Vox Populi : generating video documentaries from semantically annotated media repositories

Vox Populi: generating video documentaries from semantically annotated media repositories

Stefano Bocconi

CIP-DATA LIBRARY TECHNISCHE UNIVERSITEIT EINDHOVEN

Bocconi, Stefano

Vox Populi: generating video documentaries from semantically annotated media repositories / door Stefano Bocconi. -
Eindhoven: Technische Universiteit Eindhoven, 2006.
Proefschrift. - ISBN 90-386-0824-1. - ISBN 978-90-386-0824-2
NUR 983

Subject headings: information presentation / hypermedia / video technology / ontology / knowledge management
CR Subject Classification (1998) : H.5.4., H.5.1, I.7.2., I.2.4, H.3.1.

SIKS Dissertation Series No. 2006-27
The research reported in this dissertation has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Cover design: Alia Amin
Printed by University Press Facilities, Eindhoven, the Netherlands.

Vox Populi: generating video documentaries from semantically annotated media repositories

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven,
op gezag van de Rector Magnicus, prof.dr.ir. C.J. van Duijn,
voor een commissie aangewezen door het College voor Promoties
in het openbaar te verdedigen
op donderdag 30 november 2006 om 14.00 uur

door

Stefano Bocconi

geboren te Florence, Italië

Dit proefschrift is goedgekeurd door de promotor:

prof.dr. L. Hardman

Copromotor:
dr. F. Nack

Preface

I have recently heard that writing a PhD thesis is a life-changing experience. I certainly agree with that, and I would add that the whole PhD experience is a life-changing experience, and not only because in four years time life changes anyway.

It all started before this PhD with a personal project, making a documentary about the real opinion of the American people in the aftermath of 9-11. We thought this opinion was simply not being represented by the media that had too many interests to be truthful. Later, while reading documentary books from real documentarists I discovered how amateurish and unorganized we were, but that did not compromise the enthusiasm and good will we put in the project. I would like to thank the other members of the Interview with America group, Francesca Bertè, Simone Chiappi, Alessandro Bologna and Stefano Travelli for the incredible experience of shooting that documentary together and for allowing me to use the material for my research.

Our project did not turn out to be as successful as we had in mind. Our documentary has not told the world how American people felt about 9-11, but it has taught us several things. For one, this experience was important when Lynda Hardman, head of the then called "Multimedia and Human-Computer Interaction" group at CWI in Amsterdam, decided to offer me a PhD position. At that moment I did not have any idea what kind of research could be based on the video footage I had, and it took me more or less one and a half years to see it. At that point many pieces of the puzzle fell in place, nicely joining my experience as an amateur filmmaker (and the problems I had had while trying to produce the documentary) with scientific directions capable of providing some answers. The outcome of this endeavor can be read in this thesis, that went through the careful review of prof. Craig Lindley, prof. Anton Nijholt and prof. Matthias Rauterberg, to whom I am grateful.

During all my research I have been exposed to high research standards that, if on one hand they have created pressure and stress, on the other they have given me an idea about how research should be done. For this I am grateful to the INS2 group in general and in particular to my supervisors Lynda, Jacco, Lloyd and especially Frank, who followed my daily work and my occasional progress.

The work environment has been the most lively in my not too short work career. Professional relations were nicely mixed with personal contacts: thanks to Lynda for hosting many group events, where culture was never subordinated to food, to "the brain" Frank for filling up our agenda with all kinds of events, to my room-mate Lloyd for our special way of communicating with singing and rhymes, to Jacco for his sharp comments and insights, to Katya for her practical and at the same time emotional approach to life, to Joost for his life philosophy with which I almost totally disagree (except for the motto "denken is voor paarden"), to Yulia for the endless discussions and her love for research. I also wish all the best to the new members of the group, Michiel, Georges, Željko, Raphaël and especially Alia, who designed the cover of this

thesis. Thanks a lot for that. Even though our ”rival” group INS1 does not deserve any mention :-), I still want to give many thanks to Nina (and Volker) for all the good things that thanks to her enthusiasm and energy we did together, and to Arjen and Roberto for all the life-related philosophical discussions.

Beside a professional life, I managed to have, albeit not always, a social life, of which music has been a big part. I would like to thank Atreya and Jeroen for the opportunity they gave me to play on a professional level (with David and Peter) and to contribute at the same time to ”Project Aware!”[1], a project for human rights. I have learned a great deal of what it means to be a musician, and we have been through happy but also difficult moments together as a band. I also would like to thank Nanda (”la mia sorellina”), Manu and Simone, for all the fun we had together, playing for ourselves and for our friends. Music has never been just music in these last years. I will never forget the heart-warming experience I had spending time with all of you. I also want to thank all the people that were close to me in Amsterdam, especially Alessandro, Elena, Paola, Silvia, Niels, Laura, Pierluigi, Gabriele, Diego, Nacho, and Salvo.

Special thanks go to Hiske, who has supported me during all these years I have spent in the Netherlands. Without her I would not have done any PhD, as well as many other things. Her and her family felt and still feel like my ”Dutch” family. I also want to thank my real family, my parents and my brother for the distant love that always followed me while I was far away.

At the end of a life-changing experience there is of course a new life: I would like to thank Federica for standing by me while I find my way in a strange and not always easy country, my own Italy.

Stefano Bocconi

[1] http://www.projectaware.nl/

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Once upon a time there was a group of friends, 4 men and 1 woman, who decided to make a documentary about the consequences of 9-11 in the United States of America. After shooting for some days around the East coast, they came back with the video footage and set out to edit it into a documentary to be shown to the world (via Internet). Almost each of them had different ideas about what material should have been included in the final version and how the documentary should be shaped. The group had lots of discussions, that became arguments, that led to quarrels and destroyed their friendship. The present thesis aims at restoring their friendship (and saving future ones) by providing a method that, had it been available back then, would have made all final cut decisions unnecessary.

1.1 Motivation

We are used to viewing a film or documentary as a fixed static artifact, with an internal logic, a subject and acceptable duration. This artifact is the product of a director who crafted it for us, using footage also recorded for the purpose of making a film. This scenario does not include viewer intervention except at the last stage, when the viewer can decide to view the film or not, and even then the choice is pretty limited to watch (a part of) it or ignore it.

On the other hand, documentary makers (we call them documentarists) nowadays are experimenting with new forms of authorship. In many collaborative initiatives on the Internet, the author is not a single person, and the role of authorship is distributed among the participants. For example, the **Echo Chamber** project[1] is a collaborative filmmaking initiative to produce a documentary about how the television news media became an uncritical echo chamber to the Executive Branch leading up to the war in Iraq. Users can transcribe, tag and rate (part of) video interviews, and this information is used to select and compose the material for the final documentary. Another example is **Voices of Iraq**[2], a documentary created by distributing over 150 digital video cameras across the entire country to enable everyday people — mothers, children, teachers, sheiks and even insurgents — to document their lives and their hopes amidst the upheaval of a nation being born. Two factors make experimenting with different forms of documentaries easier than in the past. The first is that documentary making today,

[1] `http://www.echochamberproject.com/`
[2] `http://www.voicesofiraq.com/`

in particular when using digital video instead of film, has become affordable for many people: low prices for hardware (video cameras and computers) and software (video editing programs) have increased the number of potential documentarists. The second is that the Internet can be used as a cheap distribution channel, potentially accessible to everyone with a (broadband) connection.

At the same time, the Internet is changing viewers from passive to active, making them more used to interact with the content and the form of the information presented. Nonetheless, with the traditional documentary production method, this is not possible: the director crafts the same version for all viewers.

Why is this a problem? Because documentaries are meant to inform as well as entertain. For a film or documentary, the video material collected during shooting is quantitatively much more than the material that is selected for the final version. In the case of a documentary, this can mean that large amounts of footage with different themes, topics and arguments will never be seen by the viewer, sometimes only because of time limits. Moreover, when a documentary is about a matter-of-opinion issue, a documentarist has the power to build a strong argument either for or against an opinion by selecting and editing different footage from the available material. A documentarist determines a documentary's content and point of view for all viewers, where the viewers themselves would probably have made other choices.

On the other hand, making all footage available is not a suitable alternative, because a viewer is unlikely to be willing to watch hours of video with no story or theme, no apparent relation between a sequence and the following one. Documentaries still offer good narrative models to show content in a way that does not overwhelm the viewer. In case the subject is a matter of opinion, the documentarist can edit the footage so that different facts and opinions are related to each other and presented in the form of arguments pro or against the matter being exposed.

In research, several solutions have been investigated to let the viewer take the seat of the director, resulting in different automatic video generation approaches [44, 47]. Potentially, automatic video generation does not only have advantages for the viewer, but also for the documentarist. A video generation system could help the documentarist by:

- automatically presenting the material, freeing the documentarist from the need to select and edit the footage, which is a difficult task.
- generating different documentaries from the same footage, facilitating reuse of the media asset.
- dynamically evaluating the material to use for a documentary, allowing new footage to be added at a later stage and making the documentary an evolving up-to-date video document rather than a static final product

1.2 Problem definition

Both traditional documentary making and automatic video generation have their strengths: the former is capable of presenting issues to the viewer in a way that is informative and interesting at the same time, while the latter allows the documentarist to provide viewers with documentaries dynamically generated according to their interests.

