the head of the human author, but largely inaccessible to an automated system. At the same time, end-users have only a limited amount of time and energy to create personalized stories (many are busy

recording new content, rather than editing old content!). This requires a balance of complexity and functionality. We feel that our approach provides this balance. Based on user feedback as part of our four-year study, we feel that we have shown that it is possible to satisfy casual content creators while still allow extensive personalization to take place if needed. These results directly answer our research question (and fulfills the requirements on effortless interaction, personal effort and intimacy introduced in Chapter 2). We feel that the combination of automatic and manual processes is unique and powerful.

While concentrating in the creation process, we cannot forget that multimedia sharing can also stimulate user comments and reactions, which is as well part of the authoring workflow. This is the topic of next chapter, in which we present our efforts on empowering users in commenting within personalized multimedia presentations.

5

Supporting Personalized End-User Comments within Third-Party Online Videos[1]

In the previous chapters, we have reported on our efforts to empower end-users to browse a shared video collection based on personal interests and to create personalized, but still compelling, personal stories from it. In this chapter, we now shift our focus from the author to the recipient of the story.

Successful commercial video sharing systems have provided ample proof that video is a first-class Web object. Even social networks like Facebook, originally conceived for status updating, have become important distribution channels for both consumer and professionally generated video [73]. In these sharing systems, video content serves both as a medium for communicating a story (using implicit or explicit cinematic rules), and as a catalyst for communication between third-party viewers of that content [12][25][50].

[1] This chapter is based on the following papers:

R.L. Guimarães, P. Cesar, and D.C.A. Bulterman. 2012. "Let me comment on your video": supporting personalized end-user comments within third-party online videos. In Proceedings of the 18th Brazilian Symposium on Multimedia and the Web (WebMedia '12). ACM, New York, NY, USA, 253-260. DOI=10.1145/2382636.2382690 http://doi.acm.org/10.1145/2382636.2382690 (30% acceptance rate)

R.L. Guimarães, P. Cesar, and D.C.A. Bulterman. 2010. Creating and sharing personalized time-based annotations of videos on the web. In Proceedings of the 10th ACM symposium on Document engineering (DocEng '10). ACM, New York, NY, USA, 27-36. DOI=10.1145/1860559.1860567 http://doi.acm.org/10.1145/1860559.1860567 (31% acceptance rate)

Recent developments by video service providers have extended the means for third-party communication in ways that have never been possible with conventional broadcast or personal video systems. In addition to the base video content, a typical YouTube page also provides space for end-user generated comments (Figure 5.1). These include implicit forms of commentary (such as the number of views or anonymous ratings, e.g., 'like' or 'dislike'), and explicit comments for interpreted viewers.

In the case of online video on demand, textual comments are usually statically placed underneath the media player. If desired, users need to make explicit any reference to a particular event that happens within the video object (e.g., "Look at that shiny, beautiful trombone at 1:56" in Figure 5.1). In YouTube, for example, when a user writes out a particular time code in the comment, it automatically turns into a 'temporal hyperlink', that when clicked takes the interested viewer to that part of the video. However, such comments do not reproduce the *'commenting while watching'* activity people perform when consuming media together. In general, users cannot add comments that are synchronized with the video, unless the owner (who uploaded it) has given editing rights to the base video content.

Primarily, this chapter considers the scenario in which a recipient of the content – not necessarily the owner of the video or who created a personal video story, adds personalized comments that are synchronized to specific events within the video. By *personalized* we mean comments created to highlight a particular event that is interesting to, for instance, the end-user social circle. By *synchronized*, we mean that such comments will be rendered during video playback at the time such particular event happens, unlike the static comments displayed underneath, as in YouTube or Facebook. Supporting this functionality, which is aligned with the *intimacy* and *reciprocity* guidelines specified in Chapter 2, we expect to reproduce asynchronously the commenting experience people have when watching media together. In this direction, we have asked the following research question:

Question 1.5 *Does the support for timed end-user commenting within pre-authored narratives provide an identifiable improvement over current media commenting approaches?*

Motivated by a survey research on current media watching and commenting practices, this chapter reports on the design, implementation and user-centric evaluation of a video commenting paradigm for structuring synchronized

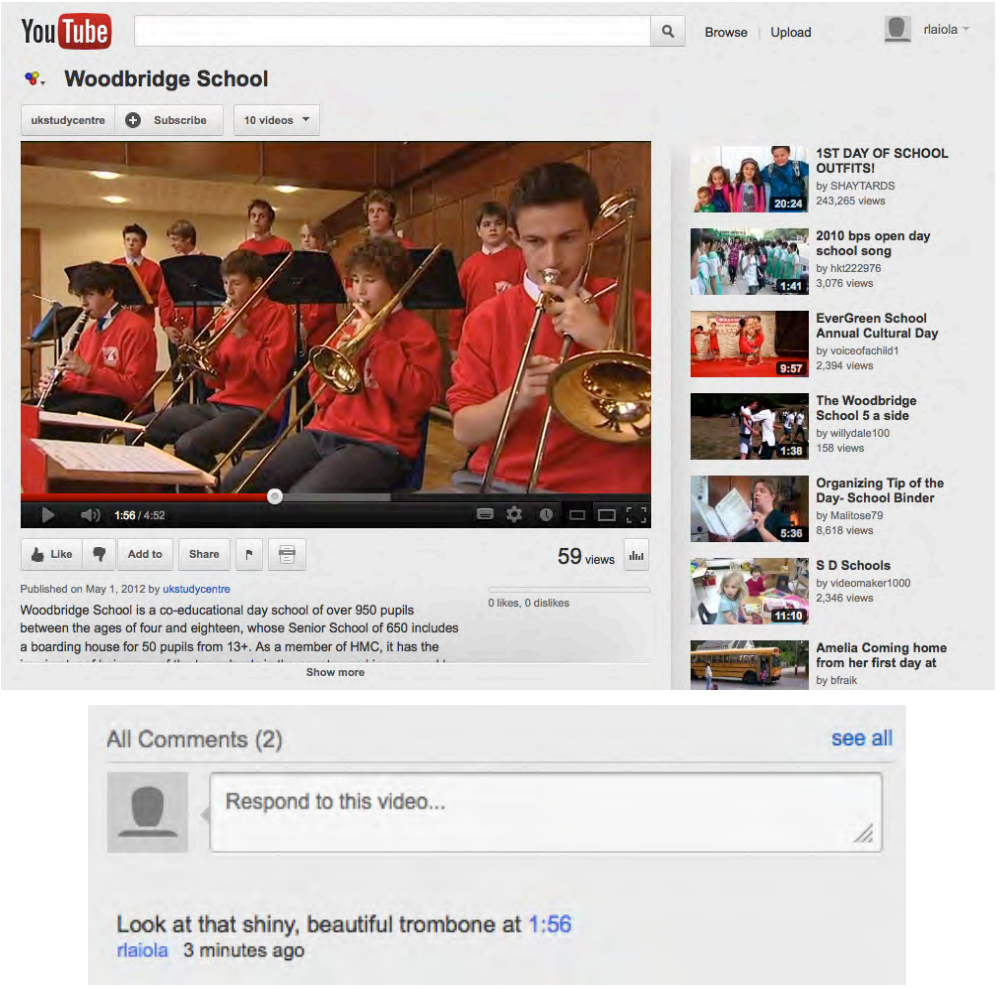


Figure 5.1. Typical end-user comment in YouTube. It appears statically underneath the third-party video.

comments within media. Our results indicate that users appreciate the functionalities of our system and find it better to comment when compared to current video commenting tools.

In order to realize our video commenting system, we also specify and describe a set of temporal transformations for multimedia documents. Our approach, unlike current solutions, allows end-users to create and share personalized timed text comments within third-party online videos. It also permits end-users to identify temporal navigation points by using hyperlinks within comments, and to associate contextual metadata (e.g., who wrote the comment and when). The benefit over current solutions lays in the usage of a rich commenting format that is not embedded into a specific video encoding format.

This chapter is structured as follows. Section 5.1 motivates our work, while Section 5.2 proposes a set of multimedia document transformations that allow end-users to add timed comments within third-party Web videos. Section 5.3 describes the design and implementation of a Web-based video commenting tool, which realizes such document transformations. In this section we also report on a predictive timing model for helping users to incidentally synchronize text comments with specific events within a video. Lastly, in Section 5.4 we present the results from the evaluation process, while in Section 5.5 we discuss the lessons learned and how these fit in the context of this thesis work.

5.1 Media Consumption and Commenting Practices

A sample group of 21 people were invited to participate in an evaluation process during the first quarter of 2012[2]. All participants were regular Internet users. Eighteen (18) people were in the 21-40-age range, while the other 3 were over 40 years old. Participants were from different nationalities including Brazilian, Chinese, Dutch, German, Hungarian and Irish.

We used semi-structured electronic questionnaires to collect users' feedback. While multiple-choice questions allowed us to explore patters and find trends (quantitative methods), open-ended questions aimed at capturing further insights into participants' opinions and perceptions. The user study was divided in 3 parts. The first part, which is the focus of this section, consisted of a questionnaire to gather background information about respondents' commenting practices when watching video content. Feedback answers were anonymous.

[2] This was an independent study and it counted with a different set of participants from the ones involved in the evaluation process discussed in the previous chapters.

5.1.1 Survey Research

Figure 5.2 summarizes the results obtained in our survey about media consumption and commenting habits. As users' practices were different, for each question we present the weighted average (colored column) and the respective standard deviation (bar). In Figure 5.2 the questions also have been clustered in two groups according to the consumption experience: synchronous (blue) and asynchronous (red) watching. For each scenario we also asked about participants' conversational and commenting practices around media.

A wide range of TV watching habits has been reported by our participants. In average, our users watch TV *every week* (Q1.1). This was the second highest frequency score among our questions. Participants also reported having the habit (between *occasionally* and *every week*) of talking with family and friends about a TV show they have watched (Q1.2). When asked about the frequency they would converse about a TV program with collocated people while watching, the average answer was *occasionally* (Q1.3). The lowest score though was obtained in the question about how often they would send tweets related to a TV show they were watching (Q1.4). As reported elsewhere [13], this activity is becoming popular over the years and, in some cases, it can be used as an interactive return channel in which the audience can influence on live TV programs.

In the second media consumption scenario, we asked our participants about the habit of watching live video feeds on the Web (e.g., Justin.tv, Ustream.tv). The average of their feedback was around *occasionally* (Q1.5). Regarding the activity of commenting on the video event while watching, we asked how often they make use of the built-in open textual chat rooms generally available on those services. Again, rather small the frequency stayed between *never* and *occasionally*, which was slightly higher than the one reported for tweet messages (Q1.6).

Regarding on demand (asynchronous) watching, we asked our participants about the usage of YouTube, Facebook and SoundCloud. Validating previous research [73], YouTube was often (between *every week* and *every day*) used by our participants to watch online videos (Q1.7). However, posting comments to the video page did not seem to be a common practice among our participants (Q1.8). In conjunction with the use of YouTube, we also witnessed a fair high frequency of video viewing on social networking sites (Q1.9). In this case though, participants habitually comment more on videos when compared to the comments added in YouTube (Q1.10). One possible explanation for this behavior is that participants are more likely to post comments within their social circle than in the open.

