Metadata-based Adaptive Assembling of Video Clips on the Web

Rene Kaiser, Martin Umgeher, Michael Hausenblas
JOANNEUM RESEARCH Forschungsgesellschaft mbH
Institute of Information Systems & Information Management
Steyrergasse 17, 8010 Graz, Austria

E-mail: {firstname.lastname}@joanneum.at

Abstract

Video content remixing—personalised or generic—is available on the Web in various forms. Most platforms target at users uploading and rearranging content, and sharing it with the community. Although several implementations exist, to the best of our knowledge no solution uses metadata to its full extent to, e.g., dynamically render a video stream. With the research presented in this paper, we propose a new approach to dynamic video assembly.

In our approach consumers may describe the desired content using a set of domain-specific parameters. Based on the metadata the video clips are annotated, the system then chooses video clips fitting the user criteria. The video clips are aligned in an aesthetically pleasing manner while the user furthermore is able to interactively influence content selection during playback.

A fictitious showcase from the sport domain illustrates the applicability of our approach. The implementation is demonstrated using an available non-linear, interactive movie production environment.

1. Introduction

Online video portals have spread like wildfire recently, incited by large communities driven by the desire for fame, social aspects or creativity alone. Therefore these platforms are designed and developed for dedicated private persons of limited skill and time. They—at least try to—live on indirect financing and advertising rather than the value of the entertaining or educating content itself.

The common idea is that video clips are uploaded by the average user, annotated and possibly modified online, then shared with other users who actively comment on it. Video editing features offered by current implementations, as Jumpcut[1], are far from what is standard in professional post production environments. The utilities are, disregarding few notable exceptions, limited to adding rudimentary metadata, adding background audio or clipping the boundaries of the video.

The result is typically a Flash stream of limited size and quality, sufficient for online editing, and consumption, respectively. Video web sites inviting users to statically assemble their own 'mashups' from a pool of clips are already present. Some systems [12] even compose clips 'by reference' instead of storing the result as an immutable file, thus not ruling out modification of clips at a later date.

Most of these deployed applications use some kind of metadata to make huge amounts of clips searchable. While metadata is not equally available for heterogeneous video resources, many professional content providers do already possess interesting related meta information.

After decades of 'lean back' consumption of predefined material, dynamically generated multimedia content impact is ramping up [5]. Consumers growing up with sophisticated technologies, as for example interactive computer games, already want to decide themselves when and what to consume, hence exhibit a 'lean forward' attitude. Further, they start to become comfortable shaping the content themselves in various ways, both before and while watching.

In this research we propose to overcome the limitations of exiting solutions by applying metadata-based assembling of video clips, hence enabling interactive views on a material. A showcase for the realm of sport videos is described, wherein the proposed principle is examined and reflected.

The following section 2 places our research in the context of related work, and discusses existing commercial solutions. Next, section 3 describes our approach using the fictitious showcase of a basketball video platform. In section 4 we present the dedicated environment, in which the showcase is realised. Finally, we reflect our experiences with the new video assembly approach in section 5, and discuss possible future work.

[1] http://www.jumpcut.com/

2. Related Work

Creating interactive media content [9] is a challenging task, as the story world is non-linear and evolves based on the user's interaction. Research in computational support for interactive narratives [18] has focused mainly on applications where the audio-visual content is computer-generated (as games and VR environments). Its final aim is the development of virtual worlds in which stories unfold and the user is able to interact with other characters and the environments of the worlds [11], whilst achieving cognitive and affective responses as those seen in conventional narrative media such as film.

However, to date, they focus mainly on wrapping up interactions in meaningful and interesting narratives, rather than on expanding traditional linear narratives towards interactivity. They are situated in the interactive rich but narrative simple area of the interaction-narrative complexity space.

Our approach may be compared with existing work on video summarisation in particular aspects. For example, a personalized abstraction of broadcast football videos through highlighting selections was reported in [1]. Bocconi [3] researched in the documentary genre how to enable an authoring process to make material dynamically available to users, without having to edit a static final cut that would select possible informative footage.

The bottom line is, however, that current video web sites, as `http://youtube.com`, `http://cuts.com`, or `http://www.motionbox.com` are far off the requirements stemming from professional production environments.