To date, there is no single approach capable of combining the advantages of human authoring with the potential benefits of automatic video generation. More specifically, there is no single approach that:

1. generates documentaries on matter-of-opinion issues which use presentation forms as a documentarist would do, showing contrasting points of view related to each other.

2. allows the viewer to select both the content of the generated documentary and, considering that information presentation on matter-of-opinion issues is not neutral, the documentary's point of view.

3. allows the documentarist to collect material to be used for documentaries, without having to specify how this material should be presented to the viewer.

The motivation for the first point is that the viewer should be familiar with the way the documentary presents the information, so that she can understand it, be informed and entertained by it. The other two points are required to shift authorship from the documentarist to the viewer: the second one concerns the choices the viewer might want to make instead of the documentarist, and the third one is about what the documentarist needs to release her role to the viewer (using the automatic generation system).

1.3 Approach

In this thesis we focus on documentaries about matter-of-opinion issues, where opinions are mainly expressed by people being interviewed. In this type of documentary the drawbacks of having a final static version are evident: especially when the number of interviewees is high, some of the interviewees' answers will not be selected, and possibly some opinions will never be displayed to the viewer. We propose an automatic video generation approach that allows the viewer to potentially see all material shot for a particular documentary, not only what a documentarist would select. The content of such a documentary is determined by the viewer choosing a particular subject she is interested in and the point of view she wants. The material is then organized according to presentation forms also used by documentarists. Automatic video generation allows the repository containing the raw footage to grow by adding relevant material, and both new and old material to be used to generate new documentaries.

To make this possible, in this thesis we develop an **annotation schema** and a **generation process**. A documentarist should not edit the material she shot in a final format, throwing away all that does not fit in it. Instead, she should describe the raw footage according to the annotation schema described in this thesis, and allow the generation process to assemble video documentaries based on these annotations and on what the viewer wishes to see. The benefits of our approach are:

- for the **viewer**, documentaries generated according to her specific requests for content and point of view, created using existing material from repositories such as broadcasters archives or documentary initiatives.

- for the **documentarist**, no need to select and edit the material, multiple usage of the same footage and the possibility of adding new material.

The price to pay for all these benefits is in the time and resources spent in annotating the footage. The annotation effort is, in general, a limiting factor to the adoption of automatic generation systems. We seek to use the simplest annotation schema that enables the flexibility of generation that we look for.

Our approach is based on some experience with documentarists, relevant literature and the lessons learned while implementing an automatic video generation model. Throughout our work the focus has been to formulate a model that could potentially solve the problem defined in section 1.2. We reduced the scope of our approach by using simple user-models and providing a simple user interface, aimed at showing the model's capabilities more than to suit end-user needs.

We test our approach on material from **Interview With America** (IWA)[3], which is an online documentary shot by a group of independent amateur documentarists. "Interview with America" is a matter-of-opinion documentary and contains interviews with United States residents from different socio-economic groups on the events happening after the terrorist attack on the 11th of September 2001. Issues discussed include the war in Afghanistan, anthrax, media coverage and social integration in multicultural societies. The repository contains about 8 hours of digital video mostly shot outdoor with hand-held video cameras.

1.4 Research Questions

From the above discussion we can now formulate specific research questions. A distinguishing point for this research is the goal of presenting information in a way the viewer is familiar with. We dedicate the first research question to this issue. We then need to define a video generation model capable of implementing such presentation forms. The second and the third research questions focus on the components of such a model: the annotation schema and the generation process. Finally, since the documentarist cannot check the quality of each generated documentary, the generation model needs to provide an indication about what it is able to generate. This issue is the subject of the last research question.

Documentary Form

RESEARCH QUESTION 1 *(*DOCUMENTARY FORM*) What characteristics of the presentation forms used by documentaries on matter-of-opinion issues must be modeled?*

In the present work we seek to model presentation forms inspired by traditional documentary making, with the assumption that these forms make the resulting video a more engaging experience for the viewer. This question addresses the problem of modeling such presentation forms. We focus on the generation of video documentaries on matter-of-opinion issues. Such documentaries show contrasting opinions and facts supporting them so that the viewer is informed on how people (and the documentarist) look at particular issues. Ideally, our video generation approach should generate documentaries with the same characteristics as human authored ones. This question is answered in chapter 2, where we describe what characteristics must be modeled.

Annotation Schema

RESEARCH QUESTION 2 *(*ANNOTATION SCHEMA*) What information should be captured in an annotation schema for an automatic video generation approach where:*

- *the viewer can specify the subject and the point of view,*

[3] `http://www.interviewwithamerica.com/documentary.html`

- *the documentarist can collect material to be used for documentaries, without having to specify how this material should be presented to the viewer,*
- *the material is presented according to presentation forms used by documentarists?*

Having determined the presentation form of the material, we need to determine a model capable of generating documentaries of this form, about the subject and point of view requested by the viewer. We decompose the generation model into two components: this research question addresses the problem of defining the first one, i.e. an annotation schema for media material capable of capturing the information necessary for viewer requested matter-of-opinion documentaries. The annotation schema must be designed as a trade off between expressivity (allowing richer presentations) and complexity (requiring more effort needed to annotate media material), since its complexity can discourage a documentarist from adopting our approach. The requirements for the annotation schema are derived in chapter 3 and the annotation schema is specified in chapter 4.

Generation Process

RESEARCH QUESTION 3 *(*GENERATION PROCESS*) How must a generation process be defined for an automatic video generation approach where:*

- *the viewer can specify the subject and the point of view,*
- *the documentarist can collect material to be used for documentaries, without having to specify how this material should be presented to the viewer,*
- *the material is presented according to presentation forms used by documentarists?*

The second component of our video generation model is a process that can manipulate the annotations and create a documentary, with the requested subject and point of view and using the presentation forms modeled in research question 3. We need to specify a process that has as inputs the annotations, the viewer-specified content and point of view, and as output a documentary. More specifically, this question addresses the technical problem of describing how the annotations specified in research question 2 can be manipulated to select and sequence video according to the presentation forms introduced in research question 3. The requirements for such a process are defined in chapter 3 and the process is defined in chapter 5, section 5.2.

Author Support

RESEARCH QUESTION 4 *(*AUTHOR SUPPORT*) How must a generation process be defined so that it can give to the documentarist an indication of the quality of the documentaries it can generate?*

In our automatic video generation approach, the selection and sequencing of the material is done by the generation process. The documentarist is unable to check the quality of the final result. Therefore, the generation process must allow the documentarist to verify whether the documentary generation will be effective or not, without having to check all possible documentaries that can be generated. This question is answered in chapter 5, section 5.3.

1.5 Research contributions

On a high level, our contribution is in defining an automatic video generation approach where rhetorical presentation patterns used by documentarists are combined with a data-driven approach. Rhetorical presentation patterns provide the viewer with an engaging viewing experience, while a data-driven approach can be applied to growing media repositories. To date, rhetoric has been achieved in a top-down manner using closed repositories, while data-driven generation approaches were unable to implement complex rhetorical presentation patterns. In details, the contributions of the thesis are the following:

Based on literature, we formulate the requirements for an automatic generation model for matter-of-opinion documentaries, in terms of presentation forms, viewer interaction and authoring support for the documentarist. Since these requirements define what the model must be capable of, but not how it should achieve it, we call them **high-level requirements**.

We specify technical requirements for an automatic video generation model in terms of properties the annotation schema must have and techniques the generation process must implement. Since these requirements are about technical details of the model, we call them **low-level requirements**.

We define an **annotation schema** that satisfies the first group of low-level requirements we set. We define a **generation process** that satisfies the second group of low-level requirements.

We implement the annotation schema and the generation process in a demonstrator called **Vox Populi**.

1.6 Guide to reading this thesis

For the 10 minute reading, read the introduction (you are nearly finished now) and the conclusions. If you have slightly more time, read also the description and the summary of each chapter. Of course, if you have the time please read the entire thesis. If you want to have a content-based access, here is the thesis outline:

Chapter 2 describes the domain of our research, video documentaries, first from a general point of view and then zooming in on the type we focus on, matter-of-opinion documentaries based on interviews. Since we aim at automatic generation of video sequences, we also describe the process of documentary making and how editing is used to communicate information. This domain analysis is used to derive *high-level requirements* for a generation model. This chapter answers *Research Question Documentary Form [1].*

Chapter 3 further specifies the generation model, by examining research related to the issues raised by the high-level requirements. We examine three areas: argument building, annotations and narrative generation, discussing previous approaches and pointing out which we can and cannot reuse. This analysis is used to formulate *low-level requirements* for the generation model, which we divide into two components, namely an

annotation schema and a generation process.

Chapter 4 describes an *annotation schema* that satisfies the requirements set in the previous chapters. More specifically, this schema models information necessary to represent the presentation forms used in documentaries, and to implement editing rules derived from the analysis of film theory in chapter 2. This chapter answers *Research Question Annotation Schema [2].*

Chapter 5 defines a *generation process* that satisfies the requirements set in chapter 2 and chapter 3. This process manipulates the annotation schema presented in the previous chapter and implements the presentation models defined in chapter 2. This part answers *Research Question Generation Process [3].* We explain how the annotation schema and the process we defined can be evaluated to provide the documentarist with feedback about the quality of the automatically generated documentaries before they are actually generated. This part answers *Research Question Author Support [4].*

Chapter 6 presents an implementation of the generation model we call *Vox Populi*. The user interface and user interaction is explained, together with the technologies used for the implementation. Based on this implementation, we examine the results of our research. Experiences in using Vox Populi are reported and an empirical evaluation is presented.