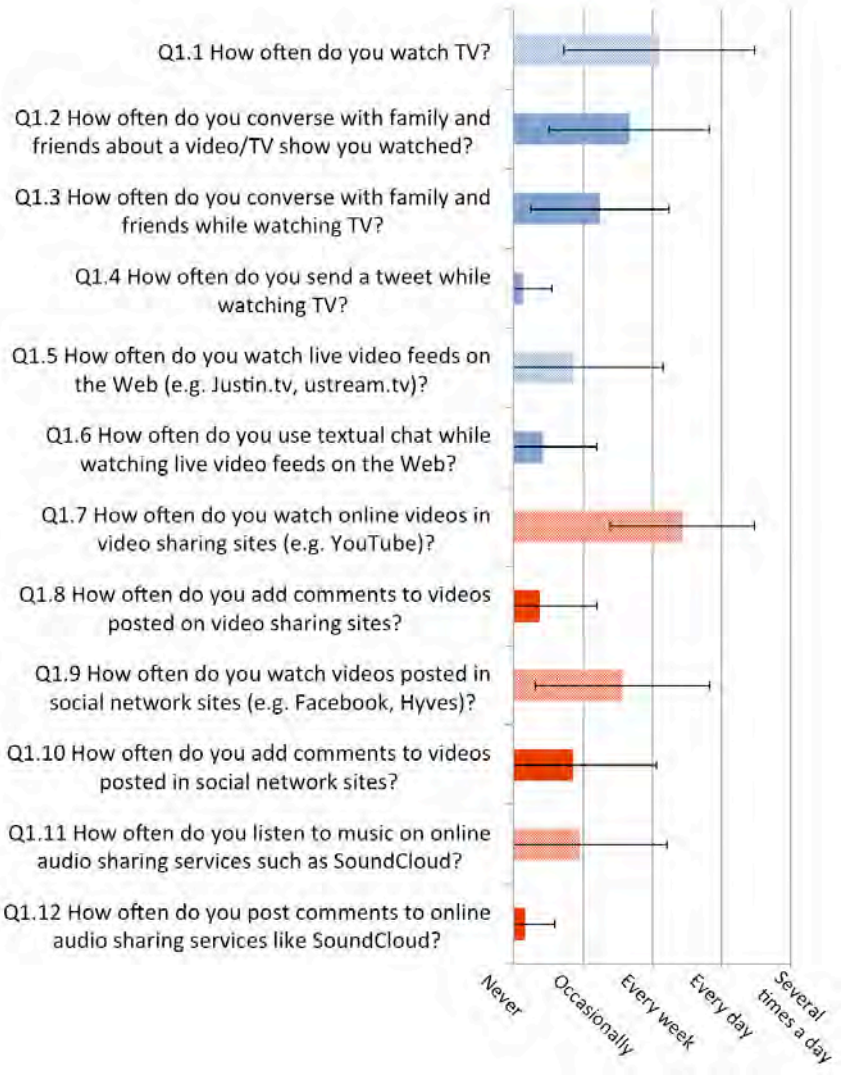


Figure 5.2. Survey research about media consumption and commenting practices. Blue columns indicate synchronous watching and related conversational habits. In red, on demand (asynchronous) consumption and commenting.

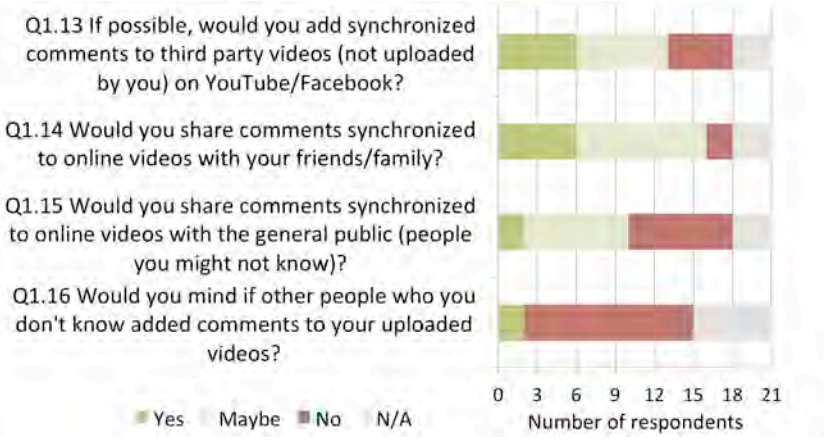


Figure 5.3. Requirements gathering: utility and usefulness of timed comments within media.

Finally, we asked our users to report on their practices using online audio streaming services such as SoundCloud. In average, respondents *occasionally* listen to music on these platforms (Q1.11) and they do not have the habit of adding timed comments to songs (Q1.12).

In the second part of the survey we asked our participants about the possibility of adding timed text comments to particular events within a third-party online video (see Figure 5.3). Thirteen (13) out of 21 participants said they would possibly (*yes* or *maybe*) add timed comments to YouTube or Facebook videos if they could (Q1.13). One of the participants expressed that such a feature would be a *"nice way to highlight sections/points of the video"*. When asked whether they would share these synchronized comments within their social circle the number raised to 16 out of 21 (Q1.14), fairly higher than the number reported for sharing comments with everyone (Q1.15). In one case, a user justified by writing *"I'm never interested in sharing my comments with the public... but to have a link that I just could send to friends"*.

At last, we asked a question related to digital rights and ownership. In this case, only 2 participants expressed they would mind if other people could add timed comments to their videos (Q1.16). In these lines, one participant highlighted the necessity of having control over the commenting activity: *"If everybody could add comments to any video it would become a real mess. Some people would use it*

to damage the image of others". This result seems to contradict the privacy issues discussed in the previous chapters, but it is important to keep in mind that here the videos were not necessarily personal (as opposed to the ones discussed in the extensive long-term evaluation process in the UK and the Netherlands).

5.1.2 Requirements Gathering

In the study presented above we looked at media consumption and commenting practices of a group of Internet users using different applications. Even though the group was small and provided results with a high variance, we obtained strong indications that people consume media and doing so, they eventually comment and share such moments within their close circle and, sometimes, with the general public. Our respondents also appreciated the utility and usefulness of synchronized comments, as they would comment on particular events within media (Q1.13-Q1.15). These results led to the specification of the requirements described below. These were used as the basis for designing and validating the online video commenting system presented in the remaining of this chapter.

i. *Retain base video integrity:* viewers should not be able to alter the base video content, either in terms of adding embedded captions/comments or of providing visual overlays on the base content — this right is reserved to the content owner;
ii. *Allow multiple-video aggregation:* in certain occasions, end-users might watch a collection of videos that are played as a continuous playlist (e.g., a personalized video story or compilation, as shown in the previous chapter). In these cases, end-users should be able to create comments that would span across the multiple videos composing the playlist;
iii. *Allow multiple-provider integration:* the user should not be locked into a single video service provider (or source) for candidate content, but should be able to populate a playlist from multiple sources;
iv. *Allow timed end-user comments:* when watching an online video, viewers should be able to add comments that are time synchronized. This feature would reproduce (asynchronously) the watching and commenting activity people have when watching media collocated;
v. *Allow micro-personalized timed comments:* end-users should be able to create different sets of time-based comments for individual

users/communities, or share these as 'broadcast' comments (similar to existing approaches in YouTube and similar systems);

vi. *Allow selective end-user viewing:* end-users might be able to select and watch comments by specific individuals and/or user communities, by topic etc. This is important because some comments might be targeted to individual users while others might be intended to the general public; and

vii. *Allow timed end-user navigation:* end-user comments should be able to include direct navigation support via timed anchors in the text content. This will allow others to navigate to other interesting parts within the same collection or to link to external media.

5.2 Media Commenting meets Multimedia Document Engineering

To address the requirements discussed above, we modeled the problem of creating timed comments within online videos from a multimedia document engineering perspective, and thus identified a set of document transformations. By document transformations we mean manipulations that can be applied to add non-embedded, flexible temporal end-user comments. Video commenting has been dealt with in many ways, ranging from the usage of models that are not timed (e.g., HTML) or are unstructured (e.g., Flash) to standards such as MPEG-7 [1] and NCL [33]. Based on our analysis [62], we rely on SMIL 3.0 [17] as the basic framework that meets our requirements. First, we create a structured multimedia document around an input video(s). The document model of SMIL 3.0 retains the base video integrity, and it allows multiple-video aggregation and multiple-provider integration. Timed text content and temporal hyperlinks allow end-users to add synchronized comments and to include timed end-user navigation points. Contextual information allows targeting timed comments to different audiences. Finally, the structured underlying model enables selective viewing.

5.2.1 Document Model

SMIL can integrate and compose a collection of audio, graphics, image, text, and video media items into a single presentation. As Web resources are distributed by nature – and might be very large in size –, in SMIL media objects are included by reference (using a URI - *Uniform Resource Identifier*). SMIL defines a single generic media object (`<ref>` element) that allows the integration of external

media resources into a SMIL presentation. However, it is also possible to use more intuitive tags when referencing external media resources (e.g., the `<video>` element is a more specific alias for the generic SMIL media reference element). Note that as an implication of the use of references, the integrity of the base media is preserved, meeting requirement i.

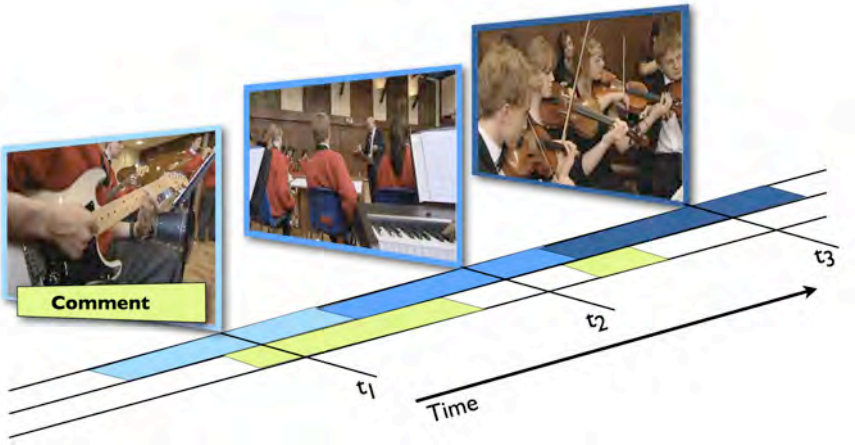
In addition, SMIL provides a powerful hierarchical composition model from which individual presentation timelines can be generated. The main temporal structuring elements are the parallel (`<par>`) and sequential (`<seq>`) containers, each of which provides a local time base for scheduling media objects (e.g., external videos) or children time containers. By using such time containers, it is possible to combine videos and comments in different temporal ways, as illustrated in Figure 5.4. In this example, three videos, stored in different video servers, are rendered as a continuous video, while the comments span across the videos. The structured temporal container behavior satisfies requirements ii and iii.

5.2.2 Timed Text Content

Unlike most text formats [15], text content in SMIL is not only constrained by its style and layout capabilities, but also by the temporal context of the presentation. For instance, text must be rendered simultaneously with related objects, and it must be hidden when these are finished.

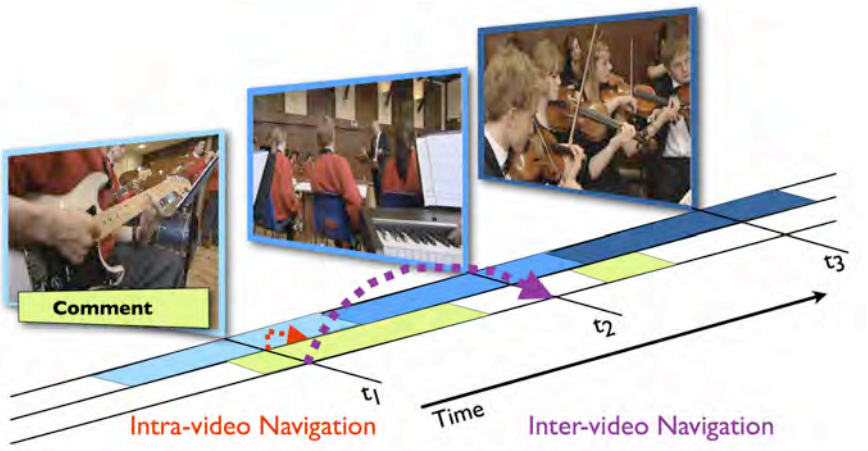
Authors can define small amounts of lightly formatted text containing embedded temporal markup within the context of a SMIL presentation. Such text may be used for labels within a presentation or for incidental comments or foreign-language subtitles. It is also possible to use large amounts of structured text (with or without temporal markup), but in this case it is recommended the use of SMILText as a text media object, or the use of objects encoded in formats such as XHTML or DFXP (*Distribution Format eXchange Profile*) [27].

