Socially-Aware Multimedia Authoring

Laiola Guimaraes, R.

2014

document version
Publisher's PDF, also known as Version of record

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/

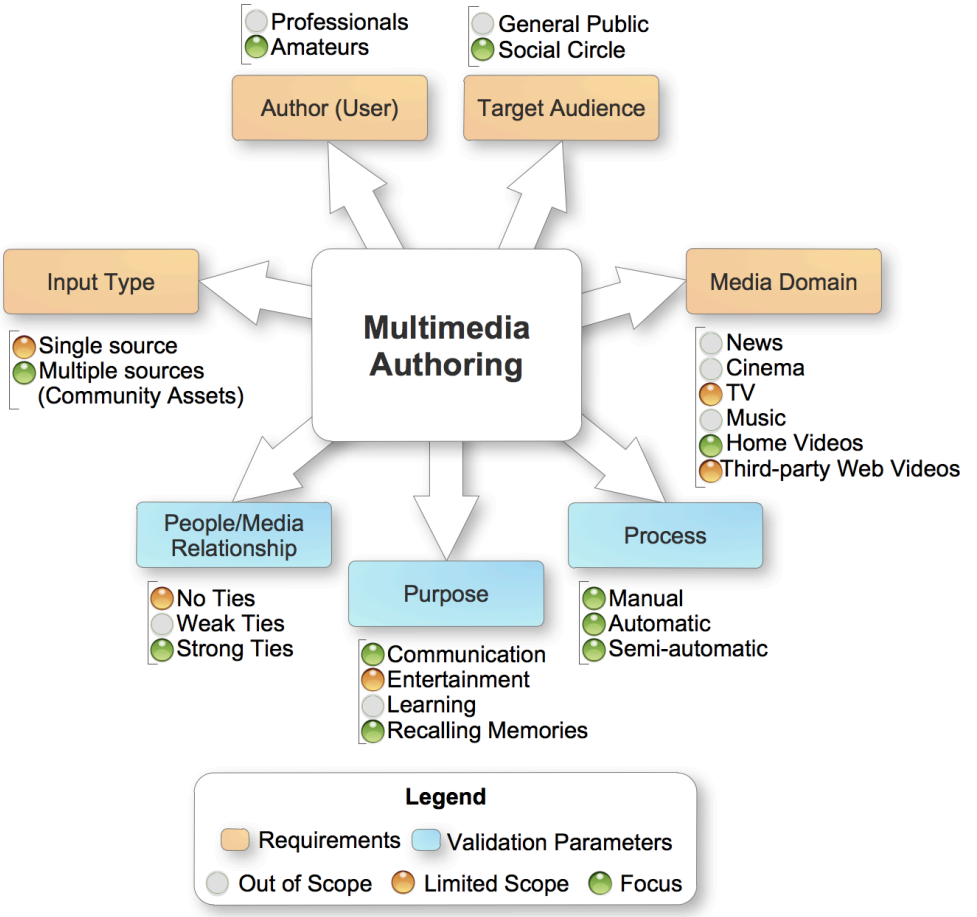


Figure 2.1. Overview of the requirements and validation parameters for socially-aware multimedia authoring systems.

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.

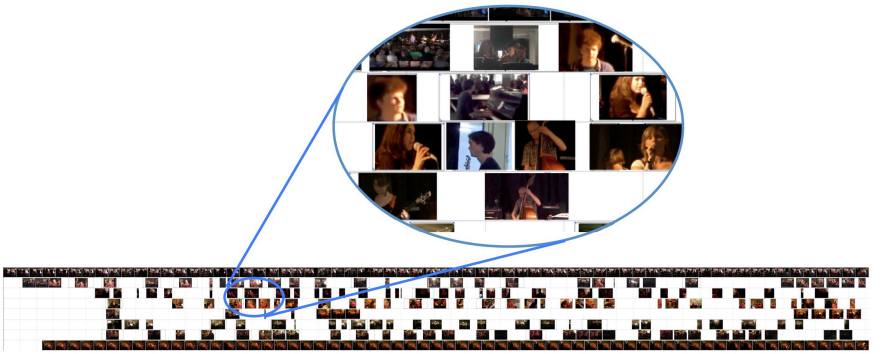


Figure 2.2. Temporal alignment of a real life data set from a concert, where a community of users recorded video clips.