Chapter 7 provides a summary of the research contributions and discusses how the research questions were answered. Alternatives solutions to our design choices are presented and their implementations are discussed as future work.

Appendix A contains a legend with the typographical conventions used in this thesis.

Appendix B is a quick reference listing the research questions and the requirements we define in this thesis.

Appendix C provides background information on video documentaries.

Appendix D contains details of the annotations used in the IWA project.

Chapter 2

Video documentaries

In this chapter we describe the focus of our thesis, namely interview documentaries about matter-of-opinion issues. More specifically, we first concentrate on the structure, looking at how information is organized in documentaries. We then describe the process of making a documentary, and in particular film editing. From this analysis we determine the structures and processes an automatic video generation approach needs to model. This discussion leads to the definition of high-level requirements at the end of the chapter.

2.1 Introduction

The goal of this chapter is to specify which characteristics of interview documentaries about matter-of-opinion issues should be modeled, as specified in *Research Question Documentary Form [1].* For an automatic video generation approach we need to focus on the *content* of documentaries and its structure, as well as on the *process* of assembling this content in a documentary. Consequently, we describe documentaries from two points of view: firstly, in section 2.2, we look at the structure of documentaries, and how information can be presented in a documentary. We describe how information can be organized, abstracting from the possible subjects documentaries can have, with special attention to interview documentaries. Secondly, in section 2.3, we examine documentary making, describing the different phases of the process. We take a closer look at the editing phase, since this phase is particularly relevant for an automatic video generation approach. Based on this analysis, in section 2.4 we then discuss aspects of documentaries an automatic video generation approach must model and we specify those aspects in high-level requirements, which we use to guide the design of the model in the next chapters.

2.2 What is a documentary?

To describe documentaries we use different literature sources: Michael Rabiger [54], Barry Hampe [33] and Sheila Curran Bernard [8], because they are documentarists and have direct experience as documentary directors, and David Bordwell [12], because he provides a film theory point of view, useful for positioning documentaries with respect to other film genres.

We use a general definition of the documentary genre from Bordwell ([12], p. 128):

> "*A documentary film purports to present factual information about the world outside the film.* "

This definition stresses that the most distinctive feature of a documentary is the intention of presenting informative content. Typically, a film labeled as *documentary* leads the viewer to assume that the persons, places and events exist and that the information presented is trustworthy ([12], p. 128). Rabiger says that documentaries explore actual people and actual situations and they always reflect a profound fascination with, and respect for, actuality ([54], pp. 3-4).

If we look at how documentaries present this factual information, Bordwell distinguishes three types of form: the **narrative form**, the **categorical form** and the **rhetorical form** ([12] p. 132). Documentaries often use more than one of these forms, but usually one type is predominant over the others. We discuss these forms in sections 2.2.1, 2.2.2 and 2.2.3. The rhetorical form is particularly relevant for the type of content we are interested in, i.e. matter-of-opinion issues, and in section 2.2.4 we examine this form's building blocks, i.e. the arguments. We then discuss in section 2.2.5 the type of documentaries we focus on in this thesis, namely interview documentaries, and in particular we examine interview documentaries where the subject of the interviews is a matter-of-opinion issue. Stories can be told in documentaries from different vantage points, called Points Of View (POV). In section 2.2.6 we describe the different POVs a documentary can have, and which of these apply to interview documentaries. The discussion about points of view leads to the investigation in section 2.2.7 whether documentaries are objective, how documentarists can misrepresent reality, and how this can happen when showing interviews.

2.2.1 Narrative form

According to Rabiger, documentaries need a story ([54], p. 4):

> "*Successful documentaries, like their fiction counterparts, also need a good story with interesting characters, narrative tension, and an integrated point of view. These elements are fundamental to all stories and are present in myth, legend, and folktales—humankind's earliest narratives.* "

A **narrative** is what we usually mean by the term story. Bordwell considers a narrative to be a chain of events in cause-effect relationship occurring in time and space ([12], p. 69). Usually the agents of cause and effect are *characters*, who play roles in a story by triggering and reacting to events. Characters are often human or human-like, but not necessarily, for example as in documentaries about animals ([12], p. 72).

A narrative usually involves a change from an initial situation to a final situation, and this change is reflected in the structure of the documentary. We call the structure level the **macro-level**. A common structure for films is to begin with an **opening** (or exposition) that introduces events and characters necessary for the viewer to understand the story, develop the plot to a **climax** and resolve (or not) the chain of cause and effect in the **ending**. Hampe uses the following structure for documentaries ([33] p. 123-126, 294-295):

- the **opening** is the point before which nothing needs to be said. The opening states the theme, asks a question, or shows something new or unexpected. It gets the documentary started and raises the expectations of the audience. Within or following the opening, the **explanation** (or **exposition**) presents the purpose of

the documentary and the problem or problems it deals with—the basic information the audience needs to understand where the documentary is heading toward.

- the **middle** builds the story by exploring conflicting elements of the situation and showing evidence in support of and in opposition to the theme. The sequence of presenting one kind of evidence followed by opposing evidence may be repeated several times as the documentarist explores a variety of subthemes.

- the **ending** is the point beyond which nothing needs to be said. The ending shows the outcome—which up to this point may have been somewhat in doubt—in which the conflicting elements are handled and resolved.

Within each part of this structure, a narrative film can be divided into **scenes**, which are distinct phases of the action occurring within a relatively unified space and time. We call the scenes level the **micro-level**.

While for Rabiger a documentary needs a story, for Bordwell, documentaries can also use two non-narrative forms—the categorical and the rhetorical form.

2.2.2 Categorical form

Categorical documentaries organize information in categories. These **categories** are groupings that individuals or societies create to organize their knowledge of the world. Categories can be strictly defined, as in science (e.g. for plants and animals), or more based on common sense, for example advanced and primitive societies ([12] p. 133). Categories and subcategories provide a form for a documentarist to use in order to organize the information she wants to convey. One classic documentary organized categorically is Leni Riefenstahl's *Olympia, Part 2* [55], made as a record of the Berlin Olympics in 1936. Riefenstahl describes the Olympics by breaking down the games into subcategories—sailing events, sprinting events, and so on, and showing the beauty of each of them[1].

Usually a documentary is not completely categorical: for example, the documentarist can introduce small-scale narratives (at the micro-level), or inject some rhetorical form into the documentary ([12] p. 134 and p. 140).

2.2.3 Rhetorical form

In using the **rhetorical form**, a documentary aims at persuading the audience to adopt an opinion about the subject, usually a matter-of-opinion issue. In a rhetorical documentary, the documentarist tries to make her position seem the most plausible by presenting different types of arguments and evidence (Bordwell [12], p. 140). An example of a rhetorical documentary is *The River* [42], made in 1937 by Pare Lorentz to promote Franklin Delano Roosevelt's policies.

According to Bordwell, rhetorical documentaries do not tend to point out other opinions, as in the case with *The River*. Rabiger, instead, discusses three different ways a documentarist can behave, depending on her respect for the audience: at the lowest level of respect there is the **propagandist**, who wants to condition the audience, showing only the evidence supporting predetermined conclusions. Moving up the scale of respect, there is the **binary communicator**, who gives "equal coverage to both sides" in any controversy. Rabiger says that this form considers the audience as a

[1] Appendix C provides more information about categorical documentaries for the interested reader.

passive mass to be informed and entertained, but not challenged to make judgments. At a higher level, is the **mind-opener**, who aims not at conditioning or diverting but at expanding the viewer's mind, by presenting something in all its complexity, never patronizing or manipulating either the subjects or the audience ([54] pp. 8-9). Hampe says that the documentarist, even when taking a strong position, should present more than one point of view ([33] p. 125):

> "*A documentary is expected to explore conflicting elements of the situation. This doesn't mean that it has to be passively neutral. But even when it takes a strong position in its theme, it should be able to acknowledge that this position isn't universally accepted. If it were, there would be little reason to make a documentary. One of the differences between information and propaganda is the willingness of the former to acknowledge that other points of view may legitimately exist, even if they are considered wrong.* "

Presenting opposing positions also has a number of advantages for the documentarist. Firstly, it can be used as a technique to make the audience want to see what will happen next. The purpose of this is to introduce some level of dramatic conflict into the structure of the documentary. Dramatic conflict is a structural tension that keeps the outcome of the documentary somewhat in doubt and keeps the audience interested ([33] pp. 298-299). Furthermore, presenting opposing positions can be used to organize the material of a documentary, as we showed in section 2.2.1 discussing the middle of a documentary. Finally, presenting opposing positions can be used as an editing technique, where both sides are presented simultaneously, for example when the voice-over contradicts what is being shown visually.

Theoretically, presenting opposing positions could be done using the categorical form as well, by presenting the category *evidence related to position A* followed by the category *evidence related to position B* and so on. Such a presentation would lack the narrative property of being a chain of cause and effect, since categories have normally no direct relation to each other. Using a more narrative approach in interview documentaries, the documentarist can present opposing positions as an ongoing debate, so that what someone says causes and is followed up by what someone else says. In this way people seem to react and reply to each other's arguments.

