Automatic Generation of Video Narratives from Shared UGC

Vilmos Zsombori[1], Michael Frantzis[1], Rodrigo Laiola Guimaraes[2], Marian F. Ursu[1], Pablo Cesar[2], Ian Kegel[3], Roland Craigie[3], Dick C.A. Bulterman[2]

[1] Department of Computing
Goldsmiths, University of London
United Kingdom

[2] CWI: Centrum Wiskunde & Informatica
The Netherlands

[3] BT Research & Technology
United Kingdom

v.zsombori@gold.ac.uk, m.frantzis@gold.ac.uk, rlaiola@cwi.nl, m.ursu@gold.ac.uk, p.s.cesar@cwi.nl, ian.c.kegel@bt.com, roland.craigie@bt.com, dick.bulterman@cwi.nl

ABSTRACT

This paper introduces an evaluated approach to the automatic generation of video narratives from user generated content gathered in a shared repository. In the context of social events, end-users record video material with their personal cameras and upload the content to a common repository. Video narrative techniques, implemented using Narrative Structure Language (NSL) and ShapeShifting Media [Ursu, 2008a], are employed to automatically generate movies recounting the event. Such movies are personalized according to the preferences expressed by each individual end-user, for each individual viewing. This paper describes our prototype narrative system, *MyVideos*, deployed as a web application, and reports on its evaluation for one specific use case: assembling stories of a school concert by parents, relatives and friends. The evaluations carried out through focus groups, interviews and field trials, in the Netherlands and UK, provided validating results and further insights into this approach.

Categories and Subject Descriptors

H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems – *Audio, Video*. H.5.2 [**Information Interfaces and Presentation**]: User Interfaces - *User-centered design*. I.7.2 [**Document and Text Processing**]: Document Preparation - *Format and notation, hypertext/hypermedia, Languages and Systems, Multi/mixed media.*

General Terms

Design, Experimentation, Human Factors

Keywords

interactive narrative, interactive storytelling, digital storytelling, computational narrativity, video, user generated content, media share, interactive television, NSL, ShapeShifting media, SMIL.

HT'11, June 6–9, 2011, Eindhoven, The Netherlands.

1. INTRODUCTION

Recent advances in non-professional video camera technologies have transformed the way we capture the important events and experiences in our lives. Cheaper, easy to carry and operate, video cameras now seem to always be available to record moments of the social events in which we partake. The richness of video makes it a very attractive recording medium for what we believe could later become valuable memories. Nevertheless, the ability to *capture* video recordings with such ease does not necessarily result in them becoming easily accessible memory objects that generate attractive and rewarding recollection experiences. It is too often the case that such recordings remain abandoned on memory cards or as downloaded files on hard drives never to be accessed again.

The main reason for this is that, as captured, a video is not final and ready for being looked at. Video, as a time-based medium, necessarily requires processing after capture, i.e. editing, and video editing is a difficult and time-consuming task. Editing, for instance, is required 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, whilst filming. The low cost of recording magnifies this problem. Editing is also required to create rewarding narrative experiences that we want to see over and over again. A simple juxtaposition of recordings that capture only fragments of social events does not necessarily create attractive mementos. But editing is not a simple process and people do not want to engage with it. The research question we pose here is: could technology help alleviate this problem? More specifically, could automatic narrative techniques be developed to help end-users compile attractive video recordings of their social events?

The above perspective puts the accent on the individual relationship between creator and content. But there is also a social dimension. In social events, such as school concerts, participants capture substantial amounts of video. Normally, each person records primarily their own children and possibly their close friends performing, whilst at the same time trying to capture sufficient contextual information for a telling background. Each personal point of view is also reflected in the style of filming: duration of shots, framing, and camera movement (pan, tilt, zoom). Not everybody films at the same time, but many might be filming concurrently. Recordings happen before, during and after

the event. All these, together, constitute a rich source of material that potentially could lead to the creation of a wide variety of movies recounting the event, varying in, for example, duration, main protagonists, story told and style. Yet, content cannot be shared in this way, as, currently, there are no tools that support the creation of stories from a common pool of video recordings of a social event.

To reiterate, our main research question is

Could video narrative concepts and techniques be employed to the generation of personalized movies from a common repository of user-generated recordings of social events?

We have approached it from three more concrete and strongly interlinked perspectives:

1. *system development*: could tools be developed for the automating preparation of the raw content and its subsequent assembly in customizable video narrations?
2. *creation of video narration*: could bona-fide narrative techniques be automated by such a system?
3. *usefulness*: do end-users like the features of the system and, ultimately, the video narrations automatically produced?

In this paper we present an affirmative answer obtained in the context of a representative use case: a school performance captured on various cameras by the children's parents, relatives and friends. To carry out this research, we built a prototype video narrative system, called *MyVideos*, using the production-independent Narrative Structure Language (NSL) and the corresponding ShapeShifting Media toolkit (authoring tool and reasoner) [Ursu, 2008a].

The paper is structured as follows. Section 2 provides the background. It briefly reiterates results of some of our interviews carried out in 2008 with focus groups in four European countries, that motivate our work, and contextualizes our approach with reference to related research. Section 3 presents the approach. It outlines the narrative concepts that underlie *MyVideos*, thus presenting our stance on the research question 2, and provides an overview of the actual system. Section 4 provides the technical details behind the implementation of *MyVideos*, directly responding to the research question 1. Section 5 describes results of our evaluations carried out with video and film professionals from UK (thus answering research question 2) and with end-users in the Netherlands (thus answering research question 3). Section 6 presents concluding remarks and directions for further work.

2. BACKGROUND

This section states that currently there are no tools that satisfy the needs of users for creating customized video narrations from content recorded in social events and stored in a shared common repository. It backs up this statement through an investigation carried out with focus groups and a survey of related work.