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

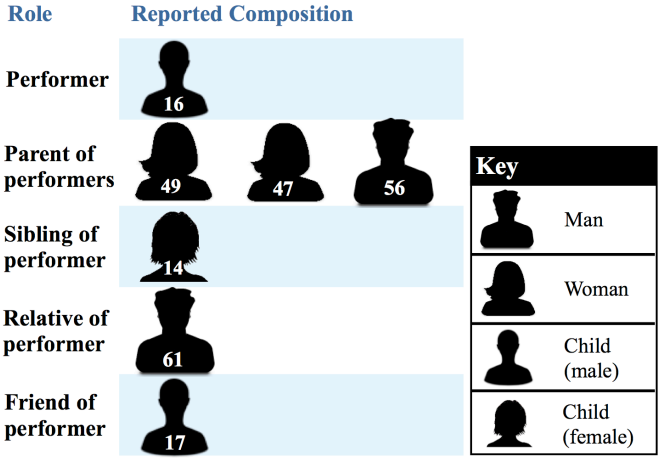


Figure 2.3. The makeup, age and gender of participants in phase 1 evaluation.

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

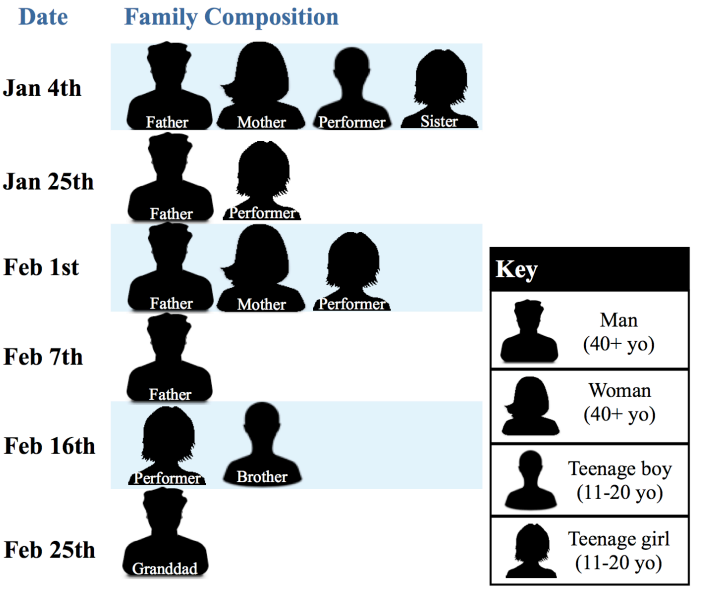


Figure 2.4. The makeup, age and gender of participants in phase 2 evaluation.

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

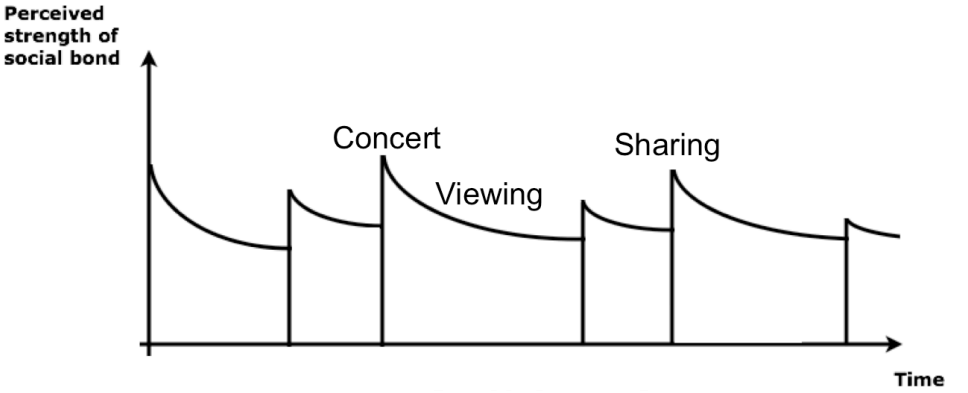


Figure 2.5. A schematic view of the perceived strength of social bond over time in relation to our scenario.