2.2.4 Arguments

Typically, when people state their opinion or position, they do so with a discourse intended to persuade or prove that their conclusions are correct. We call such a discourse an **argument**. A position or opinion is therefore supported by one or more arguments. According to Bordwell ([12] p. 141), three main types of arguments are used in the rhetorical form:

- **subject-centered arguments** are arguments about the film's subject matter. Some examples of this are appeals to common beliefs, or the use of logic and examples to prove a point.

- **viewer-centered arguments** are arguments that appeal to the viewer's emotions. Examples of topics in common arguments in this category are patriotism, religion and romantic sentimentality.

- **arguments from source** try to convince the viewer that the information comes from a reliable source, the makers are intelligent, well informed, sincere, trustworthy. The implicit argument is that since the information comes from reliable sources, the viewer should let him or herself be persuaded by it. For example the narrator could be chosen to have an authoritative voice.

These argument types correspond to the categories the Greek philosopher Aristotle (in his book "Rhetoric" [3]) used to classify means of persuasion into: logos, pathos and ethos, respectively. These means of persuasion are used by a *speaker* (or author) who tries to convince an *audience*, and are defined as follows:

- **logos**: appeals to logic or reason. The argument is based on factual data and on the conclusions that can be drawn from it. The audience should accept the argument because it sounds rational.

- **pathos**: appeals to the emotions of the audience. Pathos does not concern the truthfulness of the argument, only its appeal. The audience should accept the argument because of how it feels.

- **ethos**: appeals to the reputation of the speaker. The audience should accept the argument because the speaker is trustworthy.

In documentaries, the audience is the collection of viewers. The speaker can be the documentarist herself (when she appears or speaks in the documentary), the narrator, if there is one, or, in the case of interviews, the interviewees. In this thesis, rather than Bordwell's terminology, we use logos, pathos and ethos because of their more general applicability.

2.2.5 Interview documentaries

Up until now we have discussed documentaries based on the way they present information. In this section we specify the type of content for the documentaries we focus on, i.e. interview documentaries.

Interview documentaries (also called talking heads documentaries) record testimonies about events or social movements. There are three approaches when filming interviews. The first approach has the interviewer and the interviewee appear in the picture, revealing the process of making an interview (as in Michael Moore's "Roger and me" [46]). In the other two approaches, the interviewer's voice is often edited out and only the interviewee is in picture. In the second approach, the interviewer sits to one side of the camera and out of frame, which makes the interviewee look off-camera at an unseen interlocutor. The third approach has the interviewee answering an interviewer who sits with her head just below the camera lens, which makes the interviewee appear to be looking directly into the camera. These approaches have different effects on the audience. In the first two cases the audience is made aware of the fact there is an interviewer, even when she cannot be seen and her voice cannot be heard. In the last approach, the audience is in a direct relationship to the person on the screen ([54] pp. 183-184).

When the subject of the interviews is controversial, the way arguments are presented is particularly important for the audience to decide whom to believe, because evidence to determine the truth might be lacking. According to Hampe ([33] p. 63):

> "*You're shooting a documentary about a subject that has become controversial. One side makes charges. The other side denies them and makes countercharges. You shoot interviews with spokespersons for both sides. What evidence do you have? The fact is that while an interview is prima facie evidence that the person shown said the words that were spoken, it carries no evidence whatsoever about the truth value of the statement the person makes. But an audience, like a jury, is not above using other cues to decide whether or not to believe a speaker. His or her dress and manner, as well as the logic of the statement, can have a powerful effect on them.* "

Therefore, even though different interviews can make the same point, they do not look equally convincing to the viewer. The factors that can influence the viewer in Hampe's words can be classified using the categories we introduced in section 2.2.4: dress and manner belong to the ethos of the speaker, while the logic of the statement corresponds to her logos. Pathos is not mentioned by Hampe. Logos, pathos and ethos determine the strength of a point made by a speaker. The documentarist can deliberately choose a stronger or weaker speaker to make a particular point (Hampe [33], p. 63-64).

2.2.6 Point of view

The phrase *point of view* suggests a bias, taken as a lens on a subject under scrutiny. Point of view can also be considered more concretely as the vantage point from which a story is being told, and specifies the relation between the *storyteller* and the characters of the story ([54] p. 322). Rabiger introduces the following point of view (POV) categories[2] ([54] pp. 323-336):

- **single POV**: the story is channeled through a main character. This person may be a bystander or a major protagonist, and she is either observing, recounting or enacting events.
- **multiple POV**: the story is told by multiple characters, of which none tends to predominate. As in the Single POV, each character is either observing, recounting or enacting events.
- **omniscient POV**: the story is told by an entity who knows and sees all, more than the characters of the story know and see, and can move freely in space and time.
- **reflexive POV**: the process of film making is shown in the story itself, so that the viewer is made aware that films are creations of the filmmaker, and not objective records of unmediated life.

Documentaries do not always fall exactly into only one of these categories, but they usually adopt a major POV possibly complemented by secondary ones. The documentarist can speak directly to the audience using the omniscient POV, taking the role of the omniscient entity, or appearing as one of the characters in a single or multiple POV (as Michael Moore in "Roger and me"). A documentarist can also decide that another character or characters speak for her, adopting the single or multiple POV. A multiple

[2] We will use in the following *point of view* to mean the bias on a particular subject and *POV* to mean the vantage point from which a story is told.

POV can be used to present different and often counterbalancing viewpoints, for example, when each character represents a different social class, and the documentary shows a social process, its actors, and its outcome.

Using the reflexive POV, a documentarist can build into the film whatever doubts and perceptions would not be adequately acknowledged through showing the material on its own ([54], p. 359). Reflexivity allows the documentarist to show to the viewer her own cultural and ethical assumptions in making the documentary. Rabiger states ([54], p. 334):

> "*Aside from the issues of distortion and misinformation, there are fascinating and more abstract issues concerning the medium's boundaries. What may or may not be ethical? How, when, and why do we suspend disbelief? What deceptions does the medium practice on its makers? and so on. Such questions properly assume that film is an emerging and imperfectly understood medium, rather than a finished tool whose only use is as an informational or advocacy vehicle for a "subject".* "

A multiple POV technique being often used to make and present interviews is called **vox populi**, i.e. voices of the people ([54] p. 176). It consists of asking a number of people the same few questions, and then stringing the replies together in a rapid sequence. It is useful for demonstrating a range of opinions of people, and can be used to show either diversity or homogeneity.

Furthermore, when a documentary narrowly focuses on one or a few people (i.e. it uses a single POV), vox populi can be used as a montage technique for creating a "Greek chorus" of faces and voices ([54] p. 224), adding sequences of man-on-the-street interviews. Such a technique can be used to demonstrate where a main character belongs in relation to mass opinion and to remind the audience of the diverse opinion of the common person. From an editing perspective, long sequences of someone talking can be condensed by showing only the salient points, cutting in between them to other interviews, in this way creating a dialectical counterpoint between a main character and the ubiquitous man on the street. According to Rabiger ([54] p. 185):

> "*Two vital purposes are thus served: multiple and conflicting viewpoints can be evoked, and the material can be focused into a brief screen time.* "

As we discussed in the previous section, in the case of controversial issues, the documentarist can select the interviewees representing each of the opposing positions. A stronger or weaker speaker will make one of the opposing positions look stronger or weaker. In this way, even in a single or multiple POV the documentarist can impose her own point of view on an issue.

2.2.7 Is a documentary objective?

In documentaries about controversial issues, the documentarist expresses her point of view. Rhetorical documentaries aim at persuading the audience to adopt an opinion. A natural question is then how trustworthy a documentary is. A film labeled as a documentary leads the viewer to assume that the people and the situations shown are real and that reality is portrayed in a trustworthy manner (section 2.2). On the other hand, a documentary is a subjective construct. Even disregarding editorial decisions on the footage, an objectively recording camera has been subjectively placed somewhere and

someone has decided to start and stop it ([54] p. 7). People tend to think a documentary is objective when it balances out opposing points of view, giving an unbiased view of the events and personalities in question (the *binary communicator* defined in section 2.2.3). In contrast, a documentarist needs to interpret and to decide where the cause of justice and humanity lay in specific issues ([54] p. 6).

Even though reality is the subject of a documentary, the material shown is not necessarily a record of events shot when they were happening. The documentarist can also stage certain events for the camera to record. Documentarists regard some staging as legitimate in a documentary if the staging serves the greater purpose of presenting information. Regardless of the details of its production, the documentarist asks the viewer to assume that it presents trustworthy information about its subject ([12], p. 129). On the other hand, even when all events portrayed in the documentary took place, the result may still not be reliable. An example of this is when the order in which the events are shown implies a causality which was not necessarily present (see the discussions in [12], p. 129, or in [8] p. 64 about Michael Moore's "Roger and me" documentary).

For interviews, the documentarist must thoroughly research the subject beforehand. What the interviewee says should be checked, and it is the documentarist's responsibility to avoid including errors in the documentary ([8] pp. 72-73). Furthermore, when condensing interview material (or any material, which includes footage of people talking to others on camera) the documentarist risks taking things out of context. In this case, something may honestly seem to mean one thing, but those who were there on the shoot know that it meant something else. Not only verbal statements can be taken out of context. The documentarist can take a scene out of context and use it as evidence to give a false impression, since visual evidence can contradict what is being said in words ([33] p. 64-66).

2.3 Making a documentary

We now examine the process of documentary making to see the steps that are needed to create a documentary. This process goes through three stages, namely preproduction, production and postproduction ([54], pp. 113, 155, 241).