The SMILText also define a set of additional elements and attributes to control timed text rendering (see Figure 5.4). All SMILText content is processed in a manner consistent with other SMIL media. The SMILText profile also allows SMILText to be used as an external format. Moreover, since the `smilText` elements and attributes are defined in a series of modules, designers of other markup languages may reuse these modules when they wish to include a simple form of timed text functionality into their language. SMILText, as a text container with an explicit content model for defining timed text, meets requirement iv.




```
<smil ...>
 <head ...> ... </head>
 <body>
  <par>
   <seq>
    <video region="r1" id="v1" src="YouTube video"></video>
    <video region="r1" id="v2" src="Vimeo video"></video>
    <video region="r1" id="v3" src="Facebook video"></video>
   </seq>
   <smilText region="c1"...>
    ...
    <clear begin="t1-5s"/>
    That's an awesome guitar solo!
    <clear begin="t1+15s"/>
    <clear begin="t2+5s"/>
    Now listen to the violin performance.
    <clear begin="t2+10s"/>
    ...
   </smilText>
  </par>
 </body>
</smil>
```

Figure 5.4. SMIL document model and temporal containers.




```
<smil ...>
  <head ...> ... </head>
  <body>
    <par>
      <seq>
        <video region="r1" id="v1" src="YouTube video"></video>
        <video region="r1" id="v2" src="Vimeo video"></video>
        <video region="r1" id="v3" src="Facebook video"></video>
      </seq>
      <smilText region="c1"...>
        ...
        <clear begin="t1-5s"/>
        That's an awesome guitar solo!
        <a href="Jump to t2">Go to next video</a>
        <a href="Skip 10s">Or skip 10s</a>
        <clear begin="t1+15s"/>
        ...
      </smilText>
    </par>
  </body>
</smil>
```

Figure 5.5. Timed text content and temporal hyperlinks.

5.2.3 Temporal Hyperlinks

SMIL 3.0 Linking Modules define SMIL 3.0 document attributes and elements for navigational hyperlinking. SMIL hyperlinks may be triggered by user interaction or other events, such as temporal events. SMIL 3.0 provides only inline link elements. Links are limited to unidirectional single-headed links (i.e. all links have exactly one source and one destination resource).

As with styled time-based text comments, adding temporal hyperlinks via text content can enrich the content viewing experience for end-users and for their social circle. This association makes SMIL meet requirement vii.

It is important to highlight that our document model allows links to be added to content without violating the legal rights of any party. This is possible because navigation points within the video are encoded as a series of content events in the SMIL document. Two classes of links can be provided as illustrated in Figure 5.5:

- *Intra-video Navigation Link:* a text link that takes the viewer to another location within the active video; and
- *Inter-video Navigation Link:* a text link that takes the viewer to another piece of content, outside the active video.

5.2.4 Contextual Information

Current Web-based video solutions provide limited support for including metadata related to the comments. For example, they do not allow end-users, at authoring time, to create different views on the comments, depending on the target audience. As discussed in the previous chapter, one might not send the same set of comments to her family and for to singing teacher.

SMIL 3.0 allows associating meta-information to any element within the document body, including timed text comments. This makes it possible to provide information with semantic intent within the presentation information, by binding relevant nodes with meta-information.

As mentioned before, SMILText allows text comments to be described as single structured units that can be targeted to different audiences. Therefore, we can consider each comment entry as the smallest unit that can be tagged. In order to share a video with comments, we should add contextual metadata, such as who has created the comment, when, why, how, and to whom [60]. Support for targeted

comments might increase the authoring overhead, but it provides a level of personalization that is lacking in common Web environments.

SMIL can tackle the contextual problem, requirement v, by allowing metadata to be associated with timed text comments. Figure 5.6 illustrates this process. Here we see a master comment stream that has been composed by Dick specifically targeted for all viewers within his social circle.

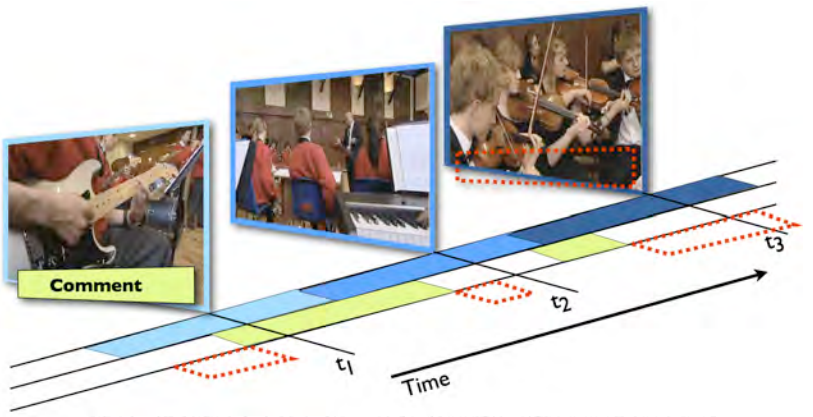
5.2.5 Selective Viewing

One shortcoming of current video captioning/commenting systems – whether closed captions or stream of comments on a Web page – is that every user visualizes the same collection of information. It is doubtful that even the most interested reader will go through the dozens of comments created by unknown individuals – but there is a much stronger incentive to view the 20 or so comments that are likely to be generated by family members or close personal friends (as indicated by the results presented in Section 5.1).

In order to deal with this problem, video commenting tools can make use of the structured nature of SMIL to selectively present content, requirement vi. Video commenting tools can enable users to — besides the traditional turn on/off all comments — select and watch comments created by a certain individual or community, about specific topics, or created on a certain day. Moreover, aggregated comments and metadata can be used for generating diagrams of hotspots within videos. All of this is possible thanks to the document model — structured text comments can be analyzed — and to the contextual information associated to the comments. Figure 5.6 illustrates a scenario in which a viewer is interested in a certain category of comments.

5.3 A Timed Text Video Commenting System

Based on the temporal transformations discussed in the previous section, we designed and implemented a video commenting system as an independent application. Our solution allows end-users to easily add timed text comments to particular events within third-party online videos. In the remaining of this section we detail the technical aspects of such commenting system.




```
<smil ...>
  <head ...> ... </head>
  <body>
    <par>
      <seq> ... </seq>
      <smilText region="c1"...>
        ...
        <div>
          <metadata name="author" content="Dick">
          <metadata name="audience" content="all">
          <clear begin="t1-5s"/>
          That's an awesome guitar solo!
          <clear begin="t1+15s"/>
        </div>
        <div>
          <metadata name="author" content="Dick">
          <metadata name="audience" content="his daughter">
          <clear begin="t3-15s"/>
          Do you remember you played the same song last year?
          <clear begin="t3+5s"/>
        </div>
        ...
      </smilText>
    </par>
  </body>
</smil>
```

Figure 5.6. Contextual information and selective viewing.

5.3.1 Infrastructure

The high-level workflow of our video commenting system is illustrated in Figure 5.7. The interaction starts when a user requests a video. For that we make use of the YouTube Data API (*Application Programming Interface*), which provides programmatic access to the videos stored in YouTube. It allows us to retrieve a set of videos matching a user-specified search term or retrieve standard feeds (e.g., most viewed today). The data is requested using AJAX (*Asynchronous JavaScript and XML*) and returned in the XML (*eXtensible Markup Language*) format, then parsed and presented to the end-user.

For video playback we use the YouTube Player API, which is exposed via JavaScript. It allows us to control not only the 'Look and Feel' of the player, but also the playback behavior of the videos in our Web application. With the current YouTube infrastructure, the client Web browser must be HTML5 compliant or have Flash Player 10.1, or higher, installed. Most importantly, the Player API provides the necessary time information to synchronize the text comments within a video. This feature is obtained by listening to specific events, which are fired accordingly (e.g., time update event). A similar infrastructure would be necessary for making the commenting tool available for videos hosted in different providers. In this case, the YouTube Data and Player APIs should be replaced and the interfaces of the new provider adapted accordingly.

Since the viewer has no rights to add comments to the base video, the timed comments are stored separately on our Web server. As mentioned previously, the actual format used to encapsulate the multimedia presentation (base video plus a layered collection of timed comments) is W3C's SMIL 3.0. In fact, timed comments are specified in SMILText, the embedded text format for use within SMIL 3.0. SMIL allows us to respect the video owner's rights and to keep a provider-agnostic enriched video. As such, comments can be shared, modified and analyzed independently.

For the synchronized playback of end-user comments we implemented a SMILText JavaScript engine that runs on the client's Web browser. Its API allows us to embed SMILText functionalities in Web pages and have the presentation controlled by an external source, in this case the YouTube video player. The SMILText engine has reasonably complete coverage of the features defined in the SMIL 3.0 SMILText External Profile. The API also provides a number of other utilities for adding and manipulating timed text content, making possible the creation of applications such as the commenting tool presented in this chapter.

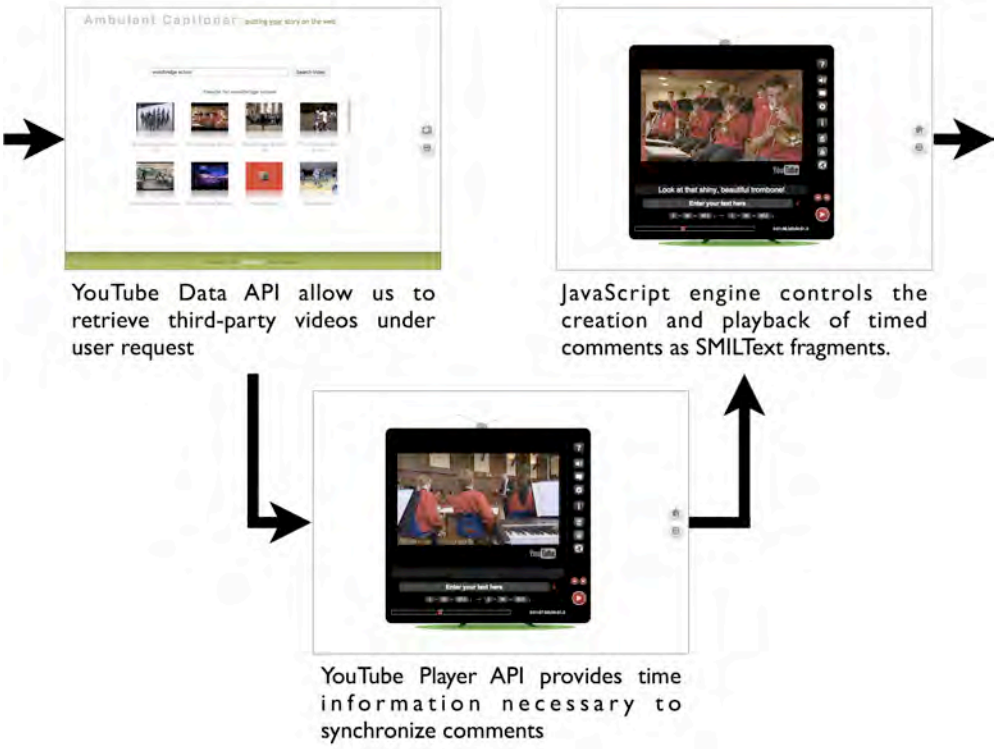


Figure 5.7. Workflow of our online video commenting system.