As for basketball content, the aspect of interactivity has recently been addressed via the 'NBA Highlight Mixer'[2], a platform that offers static assembling of still images, video clips and background music. Clips may be trimmed, simple effects and transitions may be applied. Basketball has a history of attracting fans willing to spend their time on mixing highlight videos already. However, no commercial solution is known that uses detailed statistical meta information to accomplish what we describe in this paper.

3. A New Video Assembly Approach

Existing implementations of online video assembly systems are limited to an a-priori selection of content. They require the user to browse a media repository and select a distinct sequence of video clips, which in turn are aligned by the rendering system. This approach works well for small collections of video clips and demonstrates the potential of this kind of application. However, it is obvious that manual selection does not scale well; especially considering the huge amount of video content available for certain domains.

In this work, we propose a different approach to dynamic video assembly that works for online scenarios but can be implemented in other architectures as well (desktop application, TV broadcast, mobile solutions). Instead of forcing the users to manually select a set of video clips, they describe the desired content of the result video using a set of domain-specific parameters; then, the system tries to find appropriate content and aligns it in an aesthetically pleasing manner. While consuming the result video, the user is able to further influence the presented content by interacting with the system. Where applicable, feedback information is interpreted to acquire qualitative rating data of the content.

The following describes the requirements for video and metadata input and explains the concepts of content assembly and interaction in detail. Section 3.5 then illustrates the approach with the aid of a virtual, though practical example.

3.1. Content Requirements

To allow satisfying results from an artistic point of view, we require the used material to be of consistent visual quality. If remixing an already edited movie, the material might contain editing artefacts (e.g., text overlays, cross-fades) that reduce its reusability. It is thus preferable to use the raw, unedited material where available.

The approach disregards if the clips are imported as separate video files or contained in one or more larger video files. In the latter case, the boundaries of the atomic clips have to be defined through facts extracted from metadata, or through video analysis [7].

3.2. Metadata Requirements

Metadata describing the content may be imported from different sources, however, a unified and interoperable handling must be ensured to enable sound and powerful queries on top of it. Its quantity, granularity and accuracy is fundamental. Too few, incomplete or wrong annotations directly influence the content selection and lead to less accurate output as they correlate with the users' preferences.

Multimedia metadata is available in diverse formats [6]. While manual annotation of content yields good results, the automated extraction of metadata is feasible, hence desirable [2]. Yet another approach for gathering the metadata directly from users is made possible through the so called 'human computation' approach [15].

[2] http://myvideo.nba.com (see also: http://www.gotuit.com)

3.3. Interaction

We define interaction as user input that subsequently influences content selection. A-priori stated preferences are necessary to generally define what the system should play, e.g., certain domain-specific characteristics as well as the approximate playback duration. Even during consumption, the user can interact with the host environment and is able to dynamically prompt it to provide content fitting the given criteria. Further, actions suggesting relevance feedback of the content are interpreted and added to the metadata.

3.4. Content selection & assembly

The key to our approach is the actual assembly of suitable video clips into a continuous stream. To enable interactive manipulation of the result, the clip sequence is not calculated up-front; instead, the system selects clip after clip based on comprehensive selection rules. User criteria do not likely correspond directly to metadata concepts, but are an abstraction thereof. Combining global rules with user criteria, the system uses domain-specific semantics to find suitable clips in the repository and chooses the clip that fits user criteria best, or a random one of equivalent candidates.

There are many use-cases for this approach, e.g., remixing of independent clips, generation of classic summaries, or generation of different versions of the same content. If prompted to generate a classic summary for example, the system will thus consider the remaining playback time and determine a set of most relevant unplayed clips and continue playing them in chronological order, unless the consumer's preferences change. In general, selection rules take into account:

- Global user criteria & preferences,
- User interaction,
- What has already been played, and
- Aesthetic rules and effects (e.g., fade transitions, concatenation rules based on camera movement);

Just like content selection depends on domain knowledge, rendering & transmission of the video is limited by the and the consumers' end devices.

3.5. Show Case: Online Basketball Video Assembly

To demonstrate our approach, we elaborate on a fictitious example application that allows to create highlight and summary videos of NBA basketball games. The NBA is a basketball and entertainment league with vast financial impact. For years, it has realized the commercial appeal of extended video coverage on various channels. Besides its presence on TV, footage is available on the web[3]; even entire games are broadcast via broadband. Basketball content is both spectacular and multifaceted, and therefore well suited for interactive consumption. Assuming to be a broadcast partner of the league, we have access to both game video material and high-quality metadata. Given the existence of this data, these requirements seem reasonable and realistic. Further, we feel the scenario has a considerable commercial potential in combination with personalized online advertisement.