In **preproduction**, all decisions and arrangements prior to shooting are made. For a documentary, this includes choosing a subject, doing background research, assembling a crew, choosing the equipment, and deciding the details and timetable of shooting.

In **production**, the actual material is recorded for later editing. The ratio between material shot and material used in the final documentary is between 3:1 and 60:1 ([54] p. 170). Furthermore, videotape, being less expensive than film, allows overshooting, i.e. to shoot a large amount of footage and make a selection later.

In **postproduction**, the material is edited into a coherent visual statement for presentation to an audience. In this phase, the documentarist discovers what the footage is really about as opposed to what she thinks she shot ([33] p. 98). The main activity at this stage is editing, which we discuss in section 2.3.1. Editing operates by joining shots. To understand editing we therefore explain in section 2.3.2 what a shot is, in section 2.3.3 how shots can be joined together and in section 2.3.4 the rules that assess whether two shots can be joined according to traditional film making. We then examine in section 2.3.5 what kind of documents are used to support the different stages of the documentary making process.

2.3.1 Editing

The **editing process** can be thought of as a series of approximations, each of which brings the documentary closer to the version that will communicate to an audience what the documentarist wants it to. A major task in the editing of a documentary is to cut away the parts that cannot be used or simply that there is no time to use ([33], p. 287). In interview documentaries, cutting away parts implies that some interviews cannot be included in the final version. This can lead to omitting some informative content for the viewer. The different phases in editing are ([33] pp. 99, 293, 306):

- **rough cut** contains the takes which are considered good by the documentarist. The scenes are usually too long and in a provisory order, and there is a lack of visual precision. Nevertheless, the documentary's idea is visible.
- **fine cut** has the desired length and order, awkward cuts have been polished, and sequences have been speeded up—or, occasionally, extended. Titles, music, and narration are added if desired.

2.3.2 Shots

A **shot** is a continuous strip of motion picture film, composed of a series of frames, that runs for an uninterrupted period of time. Shots are filmed with a single camera and are of variable duration. The shot has the following properties: photographic aspects, framing and duration ([12] pp. 229-289).

The **photographic aspects** describe the graphical properties of the shot, which are the *range of tonalities* due to the particular film stock used, the *speed of motion* due to the relation between the rate at which the film was shot and the rate of projection, and the *perspective relations* (the scale of the things disposed in the screen space and the depth of focus) caused by the particular choice of lens, i.e. the focal length.

The **framing** has different aspects: the **frame dimension and shape** indicates the dimension of the screen, and it is often specified by the aspect ratio, the ratio of frame width to frame height. The **onscreen and offscreen space** indicates how the framing divides the space into what is shown and what is not shown. The **angle**, **level**, **height** and **distance** describe the position of the camera with respect to the scene being shot. In interviews the first three are not relevant, since the interviewee is usually shot at a straight-on angle (i.e. not above or below her), the level is straight (perpendicular lines to the ground look perpendicular and not canted) and the camera is at the same height as the interviewee. Different camera distances, however, are used during interviews ([54] p. 186). Camera distances are usually classified, using the human body as a standard measure, as ([12] p. 262):

- **extreme close-up**: this shot isolates details, such as the lips or eyes of a face.
- **close-up**: this type of shot typically exposes the head, hands, feet or a small object. The intention is usually to highlight facial expressions, gestures or particular objects.
- **medium close-up**: a human body is shown from the head down to chest.
- **medium**: a human body is shown from the head to around the waist.
- **medium long**: the subject is framed from around the knees upwards.

- **long**: includes at least the full figure of subjects but the background dominates.
- **extreme long**: the human figure is almost invisible. Used for landscapes, bird's-eye views (e.g. of cities).

Another aspect of framing is specific to cinema and video: the **camera movement**, which produces a change in camera angle, level, height or distance during the shot. Camera movements increase information about the space of the image and give the viewer the impression she is moving around. The most common camera movements are ([12] pp. 267-268):

- **pan**: rotates the camera on a vertical axis, as if scanning horizontally the space.
- **tilt**: rotates the camera on a horizontal axis, as if scanning vertically the space.
- **tracking shot**: the camera as a whole change position, traveling in any direction along the ground—forward, backward, circularly, diagonally, or from side to side. For interviews the most relevant camera movements are forward and backward, which are called **zoom in** and **zoom out**, respectively. Although zoom technically speaking is done using zoom lens and not by moving the camera, it is still considered a camera movement because it substitutes moving the camera backward or forward.
- **crane shot**: the camera moves above ground level, i.e. it raises or descends.
- **shaking shot**: the camera, probably hand-held, is not stable. In amateur films, this parameter can give an indication of the quality of the shot.

Finally, the **shot duration** describes two different aspects: the first is the record time of the shot (i.e. short or long takes). The second aspect is the relation between the record time of the shot and the narrative time elapsed during this period. For example, in film a one minute shot can represent the arc of a day, thus the narrative time is contracted with respect to the record time of the shot. In interviews this technique is normally not used.

2.3.3 Joins

A film editor eliminates unwanted footage and cuts superfluous frames from the beginning and ending of a shot. She then joins the desired shots. Joins between shots are of two main types ([12], pp. 294-295). Firstly, shots can be joined with a **transition**, which can be a **fade** (a fade-out gradually darkens the end of a shot to black, a fade-in gradually lightens a shot from black), a **dissolve** (the end of one shot and the beginning of the next one are briefly superimposed), or a **wipe** (where the next shot replaces the previous one by means of a boundary line moving across the screen). The second, and more common, way of combining shots is the **cut**, which means juxtaposing the last frame of a shot with the first frame of the shot to be joined. While the first type of join is perceived as gradually interrupting one shot and replacing it with another, cuts are perceived as instantaneous changes from one shot to another. In classical filmmaking, fades, dissolves and wipes are often used as *punctuation* shot changes, to signal to the viewer that some time has been omitted from the film, when the director wants to present an action in such a way that it consumes less time on the screen than it does in the story (elliptical editing, [12] p. 308). This technique can also be used in video documentary.

2.3.4 Continuity rules

The viewer should perceive the join between two shots as natural, i.e. she should not feel disoriented by the sequence. Editing in traditional film making strives to ensure narrative continuity, i.e. to create a smooth flow from shot to shot that preserves spatial, temporal and graphical continuity across shots ([12] p. 310). This is called **continuity editing**.

Filmmakers have developed rules to achieve continuity editing. The most important one for **spatial continuity** is the **180° system** ([12] pp. 310-312):

> "*The scene's action—a person walking, two people conversing, a car racing along a road—is assumed to take place along a discernible, predictable line. This axis of action determines a half-circle, or 180° area, where the camera can be placed to present the action. Consequently, the filmmaker will plan, film, and edit the shots so as to respect this center line.* "

If the scene depicts a conversation between two people A and B, the axis of action is the imaginary line connecting these two people. In the 180° system, the camera can be placed anywhere as long as it stays on the same side of the axis of action, so that, for example A is always presented as looking to the right and B to the left. In this way the relative positions in the frame remain consistent, and the two characters seem to look at each other even if only one is in the picture (eyeline match). If the camera would suddenly be placed on the other side of the axis of action, the characters would look as they would have swapped position. If the scene depicts an action, the 180° system ensures consistent screen direction: when someone is moving from left to right and exits the scene to the right, we expect to see her in the next shot entering from the left, and not from the right (unless she went back).

Continuity editing also strives to sustain **temporal continuity**, by presenting the narrative according to its temporal development. Accepted common violations of this continuity are **flashbacks** and, less frequently, **flashforwards**. Flashbacks and flashforwards are signaled by a cut or dissolve ([12] p. 326). Temporal continuity requires the avoidance of **jump cuts**. A jump cut is a cut where the middle section of a continuous shot is removed, and the beginning and ends of the shot are then joined together. A jump cut causes objects to jump to a new position or, in the case of interviews, the interviewee's face to suddenly change expression and her head to suddenly be in a slightly different position. The technique breaks temporal continuity and produces a startling effect. A jump cut can occur when joining two shots that have the same subject but are not sufficiently different in camera distance and angle.

A **graphical continuity** is achieved by matching the purely pictorial qualities of two shots, such as the photographic, framing and camera movement aspects of both shots ([12] p. 297). Photographic continuity implies that the photographic properties of the shots stay constant: same type of film stock, same speed of motion and the same choice of lens, i.e. the focal length. Such continuity is more required in Hollywood-style films than in interview documentaries, since the latter can use archive material or material shot at different times with different photographic properties. Framing continuity is achieved by joining shots only when their framing distance is sufficiently similar ([48] p. 123). For example, an extreme close-up and a long shot should not be juxtaposed. A course-grained camera movement continuity is achieved by not joining shots having the same camera movement type but opposite camera movement directions, such as a pan left and a pan right (as in [56]).

Continuity rules constrain how shots can be edited together. Continuity also has consequences for the content of the shots. Traditional films (or Hollywood-style films) typically require the following continuity aspects for content ([23], p. 183, [48], p. 95):

- **continuity of actor** maintains narrative continuity by using the same actor across scenes. An actor is identified by the visual and acoustic characteristics of her body which distinguish her from all others (such as gender, age, body type, hair color/length, skin color, eye color, voice).
- **continuity of role** maintains narrative continuity by presenting the same role across scenes, usually with the same actor, although not necessarily. An example of this continuity is when a certain actor always dresses as a doctor, or when different doctors appear in different scenes (same role but different actors).
- **continuity of location** maintains narrative continuity by showing scenes as taking place in the same location.
- **continuity of action** maintains narrative continuity by requiring subsequent scenes to develop actions started in previous scenes, for example to show how a car chase ends.