5.3.2 User Interface

In order to allow end-users to comment the videos we need a user interface that hides all the complexity from them. This is achieved with our video commenting system, which wraps the video content and all the timed text comments in a multimedia presentation. The commenting interface (Figure 5.8) is composed of a video rendering area (1), a rendering space for comments (2), an input area (3) and the sidebar controls (4). In most cases, relative passive end-users simply will watch a piece of video content forwarded to them. If the content itself has embedded comments, these can be selectively turned on or off via the sidebar controls

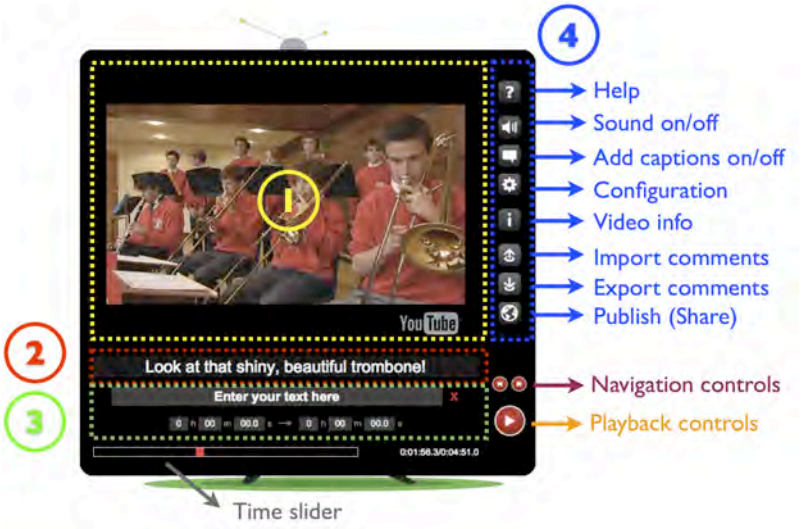
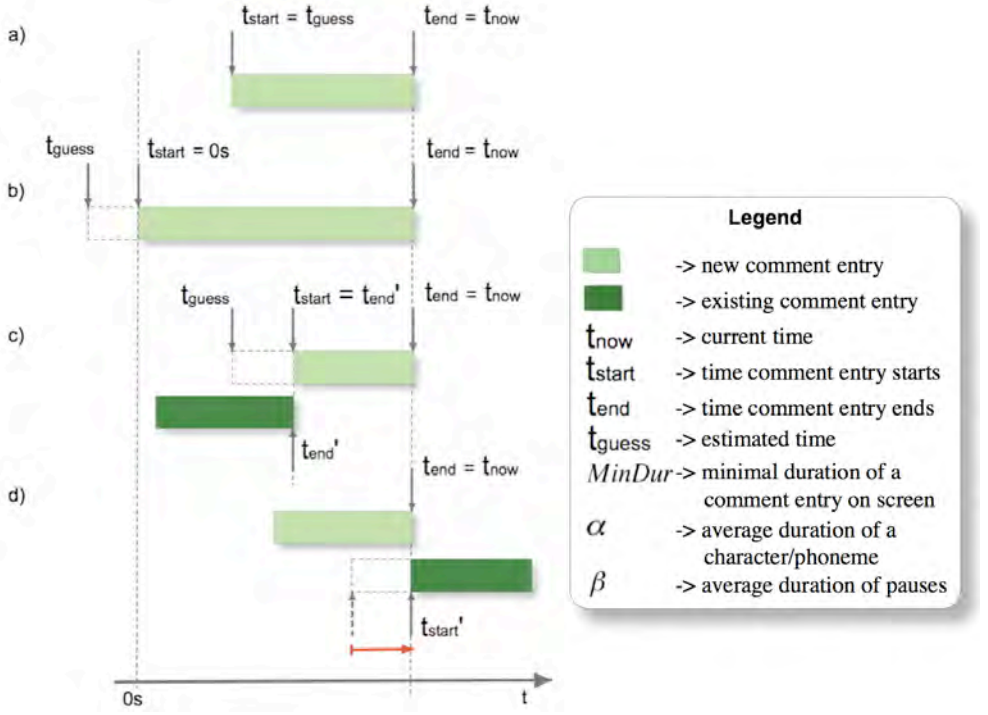


Figure 5.8. Video commenting interface.

interface. During playback users may also choose to insert new comments in the input area (Figure 5.8 (3)).

One key feature of our video commenting system is its ability to (semi-) automatically compute the temporal alignment of user-generated comments. To explain how this functionality works, consider the example in Figure 5.8. In a usage scenario, we assume users will interact after a certain moment of interest has passed (e.g., after seeing the trombone on screen). In this case, comments need to be synchronized in such a way as to avoid situations in which the comment – "Look at that shiny, beautiful trombone!" – appears after the instrument is not longer visible. Our approach for this use case is as follows. When an end-user indicates s/he wants to add a comment, the video playback is paused and the input area gains focus (Figure 5.8 (2)). As the interaction is performed right after listening to or watching an event of interest, we assume that the current moment (t_{now}) is the end of the comment ($t_{end} = t_{now}$). As an initial guess, we consider that the start time of the comment (t_{start}) is equal to the current time (t_{now}) minus a minimal duration (*MinDur*) that a comment should stay on screen for being effectively read ($t_{start} = t_{guess} = t_{now}$ - *MinDur*).

Based on our prediction model and its parameters – e.g., the number of words in a comment (*N*), the average duration of a character/phoneme of a word in a specific language (α) and the average duration of pauses (β) – t_{guess} is recalculated, being t_{start} then determined by the maximum value among t_{guess}, the end



$$t_{start} = \max\{0;\ t_{guess};\ t_{end}'\}$$

$$t_{end} = t_{now}$$

$$t_{guess} = t_{now} - \max\left\{\begin{array}{l} MinDur; \\ \left(\sum_{i=1}^{N} length(i) * \alpha\right) + (N-1) * \beta \end{array}\right\}$$

Figure 5.9. Predictive timing support for end-user comments.

of the previous existing comment (t_{end}') and zero. Figure 5.9 illustrates scenarios in which t_{start} assumes different values. In the example of Figure 5.8, the start and end time computed for "Look at that shiny, beautiful trombone!" stayed around 3min57s. When the user saves the comment, the video playback is resumed. The predictive timing functionality often provides coarse temporal support; users may fine-tune the timing if desired. In our experience, such fine-tuning is not necessary unless tightly coupled subtitles are being created.

5.4 Evaluation

As mentioned before, the survey discussed in Section 5.1 was followed by 2 other experiments. First, participants were instructed to interact with the prototype system presented in Section 5.3 and then, fill in a questionnaire to report their experiences. In the second and last part, they were asked to further explore the commenting activity by close captioning a sample video (approx. 7 minutes duration) and fill in another questionnaire. Table 5.1 summarizes the number of participants involved in each part of the evaluation process presented in this chapter. In the next sections we present the results and discuss the findings from the evaluation of our online video commenting tool.

Table 5.1. Composition of participants in evaluation process.

Evaluation Questionnaire	*Number of Respondents*	*% of Respondents*
1. Current video watching and commenting practices	21	100.0%
2. Commenting on videos with our prototype system	18	85.7%
3. Captioning videos using our prototype system	12	57.1%

5.4.1 Commenting on Videos

In general, participants' feedback regarding our video commenting tool was very positive (see Figure 5.10). When asked how much they liked the service (Q2.1), 13 out of 18 answered *some* or *a lot*. All respondents reported that our video commenting tool is helpful for adding synchronized comments to YouTube videos (Q2.2). Some expressed such appreciation by saying that *"synchronization is much better"* and *"I can easily add comments to specific moments in the video. In Facebook I think I can't. In YouTube I can but I have to type the time moment in the comment"*.

When compared to regular comment threads in YouTube or Facebook, 9 users said our tool is *better* or *much better* (Q2.3). A user justified his/her answer by saying that *"the possibility to comment on a specific moment in the video adds a lot of functionality. Instead of saying, 'after 16 seconds he does this', you can just comment at that moment. This also works quite well on SoundCloud as far as I have seen"*. On the other hand, 5 participants said they were unable to judge. One of them explained: *"I have never added comments to Facebook nor YouTube. However, the way to add comments in this (video commenting) tool is intuitive"*.

5.4.2 Close Captioning Videos

The last experiment was the most time consuming one, and for this reason, only 12 participants committed to complete it. Users were kindly requested to close caption a 7 minutes speech video. This task was first performed using our video commenting tool, and later using a standard video player and a text editor. This experiment allowed us to evaluate the effectiveness and usefulness of the time prediction algorithm provided in our commenting tool.

Using our tool, participants spent in average 61 minutes (Standard Deviation: SD = 48 minutes), and other 101 minutes (SD = 33 minutes) without it. The utility of our commenting system has also been reflected in the answers to the questionnaire (see Figure 5.11). When asked how much easier it was to add close captions with our system compared to the other method, all respondents said it was *much easier* or *easier* (Q3.1). A similar feedback has been obtained in the question regarding participants' appreciation for the predictive synchronization of captions/comments (Q3.2). In this case, 7 users reported to have liked *a lot*. In one case, one participant mentioned that *"most captions were synchronized nice to the video, and the prediction algorithm does work. It saves a lot of time having not to*

fine tune the start and end points, as you have to do with the SRT format". And another user added: *"the prediction works really good, the captions are usually where they are supposed to be!"*.

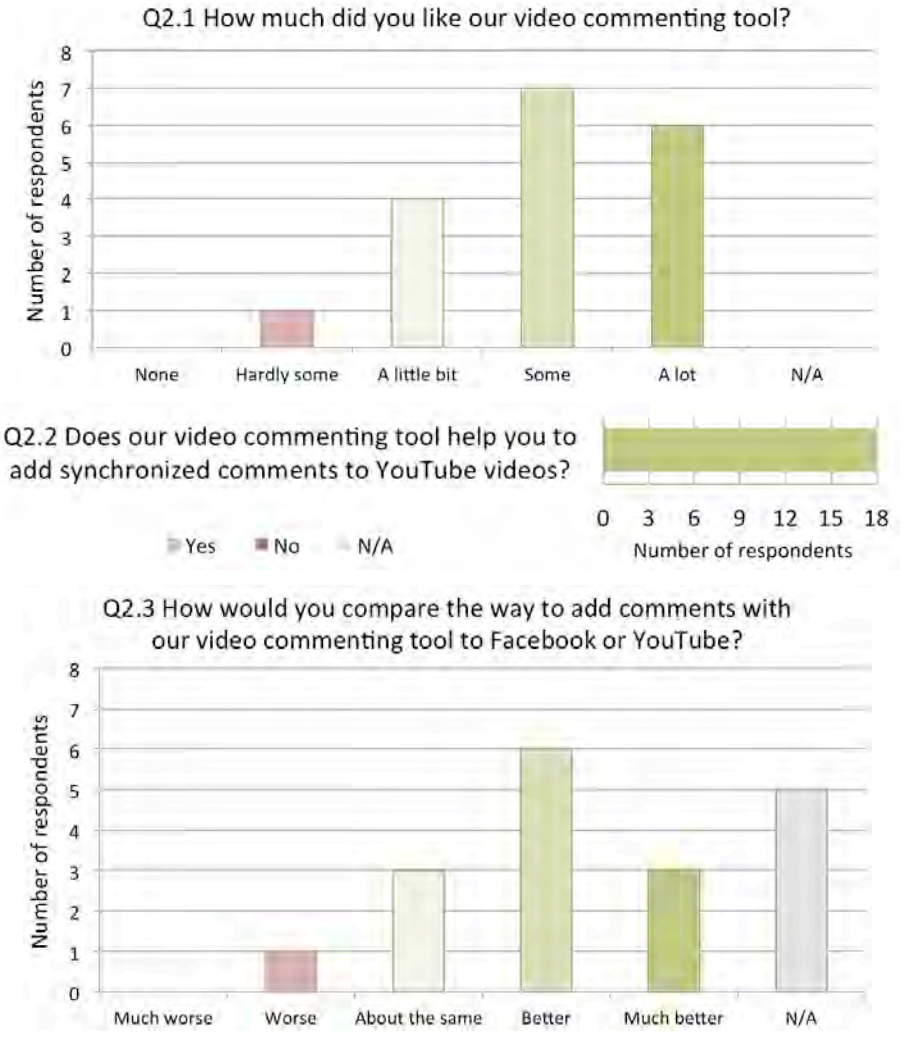


Figure 5.10. Results from the evaluation of our commenting tool.