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:

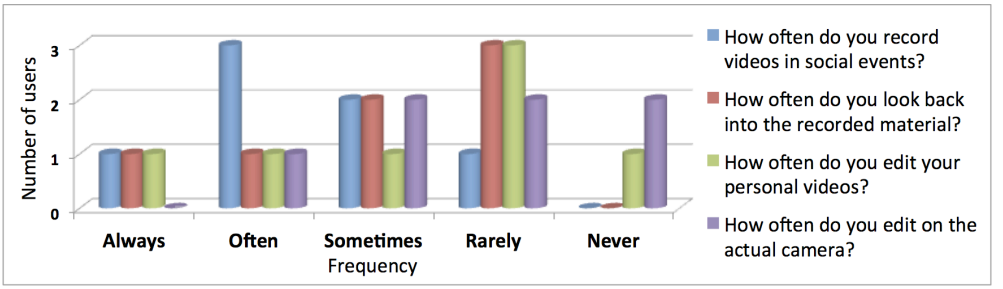
- A willingness to engage in diversified forms of recollections through recorded videos;
- A clear requirement for systems that could be trusted as ensuring privacy;
- A positive reaction to the suggestion of automatic and intelligent mechanisms for managing home videos.

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

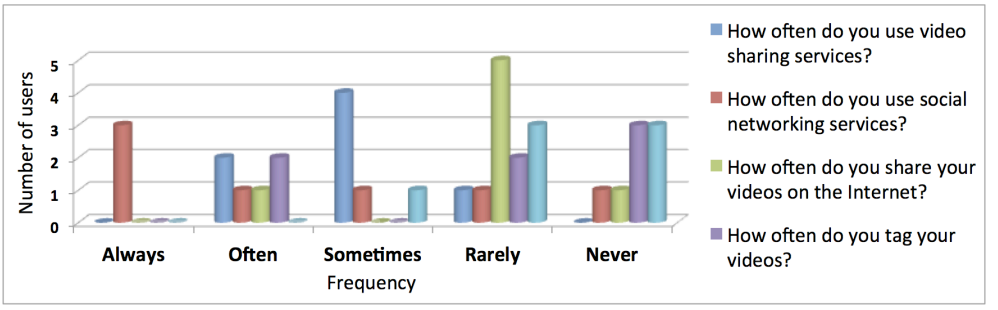
2.2.3 Requirements Gathering

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.

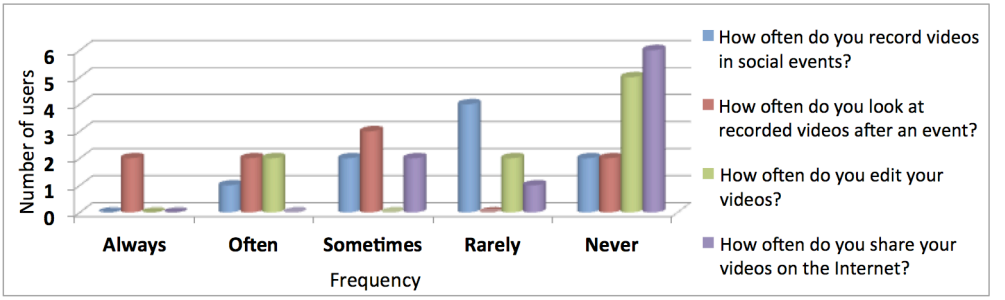


Figure 2.7. Social media habits (phase 2 evaluation).