Not all of these content constraints apply to interview documentaries. In an interview documentary it is normal to have more interviewees, interviewed in different locations. Usually no action is performed in addition to the interview. Therefore, of the above rules, only the continuity of interviewee role applies. The viewer expects to see people being interviewed across scenes. Strictly speaking, also the continuity of actor applies. In fact, once an interviewee has been introduced, it would be disorienting for the viewer to see the same interviewee in a later moment in the documentary with a different aspect, without a reason. In this case the continuity of actor rule would be broken. On the other hand, this case is not very likely to happen.

2.3.5 Documents supporting the documentary making process

In traditional documentary making several types of document are used to support the process. During preproduction, concepts, treatments and scripts can be used to describe a documentary in increasing levels of detail. A documentary is based on a **documentary concept**, which describes the reason for making it, what it is about and the effect it should have on the audience ([33] p. 94). The **treatment** is an explanation of the documentary to be made. The treatment states what is to be shot, and why. For many documentaries, the treatment is the basic shooting document. For some genres of documentaries a **script** can also be written. A script is a blueprint of the documentary, as detailed as possible, for shooting and editing. For each scene it tells what is to be shot, how it is to be shot, who is in the scene, and what is said.

In postproduction, before editing, **logs** become important, because they describe the content of the footage, since it would be otherwise impossible to remember where each shot is ([33] p. 98). A log usually contains a description or comment of a scene and a time code indicating where in the footage the scene is. Logging takes three to four times longer than the time duration of the footage ([33] pp. 280-286). A **transcript** can be useful if the audio track provides significant amounts of verbal information, as with interviews. Making transcripts can save work in the editing phase, and does not have to include literally everything that was said. Instead, the topics covered at each stage

of a filmed scene or interview can be summarized or classified into categories. This last solution has the advantage of providing content-based access to the material ([54] p. 244).

2.4 High-level requirements

In this section we discuss the aspects of documentaries that must be modeled by an automatic video generation approach using the concepts and the terms we introduced in this chapter. The high-level requirements PRESENTATION FORM [HLR 1], MONTAGE TECHNIQUE [HLR 4], CONTEXT [HLR 3] and CONTINUITY RULES [HLR 5] represent the answer to *Research Question Documentary Form [1]*, defining the characteristics of documentaries on matter-of-opinion issues that must be modeled. High-level requirements SUBJECT-POINT OF VIEW [HLR 2] and MEDIA-DRIVEN [HLR 6] relate to the characteristics of the automatic video generation model we aim for, as stated in section 1.2.

2.4.1 Presentation Form requirement

We saw that there are three presentation forms for organizing content for a documentary: a narrative form and two non-narrative forms: the categorical and the rhetorical form (sections 2.2.1, 2.2.2 and 2.2.3). Since we focus on **matter-of-opinion documentaries**, the rhetorical form is the predominant way of presenting footage.

Rhetorical form

The rhetorical form is composed by a point of view, positions and arguments. The point of view is the bias the documentarist has in presenting information. Possible points of view are:

- the **propagandist**, showing only the evidence supporting predetermined conclusions
- the **binary communicator**, who gives "equal coverage to both sides" in any controversy
- the **mind-opener**, who presents the complexity of issues, without patronizing or manipulating either the subjects or the audience.

In interview documentaries, the bias has influence on the positions presented. We consider a position as formed by an issue (such as "war in Afghanistan") and an attitude toward that issue, which can be *against*, *for* or *neutral*. A propagandist presents a controversial issue so that one side looks clearly stronger than the other, i.e. either the *against* or *for* side of a certain issue. She can achieve this goal by selecting only interviews that support her position. Even if she selects interviews supporting both positions, there will be more interviews supporting her position, or they will appear more convincing to the viewer than the interviews supporting the other position. The binary communicator strives to present information so that both *against* and *for* positions are represented as equally strong.

In using the rhetorical form, the documentarist presents opposing and supporting positions so that the interviewees seem to react and reply to each other's arguments

(section 2.2.3). Positions are expressed using the types of arguments we introduced in section 2.2.4, i.e. logos, pathos and ethos. By selecting (parts of) interviews, the documentarist can also create arguments based on logos, pathos and ethos.

Automatic approaches lack the complexity and the skills of a human documentarist, and need to follow clear rules. Showing only one side of a story (as the propagandist) or giving equal coverage to both sides (as the binary communicator) is for an automatic approach easier than presenting the complexity of an issue. Therefore, acting as a propagandist or a binary communicator is what we require from our model. Nevertheless, it is the viewer who decides whether the model should act as a propagandist or a binary communicator. By selecting the point of view, the viewer is engaged in the process of documentary making. The result is similar to what a documentarist would be able to achieve by adopting a reflexive point of view, i.e. by showing to the viewer the process of documentary making in the documentary itself, as discussed in section 2.2.6. In this way, a viewer may be able to see something in all its complexity, although our approach does not aim at explicitly modeling the mind-opener documentarist. This discussion leads to the first part of HLR 1.

Narrative and categorical forms

As we discussed in section 2.2.2, a documentarist can use one presentation form at the micro-level, and another at the macro-level. A categorical structure contains groupings of information items, where in each grouping the items are semantically associated to each other with one or more relations that do not hold (or hold to a lesser extent) across groupings. Categories can be used to organize content on the higher level (the macro-level), while within each category content can be presented in an effective and engaging way using the rhetorical form (the micro-level). As an alternative, the narrative form could provide the macro-level structure instead of the categorical form, for example using the *Opening*, *Middle* and *Ending* structure.

The above discussion leads to the following requirement:

HIGH-LEVEL REQUIREMENT 1 *(*PRESENTATION FORM*) A generation model for matter-of-opinion interview documentaries must be able to organize information using the rhetorical form. The following aspects of the rhetorical form must be modeled:*

- *Presenting the material as a propagandist or a binary communicator.*
- *Presenting supporting/opposing positions in relation to each other.*
- *Composing arguments according to logos, pathos and ethos.*

Together with the rhetorical form at the micro-level, the model may use the narrative or the categorical form at the macro-level.

2.4.2 Subject-Point Of View requirement

A documentarist needs to make a selection of footage for the final version of the documentary, with the consequence for the viewer that some information is lost. In our approach, we aim at making all the informative content in the footage directly accessible to the viewer, avoiding the selection and the bias introduced by the editorial decisions of the documentarist. Automatic video generation can be a solution to the loss of information due to a static final version, by allowing the viewer to select what

she wants. The point of view also needs to be specified since there is always a point of view (normally the documentarist's) in a matter-of-opinion documentary. In section 2.2.6 we discussed how a documentary can have different points of view and how the documentarist can express her point of view either directly or using other persons. Since the documentarist can make a position look stronger or weaker by composing arguments and selecting the speakers who support those arguments, we require the automatic approach to provide the user with the same power. This leads to the following requirement:

HIGH-LEVEL REQUIREMENT 2 *(*SUBJECT-POINT OF VIEW*) The model must allow the viewer to specify the subject and the point of view of the automatically generated documentary. Possible points of view are the propagandist's point of view and the binary communicator's point of view. In the case of the propagandist's point of view, the viewer must be able to choose the documentary position with respect to the chosen subject, either* against*,* for *or* neutral*.*

2.4.3 Context requirement

A documentarist can misrepresent reality in two ways: by shooting what has not happened (e.g. by staging a scene) and by presenting what has really happened so that it takes a different meaning (section 2.2.7). An automatic video generation approach is only concerned with the second problem, because it operates on material which has already been shot, and it must assume this material is trustworthy. Although the problem is general, we restrict its range to our field of interest, i.e. interviews. The specific problem then becomes how to model the context of statements so that an interviewee is not misquoted by taking her words out of context:

HIGH-LEVEL REQUIREMENT 3 *(*CONTEXT*) The model must be able to capture the context of statements so that an interviewee is not misquoted by taking her words out of context.*

2.4.4 Montage Technique requirement

Interview documentaries often adopt a single or multiple POV, as expressed by the interviewees. The montage technique vox populi (section 2.2.6) can be used to represent a range of arguments from different people (multiple POV) or to add to a video interview of a particular person the different opinions of the men-on-the-street (single POV). For an automatic video generation point of view, vox populi can thus be used as a presentation model for interviews:

HIGH-LEVEL REQUIREMENT 4 *(*MONTAGE TECHNIQUE*) It must be possible to present interviews with the vox populi montage technique.*

2.4.5 Continuity Rules requirement

To generate well-formed video sequences according to traditional film making practices, we require the model to apply rules of spatial continuity (the 180° system), temporal continuity (avoid jump cuts), graphical continuity (framing continuity and camera movement continuity) and continuity of role (section 2.3.4). Analogously to Auteur [48] and Media Streams [23], the model should aim for a rough cut since a fine

cut is beyond the capabilities of an automatic video generation system, since it would require the description and manipulation of the photographic properties of the shot.