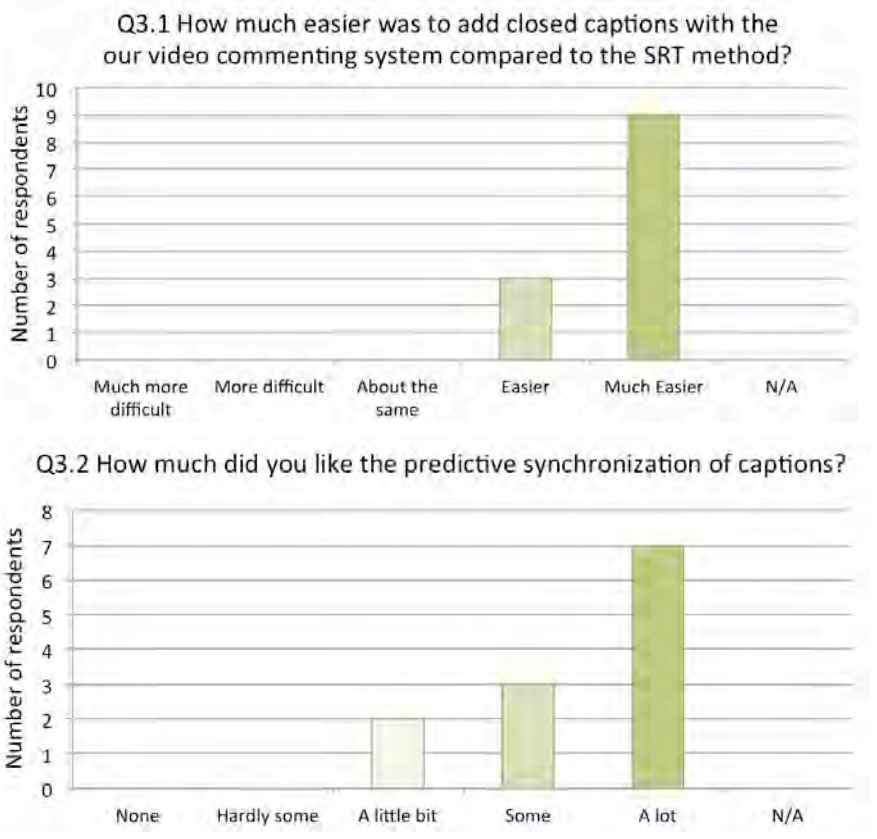


Figure 5.11. Results from the close captioning activity using our prototype system.

Although our primary objective was not to provide a close caption authoring tool, the point we want to make here is that video commenting systems like ours should not only allow users to add timed comments, but also help them by offering automatic processes that make the commenting task simpler and more intuitive.

5.5 Discussion

In this chapter we presented our efforts in supporting personalized content enrichment. Motivated by a survey on media watching and commenting practices, we introduced and evaluated a video commenting paradigm that follows the *intimacy* and *reciprocity* guidelines introduced in Chapter 2. Results from the evaluation process show that that users appreciated the functionalities of our system and would potentially use it to communicate with their close circle (requirements on intimacy and reciprocity) and, also, with the general public.

The survey research about media watching and commenting practices represents the first major contribution of this chapter. While this study is relevant to analyze user behavior; it is even more to motivate our work. Do people want to add timed comments within videos? Our results provide evidences that regular Internet users would add synchronized comments while consuming video on demand if they had the appropriate tools for doing that.

From a document model perspective, all the requirements presented in Section 5.1 are met by using a structured multimedia language like SMIL. In this work we focused on text, but a similar approach could be used for other types of user-generated enrichments [56][61]. The video commenting tool reported in this chapter also addresses the functional requirements. The transformation process starts when a video URL is given as an input. Next, our video commenting system applies a document model transformation, which respects the owners' rights by retaining the video integrity (requirement i) and allow compilations that include video clips from multiple sources (requirement iii). Timed text content is applied as soon a user clicks the input area (requirement iv). This means that given a multimedia document, our tool adds a parallel container that synchronizes comments with a particular video. Whenever a new comment entry is inserted, implicit metadata is automatically added (requirement v). As these comments can be targeted to different audiences, they can be selectively rendered (requirement vi). Multiple-video aggregation and timed end-user navigation (requirements ii and vii, respectively) can be met by integrating the personalized narratives presented in the previous chapter.

The evaluation of our video commenting system represents the second major contribution of this chapter. It shows that this paradigm brings a measurable increment over existing commenting systems. It also shows that the burden of synchronizing comments can be minimized by the use of predictive timing. These results answer our research question. Finally, we do not claim synchronized

comments should replace traditional ones, but rather be complementary. Regular comments are targeted to a fundamentally different use case than the ones offered by our system. On the one hand, in Facebook or YouTube, people can comment about a video, but also give feedback to the author or start a conversation about something unrelated. On the other hand, our video commenting system can be used to highlight interesting things for other viewers, maybe to make a point about a particular event within the video. Such textual comments should be preferably simple; otherwise viewers will have problems to read while watching a video.

6

Conclusions[1]

During the past 20 years, authoring has been part of the multimedia community's research agenda. Unfortunately, multimedia authoring has been seen as an initial enterprise that occurs before 'real' content processing takes place. This limits the options open to authors and to viewers of rich multimedia content in creating and receiving focused, highly personal media presentations. This thesis reflects on the multimedia authoring workflow and we argue that a fresh new look is required. We focused on the particular task of supporting *socially-aware multimedia authoring*, in which the relationships within particular social groups among authors and viewers can be exploited to create highly personal media experiences. Our framework is centered on empowering users in telling stories and commenting on personal media artifacts, considering the long-term social context of the user's social environment. We provided an overview of the requirements and characteristics of socially-aware multimedia authoring within the context of exploiting community content. In particular, our research involved the study of different mechanisms to allow users to explore, create, enrich and share videos based on personal relationships. Our methodology integrated knowledge from Human-Computer Interaction (e.g., focus groups/interviews for need assessment, iterative prototyping and user evaluation) and document engineering.

[1] This chapter contains extracts from the following article:

D.C.A. Bulterman, P. Cesar and R.L. Guimarães. 2013. Socially-Aware Multimedia Authoring: Past, Present and Future. ACM Transactions on Multimedia Computing, Communications and Applications (TOMCCAP), Volume 9, Issue 1s, Article 35 (October 2013), 23 pages. DOI=10.1145/2491893 http://doi.acm.org/10.1145/2491893

In this chapter we first revisit and answer the research questions of the thesis. We then reflect on the lessons learned, before concluding with a discussion of the issues that we feel can provide a fruitful basis for future multimedia authoring support. We argue that providing support for socially-aware multimedia authoring can have a profound impact on the nature and architecture of the entire multimedia information processing pipeline.

6.1 Revisiting the Research Questions

Much of the media landscape has been, and continues to be, dominated by commercially produced content. Whether image, video, audio or (to a lessor extent) text, users today have become accustomed to experiencing highly polished media messages. In spite of the dramatic impact of user contributed content sites (such as YouTube and Facebook), the amount of personal content being shared with family and friends (to say nothing of wide anonymous audiences) is minimal. A conservative estimate of media use indicates that average owners of smartphones and portable cameras capture hours of videos yearly, but that only minutes (or seconds) of content are being shared. Does this mean that user-generated content is less important? No. Personal archives have a high degree of personal value: photos of family and friends, videos of small children, audio fragments that capture the sounds of people who have played an important role in one's life. Although there may always be exceptions, it is clear, that a short video showing a child's first violin solo will not attract the same audience as, say, a slickly-produced commercial music video. This does not make the violin fragment less valuable.

In this thesis, we focused on community authoring applications, where content is contributed from many amateur sources and distributed within a relatively closed circle of viewers who have varying degrees of affinity with the content produced. We concentrated on support for situations in which both the original presentation creator and the presentation viewer play a role in determining presentation content. Given this context, we discuss and answer each of the research questions according to the work presented in the bulk of the thesis.

Question 1.1 *Can a socially-aware multimedia authoring system be defined in terms of existing social science theories and human-centered processes, and if so, which?*

In this thesis we reformulated the research problem of multimedia authoring, by investigating mechanisms and principles for togetherness and social connectivity around personal media. Our focus was on parents, family members and friends of students participating in a small-scale social event. In this scenario, parents capture recordings of their children for later viewing and possible sharing with friends and relatives. Based on a 4-year evaluation process, we specified a set of guidelines for the design and implementation of socially-aware multimedia authoring and sharing tools. We aim at nurturing strong ties and improving social connectedness by supporting *emotional intensity*, *personal effort*, and by supporting *intimacy* and enabling *reciprocity*. With these guidelines we intend to increase the feeling of connectedness, particularly among family members and friends who could not attend the social event. As shown in Chapter 2, our socially-aware multimedia authoring paradigm is aligned with the requirements needed for social communities that are not addressed by existing social media Web applications. These guidelines directly address research *Question 1.1*.

Question 1.2 *Does the functionality provided by a socially-aware multimedia authoring system provide an identifiable improvement over traditional authoring and sharing solutions? If so, how can these improvements be validated?*

To evaluate the utility and usefulness of socially-aware multimedia authoring, we realized the guidelines mentioned above in a two-phased prototype system called MyVideos. We have actively participated in the design, implementation and integration of this system, and our contributions enabled us to perform extensive field trials and these were a major part of the TA2's success[2]. Working with a test group at local high schools in two different countries (UK and the Netherlands), we investigated how focused content can be extracted from a shared repository, and how content can be enhanced and tailored to form the basis of a personalized multimedia artifact, that can be eventually transferred and shared with family and friends (each with different degrees of connectedness with the performer and his/her parents). Results from a long-term evaluation process show that all our participants (from phase 1 and 2) liked the functionality provided by our system and considered this a valid alternative to strength social interactions when apart. Therefore, using our system they would share more videos with friends

[2] The pan-European Project Together Anywhere, Together Anytime – http://ta2-project.eu.

and family. These results – complemented by more specific findings on media exploration, creation of personal memories and content enrichment (Chapters 3-5) – directly answer research *Question 1.2*.

Question 1.3 *Does a socially-aware video exploration system provide an identifiable improvement over current approaches for accessing and navigating a repository of shared media?*

While following the *emotional intensity* guideline, in Chapter 3 we discussed a two-phased design, development and experimentation of an interface for browsing a collection of user-generated videos from a shared event. Users could explore and navigate (fragments of) video clips recorded by several people based on their own personal/social interests. The design, deployment and evaluation of the system resulted in the identification of key requirements for this novel type of browsing interfaces. In particular, our approach 1) supports exploration based on the inherent event structure; 2) it makes use of contextual information to help in the navigation process; 3) it allows for flexible searches based on a combination of filters; and finally, 4) it provides a way to switch between cameras angles that might have captured different aspects of the event. Results of the evaluation process show that all participants appreciated the browsing interface and indicated that it is better than traditional tools to explore videos they care about. Therefore, they would find videos more efficiently using our system. These results clearly indicate that a socially-aware video exploration system like ours provides an improvement over current tools for accessing and navigating a repository of shared media assets, directly answering research *Question 1.3*.

Question 1.4 *Where is the balance between automatic and manual processes when authoring personalized narratives users care about?*

As for browsing a shared video collection, social relationships are key for authoring personalized stories users care about. In Chapter 4 we reported on our efforts to support the creation of personalized video stories reusing collective content. We developed a first version of an authoring system, subjected it to user testing, and then developed an improved version that follows the *personal effort* guideline of socially-aware multimedia authoring. Our initial results showed a general enthusiasm from participants, which were validated in the first evaluation phase. While the video compilations automatically produced by the initial system

were considered visually compelling, users missed the capability of personalizing those by adding their own 'imprint'. To address this limitation, we proposed a hybrid authoring approach that provides mixed support for automated creation by selecting content of personal interest and manual enhancement of personalized video stories. Based on user feedback as part of our four-year study, we have demonstrated that it is possible to satisfy casual content creators while still allowing extensive personalization to take place, if needed. These results directly answer research *Question 1.4*. We believe that the combination of automatic and manual processes provides the balance of complexity and functionality.

Question 1.5 *Does the support for timed end-user commenting within pre-authored narratives provide an identifiable improvement over current media commenting approaches?*

While concentrating in the creation process, we cannot forget that content enrichment also plays an important role in the socially-aware multimedia authoring workflow. Motivated by a survey on media watching and commenting practices, in Chapter 5 we reported on the design, implementation and user-centric evaluation of a video commenting framework that follows the *intimacy* and *reciprocity* guidelines. To realize such framework, we specified and described a set of temporal transformations for multimedia documents. Our approach allows end-users to create and share personalized timed text comments within third-party online videos. The benefit over current solutions lays in the usage of a rich commenting format – in our case SMIL [17] – which is not embedded into a specific video encoding format. The evaluation of a video commenting system that realizes our framework clearly indicates that participants appreciated our system (13/18 or 72% of the participants), and considered it helpful (100%). Our results also show that 50% of the participants considered our video commenting approach better than the one offered in YouTube and/or Facebook. These results show that our commenting framework brings a measurable increment over existing commenting systems, and directly answer research *Question 1.5*.

6.2 Reflection and Further Directions

In this thesis we provide useful insights into how a socially-aware multimedia authoring and sharing system should be designed and architected, for helping users

in recalling personal memories and in nurturing their close circle relationships. The main contribution of our work does not lie in the use of a specific technology (e.g., SMIL, NSL or Web standards) but in further understanding the fundamental trade-offs that enable better sharing of 'personal' media. Results from our evaluation process show that socially-aware multimedia authoring provides a more fruitful approach than earlier work.