> *"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

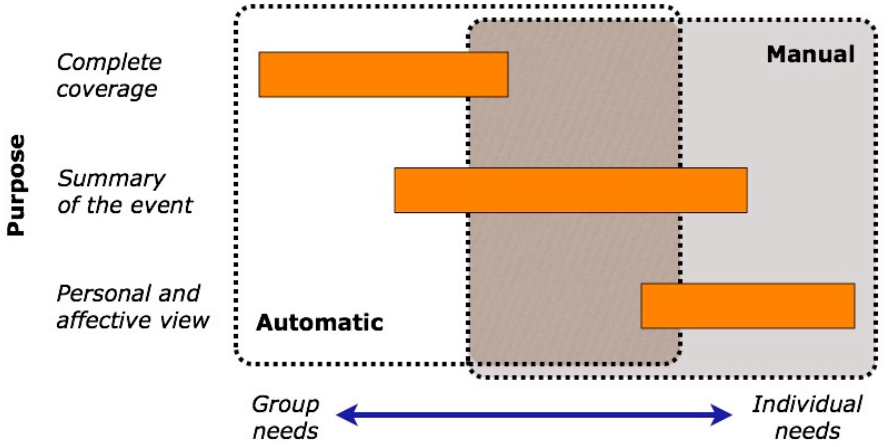


Figure 2.8. Comparison between automatic and manual generation of video compilations. Automatic methods are not sufficient for creating personal and intimate memories.

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.

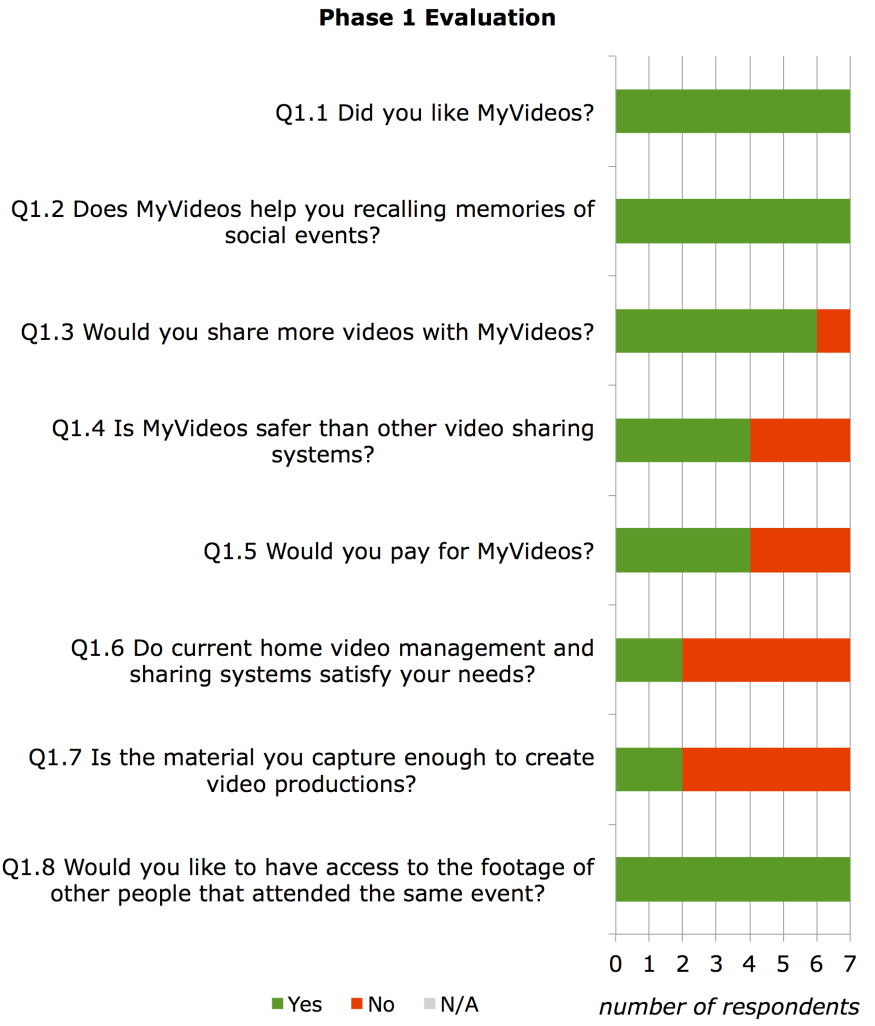


Figure 2.9. Utility and usefulness of MyVideos. Results of the questionnaire from phase 1 evaluation (Amsterdam/NL).