2.1 Motivation through Focus Groups

In 2008, we conducted interviews with focus groups – sixteen families across four countries, UK, Sweden, Netherlands, and Germany – with a view to understanding their practice regarding the use of video recordings of the social events in which they take part, their perceived limitations of current systems, and desired functionalities that could increase their involvement with video recordings. Our findings were reported in [Williams, 2009], but some of the major points are reiterated in this section, as motivation for building *MyVideos*.

The overall conclusion was that the current models and systems for video authoring and sharing known by the participants did not fit the needs of family, friends and social groups. All the participants reported that they record videos in social events, but, they barely look at this material afterwards. Most of the participants described video editing as far too lengthy and complicated, thus utterly unattractive, despite perceiving it as an essential process in the preparation of video memory objects.

When prompted about desired functionalities, the participants' responses converged to:

- a willingness to engage in varied forms of recollections and sharing through recorded video, providing suitable systems were in place to support such activities
- an overwhelmingly positive reaction to the suggestion of automatic, intelligent video compilations from a common shared pool of videos, in the form of personalized movies
- a clear requirement for systems that could be trusted to ensure privacy.

These conclusions, clearly motivating our research, represent only a product (market) perspective. The following subsection looks into related research propositions.

2.2 Related Work

The first perspective we take is that of the rich media networking, the success of which on the Web has fundamentally changed the media landscape. End-users now play an active role in the media creation and distribution chain. Media content analysis based on events [Kennedy, 2009], semantic understanding from aggregated end-user comments [Shamma, 2009] and video summarization [Truong, 2007] are just a few examples of initiatives that aim to help the end-users' access to media. As opposed to this, we aim to improve the end-users' access to their own recordings as well as their ability to share it through automatic editing of raw material. We focus on community media and the employment of sound narrative techniques to better provide for the large variety of individual needs.

The second perspective is that of repurposing rich media. Shamma [2007] proposes a community video remix. Their approach does not consider the relationship between performers, authors, and recipients as part of the same social circle. Other examples include the automatic generation of video mash-ups from YouTube content [Shrestha, 2010], social creation of photo albums [Obrador, 2010], and synchronization and organization of user-generated content from popular music events [Kennedy, 2009]. Nevertheless, none of the above solutions provide a narrative engine founded in established narrative principles for the generation of the video compilations.

The third perspective is that of configurable and interactive storytelling. Various AI approaches have been suggested, and some reviews are presented in [Cavazza, 2009], [Ursu, 2007] and [Riedl, 2006]. These techniques are predominantly aimed at *generating* narrative behaviour, leading to less structured and more emergent narratives [Ibanez, 2009], rather than applying pre-authored narrative guidelines to the automatic compilation particular narratives, which is our approach. Furthermore, the main bulk of interactive storytelling has been applied to generated

not recorded content. Nevertheless, examples of configurable or interactive video narratives exist. A recent result is reported in [Porteous, 2010], which describes a video-based storytelling prototype that is able to generate multiple story variants from a baseline video. This is more related to video summarization than to narrative aggregation of content from various sources. Vox Populi [Bocconi 2008] is another representative example, in which rhetorical documentaries are created from a pool of media fragments. Examples like these highlight a production-specific implementation approach, with the narrative structures hard-coded in general purpose programming languages. The Narrative Structure Language (NSL) and its associated ShapeShifting Media toolkit [Ursu, 2008a] constitute a production-independent framework for the authoring and delivery of configurable and interactive video narratives. Thus far, they have been employed only in professional settings (see, for example, [Ursu, 2008b; 2009]). Here, complemented by other automatic processes, such as generation of media fragments and automatic generation of annotations, they are investigated in the context of UGC.

Although a succinct review, we can conclude that our research questions have not yet been asked by the research community. The following section presents our overall approach.

3. APPROACH

We set out to build a narrative system, *MyVideos*, whose main functionality is the ability to automatically edit personalized video compilations (movies). The people who participate in a social event, a school concert, in this case, need only to upload their video recordings into the system's repository and then, each time they require a new viewing, simply set a number of configuration parameters, such as duration and main protagonist. For each such configuration, the narrative system selects the most appropriate content from the shared repository and edits it into a video compilation of professional-like quality.

This section presents an overview of the narrative principles on which the automatic compilation is based, followed by an overview of the narrative system, which is founded on ShapeShifting Media [Ursu, 2008a].

3.1 Narrative Concept

Social events and shared group activities generate experiences that could be framed in a multitude of stories. Through recounting, we relive and recreate them. The act of recounting can be seen as creating a dichotomy: the event – i.e. the object of recounting – and the storytellers, a subset of the spectators, are the two emerging concepts. In turn, they generate two perspectives on structuring the potential stories, namely:

1. the *event*: this consists of all the individual events and actions in which the participating actors are involved; for example, in the case of a school concert, individual events include the initial preparations, the sequence of songs, the reception following the performance, or more granular events, such as solos and refrains within songs and even the pauses between them; this represents a neutral, objective view, disconnected from any particular interpretation or narration; strictly logically, this concept is speculative as it has no concrete materialization: any incarnation of it inherently falls into the next category, below; nevertheless, even if it is unattainable, it is useful as it introduces a frame of reference; attainable approximations could be defined on its basis and used as structural means for storytelling
2. the *storyteller*: it acknowledges the unavoidable subjectivity of any narration and includes all the individual interpretations, through narrations, of the event; this includes the event structure, but it includes further aspects such as focus of attention (interest in a particular performer, song or instrument) and personal emotional response to the event in focus (e.g., happiness, fun, boredom, etc.)

Simplistically put, the former perspective defines a general *story space*, whereas the latter includes all the possible concrete stories (within the story space). Told stories are carried by a medium. In our case, this is video. The medium introduces another perspective, namely that of

3. the *recording*: this represents the events as they are captured on camera; they subsume the narrative intent of the particular storytellers, through the choices and viewpoints of the persons doing the filming, and, implicitly, the event structure; but they have yet another inherent structure, namely that of the video capture techniques, including aspects such as types of shots and camera movement.

