Semantic-Aware Automatic Video Editing

Stefano Bocconi
CWI
P.O. Box 94079
1090 GB Amsterdam, The Netherlands
Firstname.Lastname@cwi.nl

ABSTRACT

One of the challenges of multimedia applications is to provide user-tailored access to information encoded in different media. Particularly, previous research has not yet fully explored how to automatically compose different video segments according to a communicative goal. We propose a rhetoric-based method to support the selection and automatic editing of user-requested content from video footage. The method is applied to the domain of video documentaries to create biased sequences about a user selected subject.

Categories and Subject Descriptors: H.5.1 Multimedia Information Systems: Video, I.7.2 Document Preparation:Hypertext/hypermedia, Multi/mixed media

General Terms: Design, Experimentation, Human Factors.

Keywords: Media semantics, media rhetoric, automated video editing, video documentaries.

1. RESEARCH GOAL

A potential way of presenting the user with relevant multimedia content in an engaging way is to operate on the level of the semantics of the media source. Describing media semantics in a formal way is a requirement for such applications. The description should not only allow retrieval of relevant content, but also support the meaningful structuring of it. In other words, the application needs to know not just what a single data item is, but also with what and how to combine it to satisfy the user's request for information.

Our research goal is to investigate the media composition mechanism, i.e. how to combine media items together and present them to the user, knowing what the conveyed meaning of the combination will be. This requires understanding (part of) the semantics of the source and understanding of how these semantics change when more items are combined. To support this high-level research goal, we investigate the following two main points: the *formal modeling* of the semantics in the source domain and the *process* supporting the repurposing of the original material according to the user request. With repurposing we mean that the application selects existing content and structures it to create a coherent message, according to the user request.

At the current stage of our research we are investigating the domain of video documentary, in particular documentaries presenting historical or contemporary footage together with interviews about the subject. The original footage of a video documentary must be edited to fit a certain playing time, resulting in a loss of informative content. An application that allows the user to drive the selection, potentially makes all the content available.

The paper is structured as follows: in Section 2 we describe how previous work has addressed the topic of this research, in Section 3 we explain to what extent we have provided a solution to the two above-mentioned points and in Section 4 we discuss possible future research directions.

MM'04, October 10-16, 2004, New York, New York, USA.
ACM 1-58113-893-8/04/0010.

2. RELATED WORK

We examine related work based on the two main points defined in the previous section:

1. formal modeling of the source semantics for content retrieval
2. (a) semantic-aware structuring of video segments
   (b) automatic video editing

Media Streams [3] was developed for annotating (with iconic visual language), retrieving and automatically assembling digital video. The limitation of Media Streams is, however, that system has no awareness that showing two video segments in sequence suggests a continuity and creates a short story. Therefore, Media Streams can perform 1 and 2b but not 2a.

Other systems like ConText [2] and VOD [4] adopt the documentary form. Since the research focus of these projects was more on the syntactic aspects of the material for its automated presentation, these systems can do 1 and 2b but little of 2a.

AUTEUR [6] automatically edits video sequence with the aim of representing humorous scenes on a slapstick level. AUTEUR can perform automatic video generation through the use of domain depended rules applied to semantic annotation structures. Though AUTEUR fulfills all the above-mentioned characteristics, it is based on an ad hoc closed implementation, while we aim for the repurposing of existing video material, i.e. media content that was not created to support our process.

Very related to our research direction is Terminal Time [5], which is a system that creates in real-time biased historical documentaries based on user input. Terminal Time, too, fulfills all the above-mentioned characteristics. It is, however, as other knowledge intensive AI applications from the nineties, closed with respect to the rules and the material used.

3. CURRENT STATUS OF OUR RESEARCH

The review of related work shows that there is a lack of mechanisms that can achieve what we call "semantic-aware structuring of video segments" and are data-driven, i.e. can be applied to existing video footage, without the need of video footage to be shot explicitly to support the mechanism.

Rules (or the logic) governing the structuring should be explicitly stated and easily modifiable, and it should be possible to add new media material to the system without having to modify the process.

We tested our approach with Vox Populi [1], which is a system that generates video sequences with a bias for a particular opinion. Vox Populi uses a repository of video interviews with United States residents about the 11th of September terrorist attack and its consequences.