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.

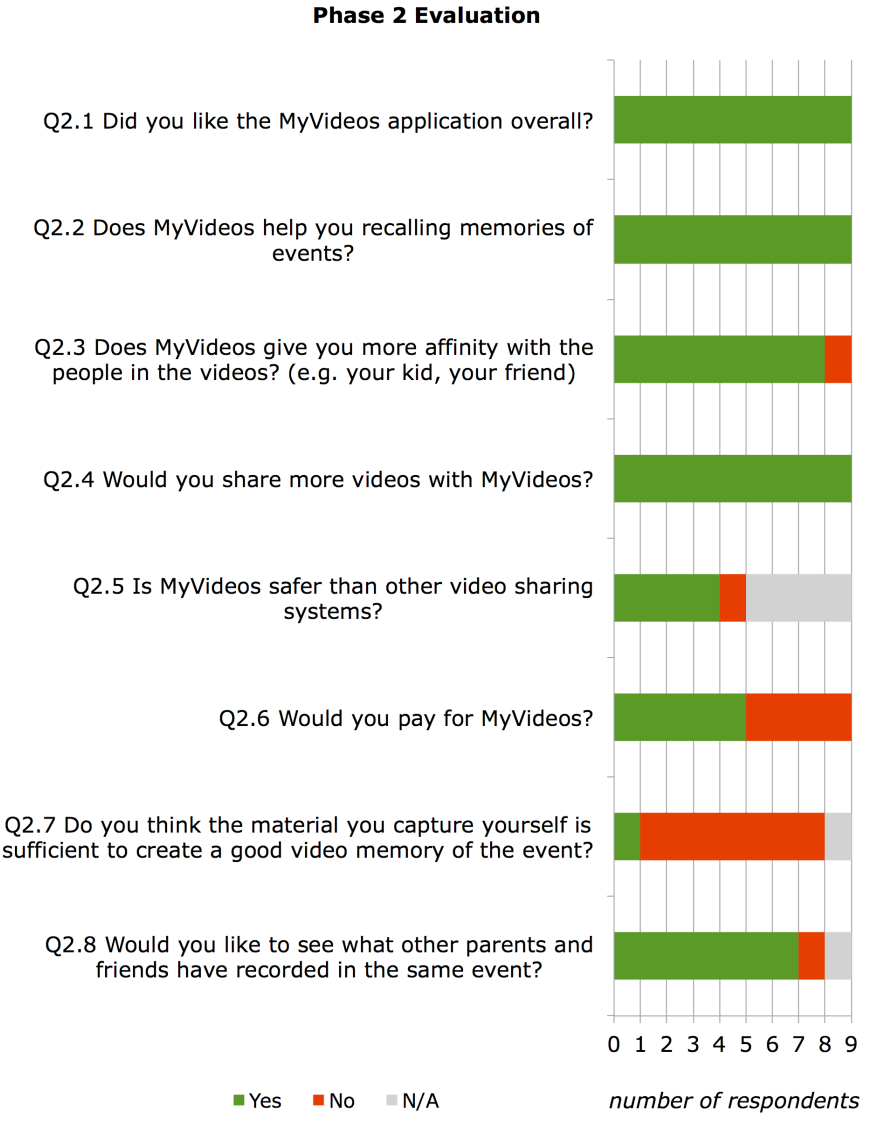


Figure 2.10. Utility and usefulness of MyVideos. Results of the questionnaire from phase 2 evaluation (Woodbridge/UK).

"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).