This conceptual framework allows us to structure the content capture process. The perspective of the *event* – the school concert – could be captured by video professionals. They could record master video and audio tracks and these should be continuous and cover the entire event. The professionals could also structure this content according to the events they identify. The resulting material could constitute the backbone of any particular narration.

Parents, relatives and friends capture individual interpretations of the event. They, most probably, are fragments of the event itself. If they are aligned in time with the master tracks, then they can inherit the event structure. For a music driven event, time alignment is possible on the basis of the audio track. Narrative intent could be captured for each individual, in a number of attributes, and propagated to all the recorded content.

The professionals' video capture techniques could be informed by the ultimate aim of automatically constructing configured narratives and thus used proactively. The stylistic structures could then be explicitly defined, in a similar manner to event annotation. The techniques employed in capturing the individual fragments may also be identified and exploited, reactively. This structured material constitutes a rich source for constructing individual narrations.

The story generation i.e. video editing, is a process intended to be automated. Good editing is about assembling the best available fit of video material to the succession of events, on one hand, and the contextual requirements, on the other. In our case, the context is represented by the identity of the person requesting the narration, from which his/her personal relationships with the individual performers in the concert can be inferred, in addition to their configuration choices. An automatic editing process thus requires an understanding of how each particular video clip corresponds to the general event structure and how it matches the configuration parameters. Automatic editing must also be aware of the cinematic rules of good practice.

In a concert, it is the music that provides the core shared narrative structure and the edits have to appear reactive to the individual musical events. For example, if an engaging solo starts, then good editing practice would demand closer shots of the solo performer. However, if a viewer expresses a desire to see a particular performer, then their close shots may need to replace some of the

close shots of the soloists. The right balance is informed by good editing practice and personal stylistic choices of the author. They also inform the granularity of the cuts, the use of the footage with regards to the capture styles and, possibly, even decisions regarding continuity of image quality (light, contrast, etc.).

We summarize the main requirements for automatic editing as the ability to:

1. structure the master tracks according to the music events
2. create media clips, from the recorded raw video content, that can be easily compiled in sequence – they are the basic building blocks of the individual narrations – and align them with the music master track
3. structure the visual content of the video fragments with regards to the general music narrative events, but also the particular interpretations
4. define a set of computational narrative structures that capture both narrative intent and good cinematic practice, which would inform each particular edit

We have experimented with two main computational narrative structures, one for the creation of complete song compilations, and the other for a video digest of a whole event. The former is sketched here in the form of a set of rules expressed in natural language and, for better readability, significantly simplified (they are incomplete and expressed considering only one configuration parameter: preferred performer):

1. Extract chosen song from master audio track and use its structure as backbone of the narration.
2. Select all the video content aligned with the selected audio fragment; the master video track provides a good foundation and a possible fall back for the event fragments that are not well covered by individual recordings
3. The set of editing points is defined as the union of all the time event boundaries; at each editing point select according to the following rules:
   a. If a solo is happening, then choose a video clip which features the soloist
   b. If there is no clip of the soloist available, then choose a clip of the preferred performer selected by the viewer
   c. If such a clip is not available, then choose a clip that best reflects the more encompassing event (if the solo appears within a fragment dominated by the strings, then choose a clip of the strings)
   d. If none of the above is possible, then choose randomly from the clips available or fall back on the video master
4. Apply exceptions to the above sequence of rules to avoid monotony
5. If the application of the above rules leads to fragments from the same video file being play continuously (e.g. in a long solo), then, after a period of time, cutaway briefly to a shot in the next priority level.

Though simplified, this rule set illustrates how the behaviour of the narrative system could be defined.

The following subsection outlines the narrative system that implements the narrative concepts presented above.

3.2 *MyVideos* Narrative System Overview

The *MyVideos* narrative system allows end-users to upload video recordings of a social event into a repository called the Media Vault and, once uploaded, to obtain automatically compiled movies configured to the following parameters (Figure 1):

- preferred people, instruments or events – the narration will be biased towards these selections
- style – selected songs or a video digest of the overall event
- duration – video compilation length, in minutes.

Figure 1. *MyVideos* configuration interface

The basic architecture of the narrative system consists of three main components: the media vault, the narrative logic and the user interface. It is illustrated in Figure 2.

The *vault* contains the video files recorded by the users during the concert on their handheld mobile cameras and a high quality soundtrack of the event. We also used a master camera that was locked off and recorded continuously, in accordance to the above mentioned narrative concept. In our trials, this was tripod-mounted, not moved during the performance, and framed to have a view of the entire event. The content is automatically transcoded to a standard format and automatically aligned to a common timeline, based on the audio track [Korchagin, 2010].

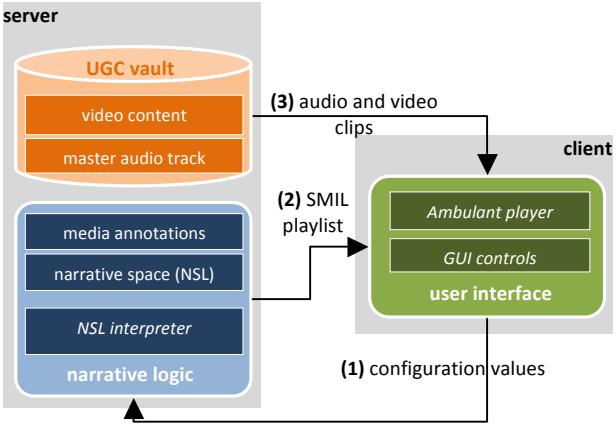

Figure 2. *MyVideos*: simplified narrative system architecture

The *narrative logic* component is based on the Narrative Structure Language (NSL), a declarative representation language for interactive video narratives [Ursu, 2008a]. The *narrative space*, expressed in NSL, embeds the narrative structures (including editing rules) that define any possible configured narration, as illustrated in the previous section. It is decoupled from the actual media, as the content is referred to in the narrative space via its characteristics captured in the *media annotations* component, thus supporting dynamic repositories of content. The narrative space is created by a professional author, who also annotates the master audio track to define the event structure that underpins the narration. The master video track, if used, can be annotated by the same person.