HIGH-LEVEL REQUIREMENT 5 *(*CONTINUITY RULES*) The model must assemble material at the rough cut using spatial, temporal, graphical and role continuity rules.*

2.4.6 Media-Driven requirement

With respect to traditional documentary making (section 2.3), an automatic video generation approach differs mostly in the postproduction phase, because in the editing phase the documentarist releases control to the automatic generation for the selection and sequencing of the video material. Automatic approaches generally focus on editing existing material rather than on shooting it[3]. Therefore, traditional documentary making up to the editing phase does not need to be substantially altered[4].

The disadvantage of an automatic video generation approach is that it requires the documentarist to annotate the footage. Manually annotating material is very time-consuming, and thus an obstacle in adopting all methods where annotations are required. In section 2.3.5 we showed that traditional documentary making already includes some sort of annotation activity, which could be either reused or modified to serve an automatic video generation approach. On the other hand, using an automatic video generation approach a documentarist has the following advantages:

- the documentarist does not need to select and edit the footage, the automatic approach does it for her;
- different documentaries can be generated from the same footage, according to different viewer requests;
- the documentarist can collect material to be used for generated documentaries, without having to specify how this material should be presented to the viewer.

While the first two bullet points are common to all automatic video generation approaches, the last is not. Therefore, we explicitly require it with the following:

HIGH-LEVEL REQUIREMENT 6 *(*MEDIA-DRIVEN*) The documentarist must be able to add material at any time to the repository used to generate documentaries without being required to specify explicitly how to present it to the viewer.*

2.5 Summary

In the first part of this chapter we defined the domain of interview documentaries on matter-of-opinion issues. We pointed out three different types of form, i.e. the **narrative**, **categorical** and **rhetorical** forms. Each of these forms represents a different way of organizing and presenting information in documentaries, and more than one form can be present in a documentary. The narrative form uses a chain of events in a cause-effect relationship. The categorical form groups information by categories. The

[3] An exception to this are the approach of Marc Davis [24] and Brett Adams [1], where the systems also give shooting directives to the user. Another possible approach is to first design the annotations and then shoot the material consequently, as suggested by Lindley [40].

[4] Note that information which is needed by an automatic approach could already be gathered during preproduction and production, as suggested in [49].

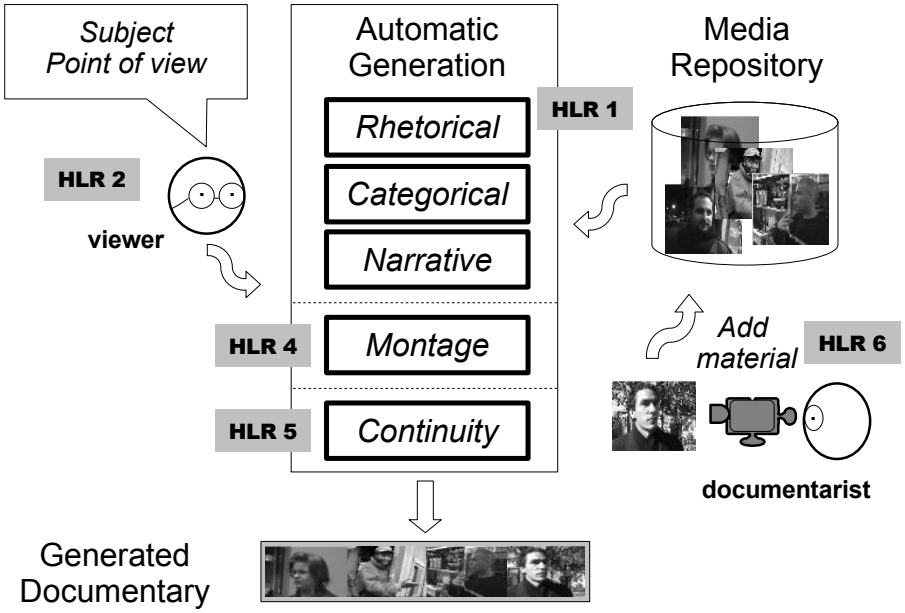

Figure 2.1: A graphical representation of the video generation model as specified by the high-level requirements in section 2.4, except the more abstract CONTEXT [HLR 3]

rhetorical form presents controversial issues using opposing positions and arguments. We explained three different types of arguments, which correspond to **logos**, **pathos** and **ethos** from classical rhetoric. We then introduced **interview documentaries** and how interviews can look more convincing to the viewer. We discuss the concept of **point of view** (POV), the different points of view that a documentary can have and how the documentarist can express her personal point of view in a documentary. We also explained **vox populi**, a single and multiple POV technique that can be used for interview documentaries. We then concluded section 2.2 discussing how objective documentaries are and the problem of **context** in presenting interview statements.

In section 2.3 we explained the process of documentary making. We focused on editing introducing the **rough cut** and **the fine cut**. We then described the properties of the **shot** (in particular the **framing**), the two types of joins (**cuts** and **transitions**) and the rules for **continuity editing**, in particular the **180° system**, the **framing continuity** and the **avoid jump cuts** rules. We concluded the section by showing how annotating the footage is common in traditional documentary making.

In section 2.4, based on the analysis of the domain, we formulated high-level requirements to specify a generation model that:

1. is able to generate documentaries on matter-of-opinion issues which use presentation forms used by documentarists;

2. allows the viewer to select the content of the generated documentary and the documentary's point of view;

3. allows the documentarist to collect material to be used for documentaries, with-

out having to specify how this material should be presented to the viewer.

In figure 2.1 we represent graphically the requirements (except the more abstract CONTEXT [HLR 3]). The high-level requirements PRESENTATION FORM [HLR 1], MONTAGE TECHNIQUE [HLR 4], CONTEXT [HLR 3] and CONTINUITY RULES [HLR 5] introduced in this chapter represent the answer to *Research Question Documentary Form [1]*, defining what characteristics of documentaries on matter-of-opinion issues must be modeled. High-level requirements SUBJECT-POINT OF VIEW [HLR 2] and MEDIA-DRIVEN [HLR 6] relate to the characteristics of the automatic video generation model we aim for. In the next chapter we will examine existing automatic approaches to see whether and how they can be used to satisfy these requirements.

Chapter 3

A model for documentary generation

In the previous chapter we analyzed the domain of documentaries, in particular interview documentaries about matter-of-opinion issues. From this analysis, we defined high-level requirements an automatic video generation model must satisfy to generate documentaries of this type. Here, we examine related work to further specify how such a model should be implemented. We investigate existing approaches that satisfy some of the high-level requirements, and determine what is still missing to satisfy all of them. To implement the rhetorical form we introduced in the previous chapter, we look at approaches that deal with arguments from the domain of argumentation theory, video generation and hypertext. We then examine how to describe video with respect to the structure, granularity and content level of the annotations. Finally, we discuss two other possible presentation forms for documentaries, the categorical and the narrative forms. From this analysis we formulate low-level requirements to guide the implementation of an automatic video generation model that satisfies the high-level requirements we set.

3.1 Introduction

The high-level requirements we set in the previous chapter define salient aspects of the domain of video documentaries that must be modeled by an automatic video generation approach. These requirements specify *what* to model, but not *how* to model it. There are several existing approaches that solve some of the modeling problems raised by the high-level requirements, although no previous work is able to satisfy them completely. In this chapter we examine how to combine different partial solutions provided by related work in a model that corresponds to our requirements.

The approach we take in this chapter is to investigate possible implementations of the PRESENTATION FORM [HLR 1] requirement, and check whether they also comply with the other requirements we set. We start from the rhetorical form (section 3.2) since it is expressed in terms that must be further specified to be modeled, namely positions, points of view and arguments. We discuss approaches that are able to implement (some of) the required aspects of this form, and show how they fail to meet other requirements, especially the MEDIA-DRIVEN [HLR 6] requirement. We then investigate whether the solutions adopted by these systems can be extended to make them suitable

for our model. This investigation leads to specify a video generation model capable of implementing the rhetorical form, in terms of the *data structure* representing the video material and *generation processes* operating on this structure.

In section 3.3 we discuss the properties of the annotations we need for our approach. We first concentrate on the structure of the annotations we need in order to create the above-mentioned data structure. We discuss the properties of different annotation structure by examining the systems that use those annotation structures. We then investigate how to segment video for automatic video generation and at what level the content in video material should be annotated.

The PRESENTATION FORM [HLR 1] requirement specifies two more presentation forms, the categorical and the narrative forms. In section 3.4 we examine whether the model we define for the rhetorical form can also support these presentation forms. We consider systems that implement the categorical and the narrative forms and discuss whether they comply to the other requirements and to our findings so far.

Finally, in section 3.5 we describe, using low-level requirements, the technical characteristics an automatic video generation model should have according to the analysis of related work and our high-level requirements. We define two components: an annotation schema needed to describe video footage and a generation process that manipulates the annotation schema to produce the final documentary. These low-level requirements will be used in chapter 4 and chapter 5 to answer *Research Question Annotation Schema [2]* and *Research Question Generation Process [3]*, respectively.

3.2 Elements for the rhetorical form

The goal of this section is to investigate possible implementations of the rhetorical form that also satisfy the other requirements. The PRESENTATION FORM [HLR 1] requirement specifies the aspects of the rhetorical form which must be modeled, i.e., presenting the material as a propagandist or a binary communicator, presenting supporting/opposing positions in relation to each other and building arguments according to logos, pathos and ethos. These three aspects are interrelated: a propagandist presents only one position or makes one position look stronger than the other, while the binary communicator strives to present contrasting positions with equal strength. Positions are presented using arguments and an argument strength determines the strength of the related position. An argument strength can be evaluated considering whether the argument is effective according to logos, pathos and ethos.