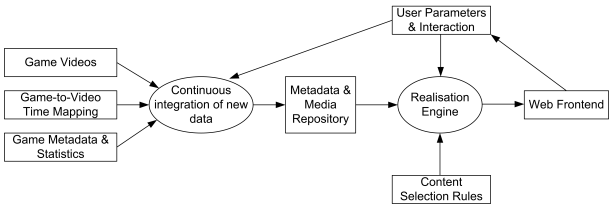

Figure 1. Showcase setup.

The setup of the showcase is depicted in Fig. 1. Raw videos are acquired along with metadata and imported into the host environment. The realisation engine is triggered by the users' input and uses inherent selection rules to choose fitting video sequences based on their preferences. Resulting video is streamed to the user via the system's frontend, a web interface. There is general consensus which perspectives are best suited for watching different actions, so the system would consider that when choosing from equivalent material.

Due to the complexity of present metadata, a wide range of use-cases can be implemented. This is realized by mapping single aspects of the users' preferences to low-level metadata concepts using domain knowledge when choosing clips.

If a video of rather short duration is requested from within a specific game, only the most important highlights will be shown. By increasing the length to the maximum, coverage may be extended to the full game. Moreover, the system can provide highlight compilations, e.g., showing 3-point shots from exceptional distance or successful defensive possessions resulting with a steal or block, probably limited to a season, a team or a certain player.

Metadata

Almost every aspect of an NBA basketball game is covered by exhaustive statistics. This includes statistics about teams, players (averages, career bests), and games (all game events by exact time, involved players, and action, e.g., free throws or turnovers) including extensive game logs[4].

[3] http://www.youtube.com/nba, http://www.nba.com/video

[4] e.g., http://sports.espn.go.com/nba/playbyplay?gameId=260122013

Comprehensive statistics—both official NBA statistics and further analysis—are publicly available on different web sites[5]. In this example, we acquire and integrate them directly from those pages or via web services.

Due to this distributed nature of the data, different data sources are integrated into one consistent relational data model (knowledge graph) which then can be queried to find relevant game events fitting user criteria.

If necessary or desired, selection of relevant clips (highlights) can be improved through video analysis. Various approaches elaborate on that, e.g., automatic extraction of shooting position as presented in [13] or the identification and tracking of the basketball as in [17].

Content

NBA games are covered extensively by multiple cameras from different perspectives. High resolution cameras are used, however, the realisation engine does not need to deal with full quality data for this online system; therefore, it is reasonable to downsample beforehand.

As seen in overlays on TV the actual game clock is available accurately for broadcasters. We require this information to be persisted and available for our system as we don't consider a posteriori extraction a reliable alternative. This information is later used to map the actual game time[6] to the video material's time codes. In this example, correct mapping information is essential for the system to operate.

While metadata is used to compute when relevant events occur, there is no exact information when an interesting sequence actually starts and ends. This has to be approximated by general rules taking into account the type of action. If results are not satisfying, this may be subject to further improvement by video analysis. Finally resulting is a set of usable video clips that can be further used.

Interaction

Relevant events are defined by the selection rules and may include spectacular actions like slam dunks, game deciding sequences like winning baskets, or clips showing the achievement of personal records. The system allows the user to describe the desired content based on the following criteria:

- Approximate duration of the result
- Preferred camera perspective selection behaviour
- Favourite play actions (e.g., slam dunks, 3-point shots, blocks, ...) or combinations thereof
- Preferred actors (teams or players)
- Pool of games: specific game (as for summaries), filter e.g., by teams, seasons, game type (regular season, playoffs, special events)

While consuming the result, the following interactions are available to the user:

- Replay current clip
- Skip current clip
- Switch camera perspective
- Show similar clips (e.g., with same player, same play action)

The feedback retrieved through user interaction is used to rate the content: skipping a clip reduces its rating, while replaying increases it. For subsequent users, the system will preferably select well-rated clips while omitting unpopular ones if possible.

4. Implementation

To show the applicability of our approach, we discuss a potential implementation of the backend in a dedicated environment: The "New Media for a New Millennium" (NM2 System).

The NM2 project aims at developing tools for the media industry that enable the efficient production of non-linear, interactive broadband media. Additionally to the production values and aesthetic pleasures of television and cinema, productions based on NM2 technologies are influenced through the interaction of the user according to their personal preferences.