The individual video clips are annotated semi-automatically. As they are aligned in time with the master audio track, all the event annotations of the master that overlap a particular video can immediately be propagate to it. A further automatic annotation is carried out by a process that can identify stylistic (e.g. shot type and camera movement) and quality (e.g. light) characteristics. This can also be used to filter out poor material (e.g. shaking camera or insufficient light). The originator of the footage is an important annotation that is easily done at upload time. Richer annotations at the semantic level, e.g., describing which person or instrument is in a close-up shot, are very important for automatic narration. They are carried out manually. Nevertheless, we have devised, in *MyVideos*, different ways of querying and clustering the content in the vault, based on existing annotations, in order to facilitate and stimulate their further provision. We have described these mechanisms elsewhere [Cesar, 2010].

Once the vault and the narrative logic have been set up, the system is ready to create personalised narrations. The approach we took was that of narratives configured at the outset. The configuration parameters are sent from the interface to the *NSL interpreter*, which then produces the corresponding narration. Narrations compiled by the NSL interpreter are represented in the declarative language SMIL [Bulterman, 2008] and returned to the client. A SMIL player (*Ambulant*) fetches the corresponding content from the vault and renders it on the client device. The delay between sending the configuration values and receiving the rendered video narration is a few seconds.

In depth technical details are provided in the following section.

4. IMPLEMENTATION

This section describes the technical details of the three key processes behind automatic narration: event annotation of the master audio track, generation of the media items – the building blocks of any customized narration – and the definition of the narrative space in NSL. It also summarises the technical details of the system's deployment on the Web. Before, though, it provides a simplified description of the main NSL constructs used for creating the *MyVideos* narrative space. For more details on NSL, the reader is kindly referred to refer [Ursu, 2008a].

4.1 NSL Generalities

The basic narrative building blocks in NSL are called *media items*. A media item represents a contiguous fragment of time based media. Its characteristics are described as *annotations*, stated in OWL syntax. Any particular narration, resulting from the interpretation of a narrative space, is a layered sequence of media items. Media items can also be "empty", in which case they are called placeholders. Media items empty or not, can have attached interactive behaviour, in which case they are called *interactive objects*. Each interaction object has a number of predefined attributes, which represent the settings that could be made in the actual user interface. NSL provides a communication mechanism between interactive objects and user interfaces.

Media items are aggregated in *narrative objects* through three main narrative structures: *link*, *layer* and *selection group*. They are recursive, in that they allow narrative objects to be aggregated into more complex ones. A *narrative space* is a top-level narrative object. Individual narrations are generated by parsing the narrative space, on the basis of the configuration values, down to the level of media items. Editing decisions are taken where they naturally occur, i.e. after the selection of each particular media item.

A *link structure* is a directed graph, possibly with cycles. The nodes are narrative objects. Each arc specifies a potential path that the narrative could take from the starting to the end node. Arcs have enabling conditions, dynamically evaluated, that specify whether the respective path can or cannot be taken. The playlist to which a link structure is resolved is the sequence of the playlists to which the parsed narrative objects are resolved. A *layer structure* has a number of layers, each being a narrative object. The playlist to which a layer structure is resolved is the aggregation in parallel (i.e. to be played concurrently) of the playlists to which the objects on each layer are resolved. Alignment between narrative objects, playing on different layers, is possible via a name referencing mechanism. The *selection group* is a collection of narrative objects, from which one is selected for play via a selection rule. The selection rule refers to annotations, user interaction and context variables.

A selection rule consists of three parts: *select*, *default* and *alternative*. The select part is used to specify which objects can be played, from those available in the group. If more than one is selected, the default part is applied and this contains a disambiguation rule. If no object is selected, then the alternative part is applied to determine what to play instead. For example, the following selection rule

select(not *played*), ***default***(*Section=Orchestra*), ***alt***(*random*)

specifies that an object that has not yet been played should play; if there is more than one such object, then one containing a shot of the orchestra (annotated with the tag Orchestra for the category Section) should be selected; if all the available objects have already been played, then play a random object. A default rule or alternative rule can itself contain a combination of select, default and alternative rules. Such recursive combinations could be employed to priorities selection criteria, as required by the narrative rules specified in the previous section.

Finally, *MyVideos* employed also NSL's mechanism for explicitly maintaining context variables. Context variables can be introduced and their values could be set dynamically, i.e. at parsing time, via a *side-effect statement* embedded in temporal annotations associated with specific time-points. For example

temporal-annotation *(point(11:22:33),*

set *(CurrentSolo), {Michael, Piano}))*

When a media item with such an annotation is placed in the playlist, the context variable is updated to the specified value when the narrative time reaches the annotation time-point.

4.2 Event Annotations and Media Items

The event structure is overlaid on the master audio track through event annotations. They are expressed as NSL side-effect statements and are employed to keep the state or context of the narration continuously updated as the narration progresses. For example, they detail what is happening in the music, i.e. when instruments start and stop playing or when a solo starts or stops. The event annotations serve an additional purpose: they are points where the narrative engine should make editing decisions. The event annotations refer to

1. song: ***set****(S, name-of-song)* – indicates the beginning a song
2. instrument: ***set****(I, {list-of-instruments})* – indicates when a particular instrument or orchestra sections becomes active
3. solos: ***set****(P, {list-of-people})* – indicates when a solo is starting, but it refers to the actual performers

where *S*, *I* and *P* are context variables. End of events can be expressed by removing elements from the list or, for the former annotation, setting the name of a song to empty.