The formal modeling of the domain semantics (the first of our research goals) is represented by two levels of time-coded annotations. The first level describes factual data as the location of the interview, the time, the interviewee's data (gender, age, but also education, profession) and is analogous to the annotations used in Media Streams. The second level describes the rhetoric used by the interviewee in making statements during the interview (the audio track), i.e. it describes formally whether the interviewee is pro, against or undecided over a certain issue and the arguments the interviewee uses to make her point. We use the following terms: **Logos**: appeals to logic or reason, **Ethos**: appeals to the reputation of the author or character, **Pathos**: appeals to the emotions of the audience, e.g. fear, sadness, contentment, joy, pride.

The user can request a subject (e.g. War in Afghanistan) and a rhetoric strategy to be used. The range of possible requests goes from simple, such "Show all opposing statements", to complex requests, such as "Attack (or support) a pro position".

The engine is aware through the annotations that a particular interview is supporting or attacking a certain issue (e.g. War in Afghanistan). Since the statements expressed in an interview are annotated rhetorically in a time-coded formal notation, the engine knows where an argument is expressed in the video. Every expression of an argument is called a statement, which is formally defined through its components.

The engine implements a simple logic that given a statement can generate other statements opposing or supporting it: for ex. from "war only solution" it can make the opposing statements "war not only solution" and "diplomacy only solution". The engine has no knowledge of terms like "war" or "diplomacy" or "military action", but relies on an RDF-encoded knowledge base that states that "war" is opposite to "diplomacy" and similar to "military action". This knowledge base can be modified or replaced to modify the way the engine generates statements.

Subsequently, the engine can look in the video material for video segments expressing the generated statements and edit them together. In doing that, elementary principles from Film Theory are used to obtain an acceptable result.

Another technique that the engine uses is to superimpose images while the interviewee is making a particular statement, where the picture is meant to provoke an emotional response in the viewer supporting or attacking the statement (in rhetorical terms, it appeals at the *Pathos* of the viewer). This is also based on the same logic described above.

This simple mechanism provides a method to automatic edit different video sequences with a knowledge (at least partially) of the semantic of the resulting video. What is appealing here is that the basic mechanism is simple and even if limited to short sequences, there is a clear purpose motivating the selection of video fragments (e.g. attacking a certain opinion).

Therefore, the proposed annotation schema based on rhetorics and the supporting/attacking logic mechanism represent respectively one possible way to achieve the goals presented in the introduction, i.e. "formal modeling of the source semantics" and "semantic-aware structuring of video segments".

4. FUTURE DIRECTIONS

In this section we present the main directions we will develop further in our research:

Use the *Ethos* of the interviewee: the voice and the way the interviewee looks, his or her social status are also influencing the viewer. The Ethos could be deducted from the first level of annotations used in our project.

Integrate more *Film Theory* techniques. Editing can be used to create effects that convey special semantics (e.g. diminishing the length of the cuts to augment the tempo or adding particular music or background sounds).

Consider cognitive theory of *emotions* like Ortony [7]. This can provide a framework to define when the viewer would feel sympathetic with the speaker and therefore more inclined in accepting his/her thesis.

Create *high level strategies* to automatically edit long sequences. Up until now the system assembles sequences based on supporting or contrasting single statements; to insure that there is a line of development in the edit result, we are looking at Storytelling to organize the material so as to provide some narrative evolution in the resulting video.

Investigate *validation techniques* and take into account user models as input to the video generation process.

Investigate *different semantic-aware composition mechanisms*, i.e. apply the ideas presented in this paper to other domains.

5. REFERENCES

[1] S. Bocconi and F. Nack. VOX POPULI: Automatic Generation of Biased Video Sequences. Technical Report INS-E0405, CWI, June 2004.

[2] G. Davenport and M. Murtaugh. ConText: Towards the Evolving Documentary. In *ACM Multimedia '95, Proceedings*, pages 377–378, November 1995.

[3] M. Davis. *Readings in Human-Computer Interaction: Toward the Year 2000*, chapter Media Streams: An Iconic Visual Language for Video Representation., pages 854–866. Morgan Kaufmann Publishers, Inc., 1995.

[4] G. Houbart. *Viewpoints on Demand: Tailoring the Presentation of Opinions in Video*. PhD thesis, Massachusetts Institute of Technology, 1994.

[5] M. Mateas. Generation of Ideologically-Biased Historical Documentaries. In *Proceedings of AAAI 2000*, pages 36–42, July 2000.

[6] F. Nack. *AUTEUR: The Application of Video Semantics and Theme Representation in Automated Video Editing*. PhD thesis, Lancaster University, 1996.

[7] A. Ortony, G. L. Clore, and A. Collins. *The Cognitive Structure of Emotions*. Cambridge University Press, 1999.