Although our research has reached its aims, there are some unavoidable limitations. First, the total amount of time spent to annotate the footage (see in Chapter 2) demonstrates that this is still a very challenging problem, especially when we consider dimly lit user-generated content with different quality, encoding etc. Although these annotations are essential in our authoring framework, this thesis does not aim at solving this problem.

As to the number of subjects participating in the evaluation, we agree that 'more is more', but note: each subject needed to agree to spend some hours per evaluation (about 1h30min recording concerts plus 2h in lab studies). We found it difficult to find high school parents who would commit to this load. We are pleased that our parents – about 25% of the concert participants! – were motivated contribute this block of time. The goals of the study make it impossible to do crowdsource testing, given the focus on common personal media. Moreover, we are not aware of other studies that provide the same breadth.

Another limitation could be that we focused on a particular use case scenario. We reiterate that our participants represent a realistic sample of users: actual family members from 2 countries (NL and UK) that have been involved in the concert recordings and prototype evaluation. We agree that generalization to other events is an important problem, but before getting there we need to start somewhere. We see this as a topic for future work.

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

6.2.1 Media Encoding and Storage

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

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

Having a logical media object be constructed out of dynamically combined physical fragments allows customized navigation to be supported. In YouTube (as in other commercial video sharing systems), dynamic mashups are not supported. End-users have to find suitable source material, cut it into shots, and assemble an encoded final video. While this solution does not impose hard requirements on delivery and rendering, it is limited in terms of adaptability, user interaction and seamless playback [36].

One approach to implementing such dynamic combination is supported by DASH (*Dynamic Adaptive Streaming over HTTP*), a system for HTTP-based streaming [70]. Although some efforts have investigated the use of DASH with Rich Media services [9][10], at present, it is typically used for storing pre-defined fragment encodings, nearly always based on support bitrate-adaptive resolutions (During presentation, the quality of the content can be adjusted based on environmental factors such as available bandwidth or end-user screen size). Similarly, dynamic media compositions could be achieved using a combination of HTML5 and W3C Media Fragments [21] and/or JavaScript code (e.g., Popcorn.js or Kaltura Video Platform).

Adaptivity in our work can leverage this support, but our main interests are in supporting a more abstract form of content selection: providing more trombone content to the father of the trombone player and more clarinet content to the mother of the clarinetist. This is a matter of dynamic content selection rather than (or at least in addition to) dynamic encoding selection. The selection (or generation) of

dynamic content requires more illusive criteria for content selection, such as a profile of the viewer in addition to profiles of the available content, and a content-wide temporal model that exposes logical divergence and convergence points for creating content streams. It also requires a container format that allows differential segment length to exist across candidate segments. To support this, the current model of content streaming would need to be revised: the seamless integration of individual content fragments (as opposed to encoding fragments) into a logical whole is a composition concept that most media servers and media container languages are as-of-yet ill-equipped to support.

6.2.2 Media Classification and Annotation

Personal media classification and annotation remains a challenge for supporting effective content sharing. For professional content, content is often highly segmented along the lines of established commercial distribution models. For personal content, the situation is vastly different. This shift in emphasis is new for multimedia, but there are many established examples in music, art and literature where the intentions of the composer, artist or writer are decoupled from the applications of the media itself.

At present, personal content annotation is driven by device-supplied metadata (e.g., clocks, location coordinates, file names, as well as objects and faces). For socially-aware multimedia, it is also necessary to encode relative social relationships among interested parties – plus to maintain those relationships over time. As with any large software system, the long-term maintenance costs of media will dominate the short-term development costs. This will require a new generation of iterative, socially-aware media classification tools. The analysis of content becomes then a continuing task, not an import activity. In the same vein, content recommendation needs to not only use such information, but also be sensitive to the context of use: are you watching alone, with your spouse, with your children, with your friends?

6.2.3 Customized Media Selection

Perhaps the most significant innovation in (broadcast) content selection occurred with the introduction of the video tape recorder. For the first time in history, it was the viewer that determined when content would be watched – on the precondition that it had been broadcast and recorded earlier. A next, but more minor, innovation

came with the introduction of the digital set-top box, which included an embedded program guide, providing the opportunity for more automated content selection and recording. The next logical development is to remove the TV guide altogether and to have the system itself recommend content for the family, which it found based on metadata encoded by the content providers.

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

6.2.4 Content-Based Navigation

One of the challenges with temporal searching along a timeline is that it is a highly unstructured activity. For instance, in a conventional YouTube interface for navigating through a video object, users can only select key frames without any higher-level narrative guidance. We note that even 1980's generation DVD technology provided more significant control through its chaptering interface. In general, the time axis provides no information on the logical structuring of the event, letting alone the performers in the concert or their relationships. Still, in the absence of any semantic structuring of content, it is often all that is available.

It will be necessary to study new mechanisms to replace timeline searching with navigation based on an overlay of structure components. One approach to provide this structure in our school concert use case is based on graphs of performers, instruments, songs or solo's. It could also be based on cinematographic classifications, such as long shots, pan shots, tight shots.

6.2.5 Ownership and Digital Rights

Reusing content brings with it questions of ownership. In printed documents, this is a solved problem: even though the base content is copyright protected, there is a

clear distinction between 'my' media and that of the original authors. For web pages and online content, the relationship is less simple.

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

When annotating a piece of media – whether it is text, audio, images, or whatever – the implication has been that the annotations are of a highly personal nature. Of course, if many of these personal notes are collected and analyzed, they could provide valuable insights into the reusability of personal media assets. Even a simple density analysis of multiple media annotations could provide interesting clues for socially-aware recommender systems.

6.2.6 Security and Privacy Concerns

Content can be used or misused by various members of a user community, depending on their (possibly time-variant) relationships. Research is required to support content access and content protection that reflect time-variant social and personal relationships.

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

6.3 Closing Thoughts

Much has changed in the 'world' of multimedia. Who would have expected twenty years ago that within two decades, it would be commonplace to not only listen to music via your computer, but buy it there as well? That books would not only be written on a PC, but that the PC and its technological 'cousins' would become a handy way to read them, or to have them read aloud. That the computer would

threaten to replace not only the television, but also the movie theatre as a venue for the shared watching of content. And, perhaps more significantly in the long term, that the computer would not only render a wide range of real and artificial images, but that it would attempt to understand them as well.

In this thesis we have outlined what we mean by *socially-aware multimedia*. We have argued that the impact of supporting user-in-the-small transcends the incremental and provides a number of (fascinating) new challenges that require fundamental research results across a wide range of multimedia disciplines.

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

Bibliography

[1] 2002. MPEG-7: The Generic Multimedia Content Description Standard, Part 1. IEEE MultiMedia 9, 2 (April 2002), 78-87. DOI=10.1109/93.998074 http://dx.doi.org/10.1109/93.998074

[2] A. Eliëns, H.C. Huurdeman, M.R. van de Watering and W. Bhikharie. 2008. XIMPEL Interactive Video - between narrative(s) and game play. In Proceedings of GAMEON, 132-136.

[3] A. Hanjalic and L. Xu. 2005. Affective video content representation and modeling. IEEE Transactions on Multimedia. 7, 1 (February 2005), 143-154. DOI= http://doi.ieeecomputersociety.org/10.1109/TMM.2004.840618

[4] A. Macaranas, G. Venolia, K. Inkpen, and J. Tang. 2013. Sharing Experiences over Video: Watching Video Programs Together at a Distance. In Proceedings of the 14th IFIP TC13 Conference on Human-Computer Interaction (INTERACT '13). Springer-Verlag Berlin Heidelberg.

[5] A. Piacenza, F. Guerrini, N. Adami, R. Leonardi, J. Porteous, J. Teutenberg, and M. Cavazza. 2011. Generating story variants with constrained video recombination. In Proceedings of the 19th ACM international conference on Multimedia (MM '11). ACM, New York, NY, USA, 223-232. DOI=10.1145/2072298.2072329 http://doi.acm.org/10.1145/2072298.2072329

[6] B. Adams, S. Venkatesh, and R. Jain. 2005. IMCE: Integrated media creation environment. ACM Trans. Multimedia Comput. Commun. Appl. 1, 3 (August 2005), 211-247. DOI=10.1145/1083314.1083315 http://doi.acm.org/10.1145/1083314.1083315

[7] B.T. Truong and S. Venkatesh. 2007. Video abstraction: A systematic review and classification. ACM Trans. Multimedia Comput. Commun. Appl. 3, 1, Article 3 (February 2007). DOI=10.1145/1198302.1198305 http://doi.acm.org/10.1145/1198302.1198305

[8] C.A. Lee, H.R. Tibbo, D. Howard, Y. Song, T. Russell, and P. Jones. 2006. Keeping the context: an investigation in preserving collections of digital video. In Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries (JCDL '06). ACM, New York, NY, USA, 363-363. DOI=10.1145/1141753.1141858 http://doi.acm.org/10.1145/1141753.1141858

[9] C. Concolato and J. Le Feuvre. 2013. Live HTTP streaming of video and subtitles within a browser. In Proceedings of the 4th ACM Multimedia Systems Conference (MMSys '13). ACM, New York, NY, USA, 146-150. DOI=10.1145/2483977.2483997 http://doi.acm.org/10.1145/2483977.2483997

[10] C. Concolato, J. Le Feuvre, and R. Bouqueau. 2011. Usages of DASH for rich media services. In Proceedings of the second annual ACM conference on Multimedia systems (MMSys '11). ACM, New York, NY, USA, 265-270. DOI=10.1145/1943552.1943587 http://doi.acm.org/10.1145/1943552.1943587

[11] C.G.M. Snoek, B. Freiburg, J. Oomen, and R. Ordelman. Crowdsourcing rock n' roll multimedia retrieval. In Proceedings of the international conference on Multimedia (MM '10). ACM, New York, NY, USA, 1535-1538. DOI=10.1145/1873951.1874278 http://doi.acm.org/10.1145/1873951.1874278

[12] D.A. Shamma, L. Kennedy, and E.F. Churchill. 2012. Watching and talking: media content as social nexus. In Proceedings of the 2nd ACM International Conference on Multimedia Retrieval (ICMR '12). ACM, New York, NY, USA, Article 12, 8 pages. DOI=10.1145/2324796.2324811 http://doi.acm.org/10.1145/2324796.2324811

[13] D.A. Shamma, L. Kennedy, and E.F. Churchill. 2009. Tweet the debates: understanding community annotation of uncollected sources. In Proceedings of the first SIGMM workshop on Social media (WSM '09). ACM, New York, NY, USA, 3-10. DOI=10.1145/1631144.1631148 http://doi.acm.org/10.1145/1631144.1631148

[14] D.A. Shamma, R. Shaw, P.L. Shafton, and Y. Liu. 2007. Watch what I watch: using community activity to understand content. In Proceedings of the international workshop on Workshop on multimedia information retrieval (MIR '07). ACM, New York, NY, USA, 275-284.

DOI=10.1145/1290082.1290120
http://doi.acm.org/10.1145/1290082.1290120

[15] D.C.A. Bulterman, A.J. Jansen, P. Cesar, and S. Cruz-Lara. 2007. An efficient, streamable text format for multimedia captions and subtitles. In Proceedings of the 2007 ACM symposium on Document engineering (DocEng '07). ACM, New York, NY, USA, 101-110. DOI=10.1145/1284420.1284451
http://doi.acm.org/10.1145/1284420.1284451

[16] D.C.A. Bulterman and L. Hardman. 2005. Structured multimedia authoring. ACM Trans. Multimedia Comput. Commun. Appl. 1, 1 (February 2005), 89-109. DOI=10.1145/1047936.1047943
http://doi.acm.org/10.1145/1047936.104794

[17] D.C.A. Bulterman and L. Rutledge. 2009. SMIL 3.0: Flexible Multimedia for Web, Mobile Devices and Daisy Talking Books. Springer-Verlag Berlin Heidelberg, 2nd ed., XXVIII, 508 pages, ISBN 978-3-540-78546-0

[18] D.J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. 2009. Mapping the world's photos. In Proceedings of the 18th international conference on World wide web (WWW '09). ACM, New York, NY, USA, 761-770.DOI=10.1145/1526709.1526812

[19] D. Kirk, A. Sellen, R. Harper, and K. Wood. 2007. Understanding videowork. In Proceedings of the SIGCHI conference on Human factors in computing systems (CHI '07). ACM, New York, NY, USA, 61-70. DOI=10.1145/1240624.1240634
http://doi.acm.org/10.1145/1240624.1240634

[20] D. Korchagin, P.N. Garner and J. Dines. 2010. Automatic Temporal Alignment of AV Data with Confidence Estimation. In Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 269-272. DOI=10.1109/ICASSP.2010.5495953
http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5495953&isnumber=5494886

[21] D. Van Deursen, R. Troncy, E. Mannens, S. Pfeiffer, Y. Lafon, and R. Van de Walle. 2010. Implementing the media fragments URI specification. In Proceedings of the 19th international conference on World wide web (WWW '10). ACM, New York, NY, USA, 1361-1364.