As all the content is aligned to a common timeline, enforcing editing decisions in these event points is achieved by using them as references for the definition of the media items. Any two consecutive annotation points define a media item for each video clip that intersects the respective time interval. In other words, all the annotation points are cutting points for sub-clipping all the video clips. If the distance between two consecutive such points is too large, further cutting points are inserted, on the music beat

4. ***cut-rhythm*** – indicates an appropriate cutting point with no significant event changes

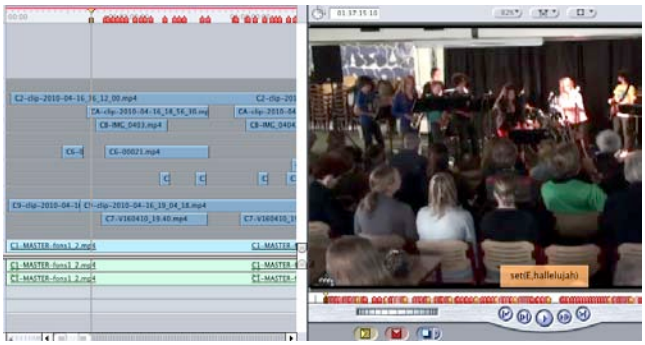

Figure 3. Screen shot of Final Cut representation and markup

For *MyVideos*, this annotation process was carried out by a video professional using FinalCut Pro (Figure 3). The annotations were marked on the common timeline, not on the audio track itself.

The annotation file exported in an XML format from FinalCut Pro is imported into *MyVideos* and, on its basis, an automatic process generates all the media items. It is possible that certain media items, namely those resulting right at the beginning or the end of the originating video clip, are too short. They are simply removed by this process.

Another automatic process propagates all the annotations associated with the (larger) media clips – such as originating camera, person or instrument in shot, or type of shot – to the media items in which they have been divided. The algorithm we implemented is simple: if a media item overlaps with an annotated interval, then it inherits that annotation. More advanced algorithms, though not required in *MyVideos*, are easy to see.

These two automatic processes produce all the basic building blocks that could be used for the automatic generation of individual narratives.

4.3 Narrative Space

This section describes the narrative structures for one style only: *selected songs*.

The narrative space consists of a sequence of two narrative objects (a link structure): a content-less selection group, called *composer*, which always selects a media item called *ComposerConfig*, followed by a selection group called *mainloop* (Figure 4).

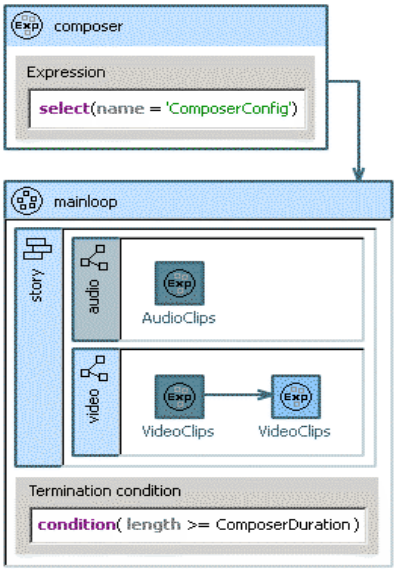

Figure 4. The *MyVideos* narrative structure in NSL

ComposerConfig is an interactive object expressed as a placeholder of duration zero. It has the mere purpose of setting the configuration values, selected at each particular viewing, in their corresponding context variables. It achieves this through a number of side-effect temporal annotations, one per configuration parameter, that transfer the value of the interactive object attribute into its corresponding context variable. The configuration context variables are constant during each compilation. They include

SelectedSongs, SelectedPersons, SelectedDuration

mainloop expresses the following narrative logic: select, in turn, all the songs chosen by the end-user and, for each song, select appropriate video content synchronized with the audio. *mainloop* consists of a two-object layer structure, called *story*. The *audio* object is the leading layer and, in turn, it is made of one selection group only, called *AudioClips*. This has the role of selecting, one by one, the songs (audio only) from the end-user's choice stored in *SelectedSongs*

select*(MediaType = Audio* and

not *played* and

Song in *SelectedSongs)*

MediaType and *Song* are annotation categories of the media items.

Disregarding, for a moment, the *video* object, this structure generates a sequence of all the selected songs. Nevertheless, this sequence could finish earlier, as there is a termination condition

condition*(length >= SelectedDuration)*

As soon as the length of the song sequence goes over the *SelectedDuration*, the compilation is terminated.

The *video* object has the role of selecting appropriate video content in sync with the audio. We have further simplified the selection criteria into:

1. Select a first 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 whose *Instruments* annotation match the instruments that are active in the audio layer, set in the context variable *I*
5. If the result consists of more clips, select those whose *Person* annotation match the persons currently playing a solo, set in the context variable *P*
6. If the result consists of more clips, choose randomly.

The first object of the *video* link structure initialises rule 1, above:

select*(MediaType = Video* and

RelativeIn = AudioClips:RelativeIn).

The second object of the video link structure is a looping selection group that implements the rules stated above:

select*(MediaType = Video* and

RelativeIn = VideoClips:last_played:RelativeOut),

default*(*

select*(Person = SelectedPersons),* ***default****(*

select*(I=Instruments),****default****(****select****(Persons=P),* ***alt****(all))),*

alt*(* ***select****(all))),*

alt*(* ***select****(all)))*

It is important to note that the same taxonomy should be used in annotating both the master audio track and the video content, as well as providing the end-user's selections.

This narrative space was authored using the authoring toolkit of the ShapeShifting Media [Ursu, 2008a].

The narrative space described here is entirely decoupled from the actual media, thus it can function as a computational format that could be applied to any collection of content recorded in a school concert, obviously, annotated using the same taxonomy.