To provide non-linearity, NM2 productions are not final edited pieces of media, rather they consist of a pool of small media units to be recombined at run-time.

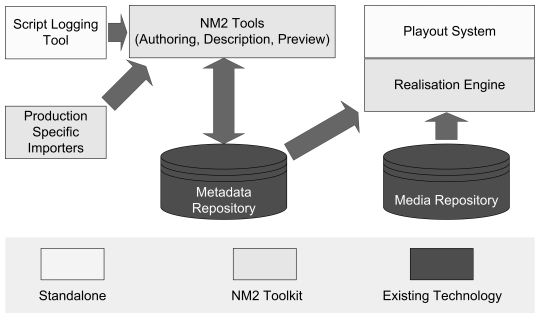

Figure 2. The NM2 System Architecture.

As depicted in Fig. 2 the NM2 system can be divided into the main areas of functionality:

[5] e.g., http://www.nba.com/statistics, http://82games.com, http://databasebasketball.com

[6] An NBA game consists of four quarters of twelve minutes each plus overtimes, but the raw material will also contain game breaks.

- The *NM2 Tools* (cf. Fig. 3 for a screenshot) enable the creation of non-linear stories; they cover the ingestion of essence (Script Logging Tool and Importers), the description of video clips, as well as the authoring, viz. the construction of possible stories.
- The *delivery system* comprises the Realisation Engine and the Playout System, respectively. The Realisation Engine is responsible for dynamically creating a playlist, based on the (author-defined) story world, and the interaction of a particular user. The Playout System takes care for rendering the actual playlist on a client device.

  Previewing in the NM2 Tools is implemented using an instance of the Realisation Engine along with a generic interaction client.
- The *Media Repository* and the *Metadata Repository*. Obviously, a media repository is needed to manage the media assets. The Metadata Repository households the low-level metadata, the production ontology, including the story world.

  Only the NM2 Tools have write access to the Metadata Repository. The Realisation Engine interprets the story world by transforming the ontological descriptions into a Prolog program. Further details are explained below.

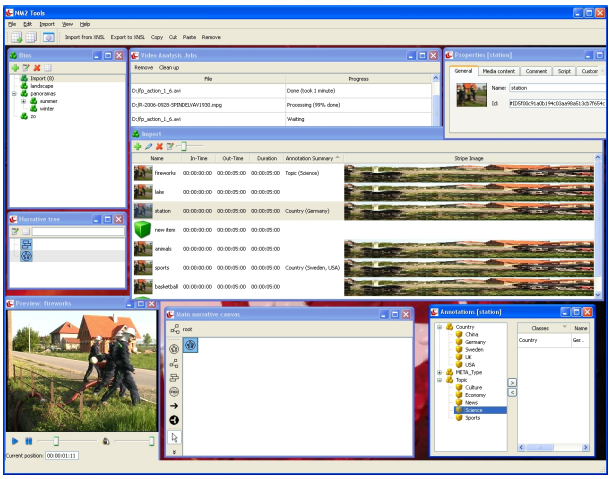

Figure 3. The NM2 Tools.

In NM2 a dynamic matching of appropriate video clips based on an expression describing the desired content is performed. In the setup discussed in this work, a remix is assembled on-the-fly. To realise the matching, a set of existing technologies is utilised, as described in the following.

MPEG-7 [8] is utilised for representing low-level features of the video clips, such as, e.g., shot boundaries; this process is typically realised using the Multimedia Mining Toolbox [2, Section 6].

OWL-DL [10] is used to formalise the domain semantics in terms of logical entities and functions as the interface to the Narrative Structure Language [14]. A *logical entity* is anything occurring in a video clip, either directly observable, or not, as a mood, or other abstract concepts [4].

The NM2 system has successfully been used and evaluated in six media productions. They are suited to a range of cross media publishing channels, including broadcast (television), broadband delivery, and DVD. The productions cover several genres, including drama, fiction, news, and a documentary. For a detailed overview on the NM2 project objectives, system capabilities and the productions, the reader is referred to [16].

5. Conclusions and Future Work

In this paper we have presented a new approach for personalised assembly of video content in an online environment. We further showed, how the fictitious showcase—a basketball remixer—can be realised using an existing non-linear movie production system, we developed over the past three years. Although the approach may not be applicable for other genres, such as drama or documentaries, our showcase illustrates the suitability for interactive sports content.