DOI=10.1145/1772690.1772931
http://doi.acm.org/10.1145/1772690.1772931

[22] D. Williams, M.F. Ursu, P. Cesar, K. Bergström, I. Kegel, and J. Meenowa. 2009. An emergent role for TV in social communication. In Proceedings of the seventh european conference on European interactive television conference (EuroITV '09). ACM, New York, NY, USA, 19-28. DOI=10.1145/1542084.1542088
http://doi.acm.org/10.1145/1542084.1542088

[23] E. Durkheim. 1971. The elementary forms of the religious life. Allen and Unwin. ISBN: 0042000033.

[24] E. Gilbert and K. Karahalios. 2009. Predicting tie strength with social media. In Proceedings of the 27th international conference on Human factors in computing systems (CHI '09). ACM, New York, NY, USA, 211-220. DOI=10.1145/1518701.1518736
http://doi.acm.org/10.1145/1518701.1518736

[25] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida and K. Ross. 2009. Video interactions in online video social networks. ACM Trans. Multimedia Comput. Commun. Appl. 5, 4, Article 30 (November 2009), 25 pages. DOI=10.1145/1596990.1596994
http://doi.acm.org/10.1145/1596990.1596994

[26] F. Shipman, A. Girgensohn, and L. Wilcox. 2008. Authoring, viewing, and generating hypervideo: An overview of Hyper-Hitchcock. ACM Trans. Multimedia Comput. Commun. Appl. 5, 2, Article 15 (November 2008), 19 pages. DOI=10.1145/1413862.1413868
http://doi.acm.org/10.1145/1413862.1413868

[27] G. Adams. 2006. Timed Text (TT) Authoring Format 1.0 – Distribution Format Exchange Profile (DFXP), W3C. Available at http://www.w3.org/TR/2006/CR-ttaf1-dfxp-20061116/. Last access on May 15th 2013.

[28] G.D. Abowd, M. Gauger, and A. Lachenmann. 2003. The Family Video Archive: an annotation and browsing environment for home movies. In Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval (MIR '03). ACM, New York, NY, USA, 1-8. DOI=10.1145/973264.973266 http://doi.acm.org/10.1145/973264.973266

[29] G. Fischer. 2001. User Modeling in Human-Computer Interaction. User Modeling and User-Adapted Interaction 11, 1-2 (March 2001), 65-86. DOI=10.1023/A:1011145532042 http://dx.doi.org/10.1023/A:1011145532042

[30] G. Rizzo, T. Steiner, R. Troncy, R. Verborgh, J.L.R. García, and R. Van de Walle. 2012. What fresh media are you looking for?: retrieving media items from multiple social networks. In Proceedings of the 2012 international workshop on Socially-aware multimedia (SAM '12). ACM, New York, NY, USA,15-20.DOI=10.1145/2390876.2390882 http://doi.acm.org/10.1145/2390876.2390882

[31] H. Becker, D. Iter, M. Naaman, and L. Gravano. 2012. Identifying content for planned events across social media sites. In Proceedings of the fifth ACM international conference on Web search and data mining (WSDM '12). ACM, New York, NY, USA, 533-542. DOI=10.1145/2124295.2124360 http://doi.acm.org/10.1145/2124295.2124360

[32] H. Sundaram and S. Chang. 2000. Determining computable scenes in films and their structures using audio-visual memory models. In Proceedings of the eighth ACM international conference on Multimedia (MULTIMEDIA '00). ACM, New York, NY, USA, 95-104. DOI=10.1145/354384.354440 http://doi.acm.org/10.1145/354384.35444

[33] ITU-T Rec. H.761, Nested Context Language (NCL) and Ginga-NCL for IPTV Services, Geneva, Apr. 2009. Available at http://www.itu.int/rec/T-REC-H.761. Last access on May 15th 2013.

[34] J.D. Weisz, S. Kiesler, H. Zhang, Y. Ren, R.E. Kraut, and J.A. Konstan. 2007. Watching together: integrating text chat with video. In Proceedings of the SIGCHI conference on Human factors in computing systems (CHI '07). ACM, New York, NY, USA, 877-886. DOI=10.1145/1240624.1240756 http://doi.acm.org/10.1145/1240624.1240756

[35] J. Ibáñez, R. Aylett, C. Delgado-Mata, and B. Molinuevo. 2008. On the implications of the virtual storyteller's point of view. Knowl. Eng. Rev. 23, 4 (December 2008), 339-367. DOI=10.1017/S0269888908000039 http://dx.doi.org/10.1017/S0269888908000039

[36] J. Jansen, P. Cesar, R.L. Guimarães, and D.C.A. Bulterman. 2012. Just-in-time personalized video presentations. In Proceedings of the 2012 ACM symposium on Document engineering (DocEng '12). ACM, New York, NY,

USA, 59-68. DOI=10.1145/2361354.2361368 http://doi.acm.org/10.1145/2361354.2361368

[37] K. Inkpen, H. Du, A. Roseway, A. Hoff, and P. Johns. 2012. Video kids: augmenting close friendships with asynchronous video conversations in videopal. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12). ACM, New York, NY, USA, 2387-2396. DOI=10.1145/2207676.2208400 http://doi.acm.org/10.1145/2207676.2208400

[38] K. Purcell. 2010. The State of Online Video, The Pew Research Center's Internet & American Life Project, June 2010. Available at http://www.pewinternet.org/Reports/2010/State-of-Online-Video.aspx

[39] K. Su, M. Naaman, A. Gurjar, M. Patel, and D.P.W. Ellis. 2012. Making a scene: alignment of complete sets of clips based on pairwise audio match. In Proceedings of the 2nd ACM International Conference on Multimedia Retrieval (ICMR '12). ACM, New York, NY, USA, Article 26, 8 pages. DOI=10.1145/2324796.2324829 http://doi.acm.org/10.1145/2324796.2324829

[40] L. Aroyo, P. Bellekens, M. Bjorkman, G.J. Houben, P. Akkermans, and A. Kaptein. 2007. SenSee framework for personalized access to TV content. In Proceedings of the 5th European conference on Interactive TV: a shared experience (EuroITV '07), Pablo Cesar, Konstantinos Chorianopoulos, and Jens F. Jensen (Eds.). Springer-Verlag, Berlin, Heidelberg, 156-165.

[41] L.A. Rowe and R. Jain. 2005. ACM SIGMM retreat report on future directions in multimedia research. ACM Trans. Multimedia Comput. Commun. Appl. 1, 1 (February 2005), 3-13. DOI=10.1145/1047936.1047938 http://doi.acm.org/10.1145/1047936.1047938

[42] L. Dib, D. Petrelli, and S. Whittaker. 2010. Sonic souvenirs: exploring the paradoxes of recorded sound for family remembering. In Proceedings of the 2010 ACM conference on Computer supported cooperative work (CSCW '10). ACM, New York, NY, USA, 391-400. DOI=10.1145/1718918.1718985 http://doi.acm.org/10.1145/1718918.1718985

[43] L. Hardman, G. van Rossum, and D.C.A. Bulterman. 1993. Structured multimedia authoring. In Proceedings of the first ACM international conference on Multimedia (MULTIMEDIA '93). ACM, New York, NY, USA, 283-289. DOI=10.1145/166266.168402 http://doi.acm.org/10.1145/166266.168402

[44] L. Kennedy and M. Naaman. 2009. Less talk, more rock: automated organization of community-contributed collections of concert videos. In Proceedings of the 18th international conference on World wide web (WWW '09). ACM, New York, NY, USA, 311-320. DOI=10.1145/1526709.1526752 http://doi.acm.org/10.1145/1526709.1526752

[45] L. Rainie, J. Brenner and K. Purcell. 2012. Photos and Videos as Social Currency Online, The Pew Research Center's Internet & American Life Project, September 2012. Available at http://pewinternet.org/Reports/2012/Online-Pictures.aspx

[46] L. Xie, A. Natsev, J.R. Kender, M. Hill, and J.R. Smith. 2011. Visual memes in social media: tracking real-world news in YouTube videos. In Proceedings of the 19th ACM international conference on Multimedia (MM '11). ACM, New York, NY, USA, 53-62. DOI=10.1145/2072298.2072307 http://doi.acm.org/10.1145/2072298.2072307

[47] M.A. Hearst. 2009. Search User Interfaces (1st ed.). Cambridge University Press, New York, NY, USA. ISBN:0521113792 9780521113793 http://dl.acm.org/citation.cfm?id=1631268

[48] M. Ames and M. Naaman. 2007. Why we tag: motivations for annotation in mobile and online media. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07). ACM, New York, NY, USA, 971-980. DOI=10.1145/1240624.1240772 http://doi.acm.org/10.1145/1240624.1240772

[49] M. Cavazza, R. Champagnat, and R. Leonardi. 2009. The IRIS Network of Excellence: Future Directions in Interactive Storytelling. In Proceedings of the 2nd Joint International Conference on Interactive Digital Storytelling: Interactive Storytelling (ICIDS '09), Ido A. Iurgel, Nelson Zagalo, and Paolo Petta (Eds.). Springer-Verlag, Berlin, Heidelberg, 8-13. DOI=10.1007/978-3-642-10643-9_4 http://dx.doi.org/10.1007/978-3-642-10643-9_4

[50] M.D. Choudhury, H. Sundaram, A. John, and D.D. Seligmann. 2009. What makes conversations interesting?: themes, participants and consequences of conversations in online social media. In Proceedings of the 18th international conference on World wide web (WWW '09). ACM, New York, NY, USA, 331-340.DOI=10.1145/1526709.1526754 http://doi.acm.org/10.1145/1526709.1526754

[51] M. Frantzis, V. Zsombori, M. Ursu, R.L. Guimarães, I. Kegel and R. Craigie. 2012. Interactive Video Stories from User Generated Content: a School Concert Use Case. In Proceedings of the 5th International Conference on Interactive Digital Storytelling (ICIDS '12). Springer Berlin Heidelberg, 183-195. DOI=10.1007/978-3-642-34851-8_18 http://dx.doi.org/10.1007/978-3-642-34851-8_18

[52] M.F. Ursu, M. Thomas, I. Kegel, D. Williams, M. Tuomola, I. Lindstedt, T. Wright, A. Leurdijk, V. Zsombori, J. Sussner, U. Myrestam, and N. Hall. 2008. Interactive TV narratives: Opportunities, progress, and challenges. ACM Trans. Multimedia Comput. Commun. Appl. 4, 4, Article 25 (November 2008), 39 pages. DOI=10.1145/1412196.1412198 http://doi.acm.org/10.1145/1412196.1412198

[53] M.K. Saini, R. Gadde, S. Yan, and W.T. Ooi. 2012. MoViMash: online mobile video mashup. In Proceedings of the 20th ACM international conference on Multimedia (MM '12). ACM, New York, NY, USA, 139-148. DOI=10.1145/2393347.2393373 http://doi.acm.org/10.1145/2393347.2393373

[54] M.O. Riedl and R.M. Young. 2006. From Linear Story Generation to Branching Story Graphs. IEEE Comput. Graph. Appl. 26, 3 (May 2006), 23-31. DOI=10.1109/MCG.2006.56 http://dx.doi.org/10.1109/MCG.2006.56

[55] M.S. Granovetter. 1973. The Strength of Weak Ties. American Journal of Sociology, 78(6): 1360-1380.