4.4 Deployment as Web Application

The *MyVideos* application has been implemented as a fully-fledged Web application. On the server side there are four main components:

- A Mongrel Web Application Server for the Ruby on Rails (RoR)[1] Application, the role of which is to present the user interface and manage user interactions
- A MySQL Database that stores the taxonomy, the media object descriptions and annotations
- The narrative engine (see previous sections), a bytecode compiled Prolog-based interpreter for NSL wrapped in a Qt/C++ application; it is invoked by the RoR Application via an HTTP call
- A Video Server that stores the recorded video clips and delivers them via HTTP video streaming.

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 is implemented using JavaScript and AJAX (*Asynchronous JavaScript and XML*). By using AJAX, partial updates of the user interfaces are performed without having to reload the entire page. 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 video clip playback, two different solutions have been used. When supported by the browser, HTML5 video elements have been used (e.g., for an iPad implementation). Otherwise, the JW player JavaScript API allows control of an embedded Flash player.

5. EVALUATIONS AND RESULTS

In order to answer the research questions 2 and 3 set in the introduction – do the tools incorporate valid filming concepts? do the end-users like the results? do the tools produce professional-like stories? – we conducted two evaluations. One consisted of interviews with a number of video broadcast professionals from UK. The other one was a long term user trial carried out with families from The Netherlands. Our findings are reported in this section.

5.1 Video Professionals Perspective

A series of interviews were conducted with three video broadcast professionals from UK: a *director* of current affairs TV programming with over 20 years experience, a broadcast TV *editor* with similar levels of experience and a corporate video editor and *producer* of 10 years experience. The editor and producer had had in the past experience of dealing with a significant amount of amateur/user-generated footage. The director, as a TV journalist, had expertise in assembling stories from content representing social events and was, therefore, also interested in processing own material in a manner similar to *MyVideos*. The interviews followed a basic structure of:

1. Interviewees were asked how they would envisage filming and editing stories with multiple cameras from family and social group events.
2. They then had explained to them the specific scenario of the School Concert and shown the *MyVideos* interface.

[1] The technologies mentioned in this section, if unknown, could very easily be identified via a simple Internet search, therefore they will not be web-referenced.

3. Finally, the method of footage preparation and timeline annotation was presented to them in a FinalCut system and their opinion sought on how effectively they thought the rules could function. They were not shown automatically compiled edits.

All the subjects expressed a belief in the validity of the model both in terms of the demand for and use of a *MyVideos* style configurable compilations which could be reedited, and also the model of annotation and structures used to generate the compilations.

The editor referred to his own experiences of working for friends, for example on wedding videos, and how it was an iterative process – *"you never really know what's in the mind of a parent"*. Inevitable there would be a process of re-authoring according to the participants wishes after a first cut, for example, to remove an offending individual. In response to the idea of automatic edits that were reconfigurable after the fact – *"I would absolutely feel people would like that. Not many people have decent editing equipment at home, but loads of people shoot stuff… and most of it just sits there"*. The corporate video producer added that *"A parent would never miss their own child"*, but "*having a library from which your particular type of edit could be generated could be attractive"*. The director, as a TV professional, saw it only as a niche market to give the people the ability to edit the material themselves, which he contrasted with the ability to automatically generate good quality edits.

When it came to the manner in which the footage was prepared and annotated on a timeline all the respondents expressed a similar instinctive approach to the one we had taken. Regardless of their different levels of knowledge in the relevant tools (the director would be more used to working with professional editors with technical expertise), when asked to imagine themselves dealing with all this user-generated footage in a professional context, they all wanted to have the ability to think in terms of time, timeline and if possible a master audio track – the director: *"If it's possible to download it onto a file or something that has timecode… that's the key thing…You need somehow the ability to recognize when things happen"*. The producer: *"Could I synchronise the time code across the cameras… I would lay the mastertrack down first*" and most importantly*, "for any concert or music video the song's your basic structure - I would have all of my categorising and prioritisation decisions based around the audio."* So when attempting to produce a best edit, all of the interviewees said they would try and construct the footage into what is broadly referred to as a 'multi-cam' shoot, in which footage is setup in parallel and cameras are chosen according to events that occur on the master timeline, in this case the audio. This matched the approach taken in the *MyVideos* automatic authoring system. When asked to come with up basic rules that could be applied their thoughts were similar to those we attempted to set out in our narrative concept.

When asked whether they see any types of annotations as more important than others, they all expressed more of an interest in the ability to have knowledge of shot type and quality, rather than of which individuals were in the shot – *"Something that could distinguish between shot sizes, I could see that being useful… something that makes sense to you as you're editing"* – and shot quality, for example – *"If I had 9 cameras… then I would look through quickly to see which one had been shot well… you can tell pretty quickly…"*. They expressed less interest in the ability to explore the media clips according to the people and instruments in the shot. This is an interesting viewpoint, as it seems to suggest that video quality and syntax choices happen before semantic choices, but obviously, this is only in the context of synchronized footage. Furthermore, quality and shot type are attributes easier to extract automatically than people and instruments. It was also suggested that quality measures may not necessarily need to be applied per clip, but rather at the level of cameras, as there is not that much variation within the content originating from one camera. Quality measures could then be built into the narrative rule structures via camera choices.

However, it is the reliable detection of shot type or size that would represent the best step towards improving annotation for the purpose of automatic authoring, and also the best means to start introducing the concept of a style choice, as well. All the interviewees, and especially the editor, made the point of how varying the shot size would be one of their priorities to create attractive editing. He wanted to avoid jump cuts and the best way of doing this was being able to know which types of shots were available - *"Going from a wider shot to a closer one looks pretty"* as oppose to jumping between medium shots. However, he acknowledged that a lot of modern editing actively employs jump cuts as stylistic device, so, from the automatic authoring perspective, a reliable awareness of the shot type would allow it to have rules that could enforce a conventional styles of editing or move towards a modern style of editing, depending on the preferences of the end-users.