Future work may include support for other use cases, as for example assistance for defensive strategy analysis. There, the system could assemble all video clips where opposing player A scored from a specific area while another opposing player B was also involved in the play and credited an assist. Also, the already available metadata can be used to display the current score, statistics or any description of special events.

While available metadata seems to be sufficient for the basketball example, automated metadata extraction through video analysis can be used to enrich or prove existing information in other domains.

6. Acknowledgements

The work presented herein has been partially funded under the 6th Framework Programme of the European Union within the IST project “New Media for a New Millennium (NM2)”. The authors would like to gratefully acknowledge Werner Bailer for his support, discussion, and feedback on various revisions of this paper.

In memory of Dr. Jonathan James Cook.

References

[1] N. Babaguchi, Y. Kawai, T. Ogura, and T. Kitahashi. Personalized abstraction of broadcasted american football video by highlight selection. *IEEE Transactions on Multimedia*, 6(4):575–586, 2004.

[2] W. Bailer, P. Schallauer, M. Hausenblas, and G. Thallinger. MPEG-7 Based Description Infrastructure for an Audiovisual Content Analysis and Retrieval System. In *Proceedings of SPIE - Storage and Retrieval Methods and Applications for Multimedia*, volume 5682, pages 284–295, 2005.

[3] S. Bocconi. *Vox Populi: generating video documentaries from semantically annotated media repositories*. PhD thesis, Technische Universiteit Eindhoven, 2006.

[4] M. Hausenblas. Applying Media Semantics Mapping in a Non-linear, Interactive Movie Production Environment. In *1^{st} International Conference on New Media Technology (I-Media '07)*, Graz, Austria, 2007.

[5] M. Hausenblas and F. Nack. Interactivity = Reflective Expressiveness. *IEEE MultiMedia*, 14(2):1–7, 2007.

[6] M. Hausenblas, R. Troncy, C. Halaschek-Wiener, T. Bürger, and O. Celma. Multimedia Semantics on the Web: Vocabularies. W3C Incubator Group Report, W3C Multimedia Semantics Incubator Group, 2007.

[7] S. Liu, M. Xu, H. Yi, L.-T. Chia, and D. Rajan. Multimodal semantic analysis and annotation for basketball video. *EURASIP Journal on Applied Signal Processing*, pages Article ID 32135, 13 pages, 2006. doi:10.1155/ASP/2006/32135.

[8] MPEG-7. Multimedia Content Description Interface. Standard No. ISO/IEC 15938, 2001.

[9] J. H. Murray. *Hamlet on the Holodeck: The Future of Narrative in Cyberspace*. The Free Press, 1997.

[10] OWL. Web Ontology Language Reference. W3C Recommendation, 10 February 2004.

[11] M. Riedl and R. Young. From Linear Story Generation to Branching Story Graphs. *IEEE Journal of Computer Graphics and Applications*, pages 23–31, 2006.

[12] R. Shaw and P. Schmitz. Community annotation and remix: a research platform and pilot deployment. In *HCM '06: Proceedings of the 1st ACM international workshop on Human-centered multimedia*, pages 89–98, New York, NY, USA, 2006. ACM Press.

[13] M. Tien, H. Chen, Y. Chen, M. Hsiao, and S. Lee. Shot classification of basketball videos and its application in shooting position extraction. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007)*, pages I–1085 – I–1088, 2007.

[14] M. F. Ursu and J. Cook. D5.3: Languages for the representation of visual narratives. Deliverable to EC (permission required), NM2 consortium, 2005.

[15] L. von Ahn. Games with a Purpose. *Computer*, 39(6):92–94, 2006.

[16] D. Williams, M. Ursu, J. Cook, V. Zsombori, M. Engler, and I. Kegel. ShapeShifted TV – Enabling Multi-Sequential Narrative Productions for Delivery over Broadband. In *The 2^{nd} IET Multimedia Conference, 29-30 November 2006*. ACM Press, 2006.

[17] L. Wu, X. Meng, X. Liu, and S. Chen. A new method of object segmentation in the basketball videos. In *Proceedings of the 18th International Conference on Pattern Recognition (ICPR 2006)*, pages 319–322, 2006.

[18] M. Young. Notes on the Use of Plan Structures in the Creation of Interactive Plot. In *In the Working Notes of the AAAI Fall Symposium on Narrative Intelligence*, 1999.