[56] P. Cesar, D.C.A. Bulterman, D. Geerts, J. Jansen, H. Knoche, and W. Seager. 2008. Enhancing social sharing of videos: fragment, annotate, enrich, and share. In Proceedings of the 16th ACM international conference on Multimedia (MM '08). ACM, New York, NY, USA, 11-20. DOI=10.1145/1459359.1459362 http://doi.acm.org/10.1145/1459359.1459362

[57] P. Cesar, D.C.A. Bulterman, R.L. Guimarães and I. Kegel. 2010. Web-Mediated Communication: in Search of Togetherness. In Proceedings of the 2nd Web Science Conference (WebSci 2010). Available at http://journal.webscience.org/371/

[58] P. Obrador, R. Oliveira, and N. Oliver. 2010. Supporting personal photo storytelling for social albums. In Proceedings of the international conference on Multimedia (MM '10). ACM, New York, NY, USA, 561-570.

DOI=10.1145/1873951.1874025
http://doi.acm.org/10.1145/1873951.1874025

[59] P. Shrestha, P.H.N. de With, H. Weda, M. Barbieri, and E.H.L. Aarts. 2010. Automatic mashup generation from multiple-camera concert recordings. In Proceedings of the international conference on Multimedia (MM '10). ACM, New York, NY, USA, 541-550. DOI=10.1145/1873951.1874023 http://doi.acm.org/10.1145/1873951.1874023

[60] R. Fagá Jr., B.C. Furtado, F. Maximino, R.G. Cattelan, and M.G.C. Pimentel. 2009. Context information exchange and sharing in a peer-to-peer community: a video annotation scenario. In Proceedings of the 27th ACM international conference on Design of communication (SIGDOC '09). ACM, New York, NY, USA, 265-272. DOI=10.1145/1621995.1622048 http://doi.acm.org/10.1145/1621995.1622048

[61] R.G. Cattelan, C. Teixeira, R. Goularte, and M.G.C. Pimentel. 2008. Watch-and-comment as a paradigm toward ubiquitous interactive video editing. ACM Trans. Multimedia Comput. Commun. Appl. 4, 4, Article 28 (November 2008), 24 pages. DOI=10.1145/1412196.1412201 http://doi.acm.org/10.1145/1412196.1412201

[62] R.L. Guimarães, P. Cesar, and D.C.A. Bulterman. 2010. Creating and sharing personalized time-based annotations of videos on the web. In Proceedings of the 10th ACM symposium on Document engineering (DocEng '10). ACM, New York, NY, USA, 27-36. DOI=10.1145/1860559.1860567 http://doi.acm.org/10.1145/1860559.1860567

[63] R.L. Guimarães, P. Cesar, D.C.A. Bulterman, I. Kegel, and P. Ljungstrand. 2011. Social Practices around Personal Videos using the Web. In Proceedings of the ACM International Conference on Web Science. Available at http://journal.webscience.org/437/

[64] R.L. Guimarães, R. Kaiser, A. Hofmann, P. Cesar, and D. Bulterman. 2010. Video analysis tools for annotating user-generated content from social events. In Proceedings of the 5th international conference on Semantic and digital media technologies (SAMT '10), Thierry Declerck, Michael Granitzer, Marcin Grzegorzek, Massimo Romanelli, and Stefan Rüger (Eds.). Springer-Verlag, Berlin, Heidelberg, 188-189. DOI= 10.1007/978-3-642-23017-2_14 http://www.springerlink.com/content/3x57j441142825h9/

[65] R. Lienhart. 1999. Abstracting home video automatically. In Proceedings of the seventh ACM international conference on Multimedia (Part 2) (MULTIMEDIA '99). ACM, New York, NY, USA, 37-40. DOI=10.1145/319878.319888 http://doi.acm.org/10.1145/319878.319888

[66] R. Pea, M. Mills, J. Rosen, K. Dauber, W. Effelsberg, and E. Hoffert. 2004. The Diver Project: Interactive Digital Video Repurposing. IEEE MultiMedia 11, 1 (January 2004), 54-61. DOI=10.1109/MMUL.2004.1261108 http://dx.doi.org/10.1109/MMUL.2004.1261108

[67] R. Rettie. 2003. Connectedness, awareness and social presence. In Proceedings of the 6th Annual International Workshop on Presence. Available at http://eprints.kingston.ac.uk/2106/. Last access on May 15th 2013.

[68] S. Bocconi, F. Nack, and L. Hardman. 2008. Automatic generation of matter-of-opinion video documentaries. Web Semant. 6, 2 (April 2008), 139-150. DOI=10.1016/j.websem.2008.01.004 http://dx.doi.org/10.1016/j.websem.2008.01.004

[69] S. Eisenstein. 1949. Film Form: Essays in Film Theory. New York: Hartcourt, 279 pages, ISBN 0156309203

[70] S. Lederer, C. Mueller, C. Timmerer, C. Concolato, J. Le Feuvre, and K. Fliegel. 2013. Distributed DASH dataset. In Proceedings of the 4th ACM Multimedia Systems Conference (MMSys '13). ACM, New York, NY, USA, 131-135.DOI=10.1145/2483977.2483994 http://doi.acm.org/10.1145/2483977.2483994

[71] S.U. Naci and A. Hanjalic. 2007. Intelligent browsing of concert videos. In Proceedings of the 15th international conference on Multimedia (MULTIMEDIA '07). ACM, New York, NY, USA, 150-151. DOI=10.1145/1291233.1291264 http://doi.acm.org/10.1145/1291233.1291264

[72] S. Whittaker, O. Bergman, and P. Clough. 2010. Easy on that trigger dad: a study of long term family photo retrieval. Personal Ubiquitous Comput. 14, 1 (January 2010), 31-43. DOI=10.1007/s00779-009-0218-7 http://dx.doi.org/10.1007/s00779-009-0218-7

[73] The Nielsen Company. 2010. How People Watch – A Global Nielsen Consumer Report. Available at http://www.nielsen.com/us/en/insights/reports-downloads/2010/How-We-

Watch-The-Global-State-of-Video-Consumption.html. Last access on May 15th 2013.

[74] U. Westermann and R. Jain. 2007. Toward a Common Event Model for Multimedia Applications. IEEE MultiMedia 14, 1 (January 2007), 19-29. DOI=10.1109/MMUL.2007.23 http://dx.doi.org/10.1109/MMUL.2007.23

[75] V.K. Singh, J. Luo, D. Joshi, P. Lei, M. Das, and P. Stubler. 2011. Reliving on demand: a total viewer experience. In Proceedings of the 19th ACM international conference on Multimedia (MM '11). ACM, New York, NY, USA, 333-342. DOI=10.1145/2072298.2072342 http://doi.acm.org/10.1145/2072298.2072342

[76] V. Zsombori, M. Frantzis, R.L. Guimaraes, M.F. Ursu, P. Cesar, I. Kegel, R. Craigie, and D.C.A. Bulterman. 2011. Automatic generation of video narratives from shared UGC. In Proceedings of the 22nd ACM conference on Hypertext and hypermedia (HT '11). ACM, New York, NY, USA, 325-334. DOI=10.1145/1995966.1996009 http://doi.acm.org/10.1145/1995966.1996009

[77] Z. Yu, N. Diakopoulos, and M. Naaman. The multiplayer: multi-perspective social video navigation. In Adjunct proceedings of the 23nd annual ACM symposium on User interface software and technology (UIST '10). ACM, New York, NY, USA, 413-414. DOI=10.1145/1866218.1866246 http://doi.acm.org/10.1145/1866218.1866246

Summary

Creating compelling multimedia productions is a non-trivial problem. This is true for both professional and personal content. For professional content, extensive production support is typically available during creation. Content assets are well structured, content fragments are professionally produced with high quality, and production assets are often highly annotated (within the scope of the production model). For personal content, nearly none of these conditions exist: content is a collection of assets that are structured only by linear recording time, of mediocre technical quality (on an absolute scale), and with only basic automatic annotations. These conditions limit the options open to casual authors and to viewers of rich multimedia content in creating and receiving focused, highly personal media presentations. The problem is compounded when authors want to integrate community media assets: media fragments donated from a potentially wide and anonymous recording community. In this thesis we reflect on the traditional multimedia authoring workflow and we argue that a fresh new look is required. Our experimental methodology aims at meeting the requirements needed for social communities that are not addressed by traditional authoring and sharing applications. We focus on the particular task of supporting *socially-aware multimedia authoring*, in which the relationships within particular social groups can be exploited to create highly personal media experiences. Our framework is centered on empowering users in telling stories and commenting on personal media artifacts, considering the long-term social context of the user. The work has been evaluated through a number of prototype tools that allow users to explore, create, enrich and share rich multimedia artifacts. Results from our evaluation process provide useful insights into how a socially-aware multimedia authoring and sharing system should be designed and architected, for helping users in recalling personal memories and in nurturing their close circle relationships.

Nederlandse Samenvatting

Goede, aantrekkelijke multimedia producties maken is een ingewikkeld probleem. Dit geldt voor zowel professionele als persoonlijke mediadocumenten. Voor professionele producties is in de regel een uitgebreid instrumentarium ter beschikking. De fragmenten zijn goed gestructureerd, professioneel gemaakt, van uitstekende kwaliteit en vaak van annotaties voorzien (binnen de kaders van het productie model). Voor persoonlijke mediadocumenten gelden bijna geen van deze condities: het materiaal is doorgaans een verzameling fragmenten met als enige structuur de lineaire opnametijd, een (op een absolute schaal) matige technische kwaliteit en minimaal van annotaties voorzien. Deze condities beperken de mogelijkheden voor minder ervaren producenten en consumenten om verrijkte multimedia presentaties te maken die toegespitst en in hoge mate persoonlijk zijn. Dit is nog lastiger als men ook gemeenschappelijk materiaal wil gebruiken: media fragmenten beschikbaar gesteld door andere, potentieel uiteenlopende en vaak anonieme bronnen. In dit proefschrift bekijken wij de traditionele manier van werken en stellen dat een nieuwe kijk nodig is. Onze experimentele aanpak is gericht op de behoefte van groepen mensen die niet geboden wordt door de gebruikelijke programmatuur voor het bewerken en verspreiden van multimedia. Wij kijken in het bijzonder naar ondersteuning voor het vervaardigen van 'sociaal-bewuste multimediale producties', waarbij relaties binnen bepaalde sociale groepen kunnen worden gebruikt om zeer persoonlijke multimediale ervaringen te creëren. Onze opzet is erop gericht om mensen in staat te stellen om hun persoonlijk verhaal te vertellen en commentaar te geven bij persoonlijke media artefacten, waarbij de bestendige sociale context van de gebruiker een rol speelt. Dit werk is geëvalueerd door een aantal prototypen te ontwikkelen, waarmee mensen hun multimedia producten kunnen overzien, bewerken, verrijken en verspreiden. Resultaten van deze evaluatie leveren goed bruikbare inzichten op hoe men het beste systemen kan ontwerpen voor sociaal-bewuste media productie en verspreiding waarbij gebruikers in staat worden gesteld om hun persoonlijke herinneringen op te halen en daarmee de relaties in hun eigen kring te onderhouden.

Curriculum Vitae

1981 Born on July 9th in Vitória/ES, Brazil.

1996–1999 Technical High School (Electrotechnics), Federal Center for Technological Education of Espírito Santo (CEFETES), Vitória/ES, Brazil.

1999–2004 Engineer (Computer Engineering), Federal University of Espírito Santo (UFES), Vitória/ES, Brazil.
Final project: *Desenvolvimento da Infraestrutura Computacional de Gerenciamento e Comunicação de Dados do Ambiente MultiJADE*, supervised by prof.dr. F.M. Varejão.

2005–2007 M.Sc. in Informatics (Computer Networks and Hypermedia Systems), Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Rio de Janeiro/RJ, Brazil.
Dissertation: *Composer – An Authoring Tool of NCL Documents for Interactive Digital TV*, supervised by prof.dr. L.F.G. Soares.

2007–2012 Ph.D. student in the Distributed and Interactive Systems group, Centrum Wiskunde & Informatica, Amsterdam, The Netherlands.
Thesis: *Socially-Aware Multimedia Authoring*, supervised by prof.dr. D.C.A Bulterman and dr. P.S. Cesar.