The editor, when discussing a basic set of editing rules, said he would start with a wide and *"prioritise mediums and close-ups, people on the stage… interspersed with every 5-6 shots with a reverse of the audience"*. Shots of the audience, rather than taken by the audience, is an element in the overall story telling that we did not take into account in the *MyVideos* test bed. We concentrated on the concert itself, and of the participants rather than the viewers. It is important to note that, in these shared, group experience, the viewers are very much part of the experience as well. When the TV director/journalist was asked what would be a way he would delve into the story of the concert he said, *"I would give a camera to at least a couple of the parents, and I would film the parents filming the concert… because after all it's the relationship between the child and the parents that are important here…"*. This is an important clue to when it comes to compiling the best stories: it is not just the video clips of the performers that might be important, and even the best edit of the concert itself might not perfectly suit the goals of building community and shared experiences.

5.2 End-Users Perspective

A selected set of parents from a high school in Amsterdam has been actively collaborating with this research. In December 2009, the parents were invited to a focus group that took place in Amsterdam. In April 2010 they recorded (together with friends, relatives and some researchers) a concert of their children – the Big Band. The parents that took part in the recordings provided the research team with all the material: in total around 210 media objects were collected from a concert lasting about 1h and 35 minutes. Twelve cameras were used: ten were mobile, two were fixed master cameras.

Seven people have subsequently been recruited for a long term user study. Primarily, they were relatives and friends of the children that performed in the high school concert in Amsterdam. The participants were of a variety of ages, and all of them of a

quite high socio-economic status. Participants' occupations included student, social scientist, software engineer, art designer and visual artist. All the trialists were Dutch. The average age of the participants was 37.1 years (SD = 20.6 years); 3 participants (42.8%) were female. Three of the participants were students (undergraduate). Among our participants, 3 had children (ranging in age from 14 to 17 years). All participants were currently living in the Netherlands, apart from an uncle, who lives in the US (the only one that was not present in the concert recordings).

Participants kindly volunteered themselves for their participation, and the experiments were conducted over a two–month span in the summer of 2010 (Jul-Sep). We are aware that we have a small sample of only 7 people in our study, from which it is difficult to draw conclusive results. Nevertheless, we got a strong sense that this group provides a deep understanding of the ways in which people currently edit and share personal videos. First, because we involved three sets of parents who recorded their kids in the concert who had been contributors to a previous focus group during a 10 month period dating from December 2009 [Cesar, 2010]. Furthermore, we should emphasize that it is a quite arduous work, given the time and personal constraints, to find and involve people, get them to record even a small concert, process all the recorded video materials, and finally have these participants experiencing the actual system. The goals of the study make it almost impossible to do crowd-source testing, since users of the system should be people that care about the actual content of the videos.

We followed a semi-structured approach for data collection, including in person interviews (each trialist individually), and interaction with the *MyVideos* Web application. In this section we report only on the users' experiences related to the automatic generation of video compilations.

All our participants reported they record videos in social events, such as family gatherings, vacation trips. However, most of them said they barely look at the recorded material afterwards. For most of the trialists, video editing is time consuming and way too complicated. Nevertheless, some are familiar with video editing tools such as iMovie and Windows MovieMaker. One mother reported she carries out minor editing tasks on a device (e.g. clipping) because her personal computer cannot handle the video material efficiently.

Regarding their interaction with the *MyVideos* system, we conducted an exploratory study. During this study our participants were not given specific assignments, but they were instructed to interact with the system and we asked questions and investigated opinions as they were involved in the interactions. As mentioned before, in general our trialists had a cinematographic and utilitarian view of the system. As we expected, they stated that they are not willing to spend a lot of time authoring videos.

Our participants did appreciate the generated compilations. Overall, the composed videos were considered *visually compelling*, and some participants *wanted to share them straight away.*

The following conversation during the user trial is illustrative for the general reaction of the participants to the results provided by the automatic narrative engine. Before using the system: *"I am expecting to be very pleased".* When watching the compilation*: "…I did not mention the drummer… you will not see him. I guess… but I did ask for the bass… I am very happy about it…"* Afterwards, discussing the system*: "It is an enjoyable product…It feels like a human has made it!"* (Friend of a performer)

When prompted, our participants, mainly the teenagers, also described that it would be interesting to have an *exclusion filter*, which would allow the removal of a specific person from the video compilation. However, such functionality was not seen as critical for using the system. *"You only want to have your friends... if you had a fight with somebody, and you don't want to remind your parents about that, when you create a video of the concert you wouldn't have this person in the whole video at all."* (Performer)

In general, the automatic authoring tool was seen as an effortless way of getting semi-professional movies made. Users expressed their interest on the possibility of fine tuning and personalizing the generated production afterwards. *"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) This opens up a direction for future work.

6. CONCLUSION AND FUTURE WORK

We set out to build a narrative system, *MyVideos*, whose main functionality was the ability to automatically edit personalized video compilations (movies) from shared repositories of recordings of social events. Our quest was motivated by peoples' inability and lack of interest in video editing, which, in turn, causes significant amounts of video recordings to remain abandoned on memory cards or as files on hard drives never to be accessed again. The research questions we asked in this context were: Could such a system be built with current technology? Could professional video narrative techniques be automated by such a system? Do end-users like the automatically produced video narrations?

Focusing the scope of the questions to a specific type of social event – a school concert – we found positive answers to all our research questions. On the basis of Narrative Structure Language (NSL) and ShapeShifting Media, we successfully implemented a prototype narrative system, *MyVideos*, that is able to generate video compilations customized to the preferences of the end-users. Our narrative approach, conceptual and implemented in the prototype, has been validated by video broadcast professionals from the UK. The prototype and automatically generated video compilations have also been assessed positively in a long term user trail carried out with families from The Netherlands.

Encouraged by these positive results we have a number of further developments we plan to explore, which include:

- understanding the response of the end-users to various levels of complexity and sophistication of narrative spaces
- investigating other potential applications for this paradigm, in areas or scenarios that involve user generated content
- providing better configuration means, e.g. allowing parts of the edited video to be kept or rejected, similarly to a "fruit machine" behaviour, or supporting real-time interaction
- exploring a better integrated hybrid approach in which automatic and manual processes complement and stimulate each other (e.g. use manual editing for fine tuning and automatic editing to stimulate manual annotation)
- integrating such functionality with social media platforms

Finally, it is worth noting a "side effect" of this research: the validation of ShapeShifting Media in a UGC context.

7. ACKNOWLEDGMENTS

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. Special thanks to all the partners involved in the *MyVideos* demonstrator.

8. REFERENCES

[1] Bulterman, Dick C. A. and Rutledge, Lloyd W. 2009. *SMIL 3.0: Flexible Multimedia for Web, Mobile Devices and Daisy Talking Books*. Springer Publishing Company, Incorporated, ISBN: 978-3-540-78546-0.

[2] Bocconi, S., Nack, F. and Hardman, L. 2008. Automatic generation of matter-of-opinion video documentaries. *Special Issue on Semantic Multimedia, Journal of Web Semantics* (JWS), 6(2), pp. 139 – 150. DOI=http://doi.acm.org/10.1145/1026633.1026636

[3] Cesar, P., Bulterman, Dick C. A., Guimarães, R. L. and Kegel, I. 2010. Web-Mediated Communication: in Search of Togetherness. In *Proceedings of the Web Science Conference* (WebSci10), Raleigh, North Carolina, USA.

[4] Cavazza, M., Champagnat, R., Leonardi, R. and the IRIS Consortium, 2009. The IRIS Network of Excellence: Future Directions in Interactive Storytelling. In *Proceedings of the Second International Conference on Interactive Digital Storytelling* (ICIDS 2009), Springer LNCS 5915, pp. 8-13. DOI=http://dx.doi.org/10.1007/978-3-642-10643-9_4

[5] Ibanez, J., Aylett, R., Delgado-Mata, C. and Molinuevo, E. 2009. On the Implications of the Virtual Storyteller's Point of View. *The Knowledge Engineering Review*, 23(4):339-367. DOI= http://dx.doi.org/10.1017/S0269888908000039

[6] Kennedy, L. and Naaman, M. 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=http://doi.acm.org/10.1145/1526709.1526752

[7] Korchagin, D., Garner, P. N. and Dines, J. 2010. Automatic Temporal Alignment of AV Data with Confidence Estimation. In *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing*, Dallas, USA, 2010

[8] Obrador, P., de Oliveira, R. and Oliver, N. 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=http://doi.acm.org/10.1145/1873951.1874025

[9] Porteous, J., Benini S., Canini L., Charles, F., Cavazza, M. and Leonardi, R. 2010. Interactive Storytelling via Video Content Recombination. In *Proceedings of the 18th ACM International Conference on Multimedia* (Short Papers), Firenze, Italy, October 25-29, 2010. DOI=http://dx.doi.org/10.1145/1873951.1874334

[10] Riedl, M. and Young, R. M. 2006. From Linear Story Generation to Branching Story Graphs, *IEEE Journal of Computer Graphics and Applications*, May/June: 23–31. DOI= http://doi.ieeecomputersociety.org/10.1109/MCG.2006.56

[11] Shamma, D. A., Shaw, R., Shafton, P. L. and Liu, Y. 2007. Watch what I watch: using community activity to understand content. In *Proceedings of the international workshop on Workshop on multimedia information retrieval* (MIR '07). ACM, New York, NY. DOI=http://doi.acm.org/10.1145/1290082.1290120

[12] Shamma, D. A., Kennedy, L. and Churchill, E. F. 2009. Tweet the Debates: Understanding Community Annotation of Uncollected Sources. In *Proceedings of the first SIGMM Workshop on Social Media*, pp. 3-10. DOI= http://doi.acm.org/10.1145/1631144.1631148

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

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

[15] Ursu, M. F., Cook, J. J., Zsombori, V., Zimmer, R., Kegel, I., Williams, D., Thomas, M., Wyver, J., Mayer, H. 2007. Conceiving ShapeShifting TV: A computational language for truly-interactive TV. In *Proceedings of the 5th European Interactive TV Conference* (EuroITV'07), published as *Interactive TV: A Shared Experience, LNCS 4471*, P. Cesar, K. Chorianopoulos, and J.F. Jensen, Eds. Springer, 96–106

[16] Ursu, M. F., Kegel, I., Williams, D., Thomas, M., Mayer, H., Zsombori, V., Tuomola M. L., Larsson, H., and Wyver, J. 2008a. ShapeShifting TV – Interactive Screen Media Narratives. *Multimedia Systems Journal, Volume 14, Number 2 / July 2008*, ACM/Springer, pages 115-132. DOI=http://dx.doi.org/10.1007/s00530-008-0119-z

[17] Ursu, M.F., Thomas, M., Kegel, I., Williams, D., Tuomola, M., Lindstedt, I., Wright, T., Leurdijk, A., Zsombori, V., Sussner, J., Maystream, U., and Hall, N., 2008b, Interactive TV Narratives: Opportunities, Progress and Challenges, *ACM Transactions on Multimedia Computing, Communications and Applications*, 4 (4) : Article 25, pp. 25:1–25:39.

[18] Ursu, M.F., Zsombori, V., Wyver, J., Conrad, L., Kegel, I., and Williams, D., 2009. Interactive Documentaries: A Golden Age, *ACM Computers in Entertainment (CiE)*, Volume 7, Issue 3, Special Issue: TV and Video Entertainment Environments, Section: Production, Interactivity, and Narrative. September 2009

[19] Williams, D., Ursu, M. F., Cesar, P., Bergstrom, K., Kegel, I. and Meenowa, J. 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=http://doi.acm.org/10.1145/1542084.1542088