Arguments represent therefore the building blocks for the rhetorical form. In this section we examine related work that has dealt with arguments. Even if we focus on arguments, we also explain how an argument structure can be used to model the other two aspects of the rhetorical form, namely presenting the material as a propagandist or a binary communicator and presenting supporting/opposing positions (HLR 1). Initially, in section 3.2.1 we discuss arguments based on **logos**, since no existing approach has modeled pathos and/or ethos for argument building (this also means that no existing approach satisfies HLR 1). Argumentation studies provide us with a model for arguments and operations to manipulate arguments. We then examine arguments in automatic video editing and hypertext to discuss data structures and processes that can support argument representation and manipulation.

Arguments can also be based on **pathos** and/or **ethos**. As we said above, no existing video generation approach has modeled pathos and ethos for argument building. We therefore examine in section 3.2.2 a different discipline, COGNITIVE THEORY, and

describe a theory of emotions, the **OCC MODEL** [52]. This theory provides us with the means of estimating how effective an argument is on the pathos (section 3.2.2) and ethos (section 3.2.3) planes but, differently from arguments based on logos, does not support the dynamic creation of arguments by juxtaposing clips.

Finally, in section 3.2.4 we summarizes all the elements from previous approaches that we can use to further specify a video generation model for the rhetorical form.

3.2.1 Arguments based on logos

In this section we introduce a model for arguments, the model of **TOULMIN** [71], which allows us to represent the structure of arguments independently from the subject being argued for. Arguments can be represented by a single video clip, e.g. an interviewee arguing a statement, but they can also be dynamically created by juxtaposing clips that express contrasting statements. We therefore examine rules to manipulate arguments from argumentation theory with Verheij's **CUMULA** system ([17], p. 369). To investigate whether a logic-based approach is feasible for our case, we examine argument generation in the domain of automatic video generation with Michael Mateas's **TERMINAL TIME** [44]. As we will show, logic-based approaches require a modeling complexity which is against the MEDIA-DRIVEN [HLR 6] requirement. We therefore discuss other less complex approaches to argument generation: Warren Sack's **SPLICER** [58] in the domain of video generation and **SCHOLONTO** [63] in the domain of scholarly hypertext. Both systems create arguments based on the relations between clips (Splicer) or documents (ScholOnto), and not on a model of the content. Particularly the latter, analogously to Toulmin's approach, aims at modeling the structure of arguments and not their subject, but it is not media-driven, and does not satisfy the MEDIA-DRIVEN [HLR 6] requirement. systemAutomatic link generation [19] can be integrated with ScholOnto's approach, so that ScholOnto's argument structure can be created in a media-driven way, according to HLR 6. This structure can then be manipulated using CUMULA's rules to dynamically create other arguments.

3.2.1.1 The model of Toulmin

The model of **TOULMIN** [71] is commonly used in argumentation studies to diagram the domain independent way an argumentation works. This model is not concerned with the soundness of the argumentation but describes the general structure of rational argumentation, by identifying the different discourse parts used to make a claim and their role.

According to Toulmin, an argument can be broken down into the following functional components (figure 3.1):

- a claim is a statement being argued for, the conclusion of the argument, concerning a potentially controversial issue, for example, "*war is the right solution*"
- the data[1] are facts or observations about the situation under discussion, and are the basis for making the claim, for example, "*we have been attacked*"
- a warrant is the chain of reasoning that connects the data to the claim, usually based more on common sense than on strict logic, for example, "*if you are attacked, then you must react with violence*"

[1] In literature the data are also called the grounds.

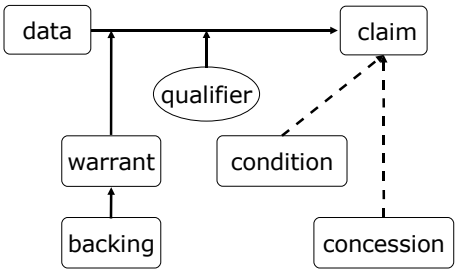

Figure 3.1: The model of Toulmin. Dashed lines connect the claim with the rebuttals parts of the argument.

- a backing is the theoretical or experimental foundation that justifies the warrant, for example, "*waging a war prevents future attacks, since a war damages the opponent and causes a reduced capability to attack again*"
- a qualifier expresses the degree of certainty for the claim, for example "*always, sometimes*"
- rebuttals are possible exceptions to the claim, which can be used as counterarguments to it. They can be:
  - concessions that contradict but are less strong than the claim, for example, "*even though war kills innocent people*"
  - conditions that, if true, could invalidate the claim, for example, "*as long as no innocent people are killed*".

Not all parts need to be explicitly present in an argument. The backing is often implicit in the common cultural background of the speaker and the intended audience. In general, Toulmin is suited for generic argumentation but has been shown to provide insufficient modeling power for specialized fields of argumentation. For example, Newman and Marshall [50] found a number of shortcomings in using Toulmin to model legal argumentation. At the same time, they were able to propose extensions to the model to tackle the shortcomings. Therefore, Toulmin can be extended in cases where the model is not able to capture a particular argumentation type, and this feature makes the model a good design choice. In our case, Toulmin can be used to analyze and encode the arguments used by interviewees when expressing their positions.

3.2.1.2 CUMULA

Toulmin provides only a static representational schema, and does not model how to create an argument. The latter is the focus of argumentation theory, which models the process humans use to arrive at conclusions by disputing them[2]. Argumentation theory uses NONMONOTONIC logic, i.e. a logic where conclusions drawn may be later withdrawn when additional information is obtained. In nonmonotonic logic, the process of

[2] An overview of logic models for argumentation is given in [17].

inferring a conclusion can be defeated by new information, therefore the reasoning is called DEFEASIBLE ([17] pp. 337-338).

Several argumentation models have been defined, as discussed in [17]. Verheij's **CUMULA** ([17], p. 369) provides a group of basic operations to counterargue a conclusion. Verheij considers arguments as being composed of reasons, from which conclusions can be drawn. An argument is then "*reason*, so *conclusion*". If we compare this structure to the model of Toulmin, we see that Verheij's conclusion corresponds to Toulmin's claim, while Verheij's reason is decomposed with Toulmin into data, warrant, backing and qualifier. Toulmin offers, therefore, more representational granularity than the "*reason*, so *conclusion*" structure (as also stated in [7]).

CUMULA's approach is based on formal (non-monotonic) logic to establish the winning argument. A conclusion can be defeated by the following actions:

- **rebuttals**: stating a conclusion opposite to the given one. For example, "*We are attacked. So, we wage a war*" has rebuttal "*Terrorism has economic grounds. So, we do not wage a war*".

- **undercutters**: attacking the connection existing between a reason and a conclusion. For example, "*We are attacked. So, we must wage a war*" has undercutter "*We do not know who attacked us*".

- **sequential weakening**: repeating the steps that lead to the conclusion, which is valid once, but cannot be carried out forever. For example, "*Casualties are normal in a war. So, one more casualty is normal*".

- **parallel weakening**: an argument that contains more than one reason for its conclusion. For example, "*We are attacked. So, we should wage a war*" defeated by "*We do not know who attacked us; war has never solved anything. So, we should not wage a war*".

The first item is analogous to Toulmin's rebuttals, with the difference that rebuttals in Verheij's model defeat the conclusion, while in Toulmin they just represent possible counterarguments to the claim, without invalidating it. This difference is because CUMULA models the process of argumentation, in which conclusions can be defeated, while Toulmin models a static argument, which must contain an undefeated claim (otherwise it would not be an argument). If we apply CUMULA's four defeat actions to the model of Toulmin, we see that *rebuttals* are directed to the claim, while *undercutters* are directed to the parts which support the claim, i.e. data, warrant and backing. The other two actions, *sequential weakening* and *parallel weakening* can be directed to both the claim and the parts supporting it. Sequential weakening is applied when more arguments are directed to the same Toulmin part, while parallel weakening implies that more arguments are directed to different parts of the argument. Toulmin's rebuttals do not need to be defeated, since they already represent counterarguments.

For the rhetorical form we aim to implement, using Toulmin together with CUMULA's four defeat actions can potentially provide a representational schema for arguments and a process to create and compose them. We still need to determine when in the domain of video generation an argument can be formed, i.e. when a defeat action can be applied to one of the parts defined by the model of Toulmin. We discuss this issue with the following two systems.

3.2.1.3 Terminal Time

TERMINAL TIME [44] constructs ideologically-biased historical video documentaries in response to an audience's feedback. Multiple choice questions are posed periodically on the projection screen, and the audience chooses the answers by clapping. The answer generating the loudest clapping wins. After each question, Terminal Time manipulates the presentation of the historical facts in the documentary to mirror and exaggerate the ideological position implied in the audience's answers.

Terminal Time is modeled as a rhetorical narrator, who has *a rhetorical goal* (for example an anti-religious rationalist narrator would have the "show religion is bad" goal) and "spins" the events narrated to support her position. The system follows a top-down approach. It first starts with the rhetorical goal, then it creates a version supporting this goal and finally it presents the version to the public with a generated voice-over narration plus selected historical footage.

In Terminal Time, several rhetorical goals and sub goals are modeled. For example, the goal "show religion is bad" has, among others, the subgoal "show-thinkers-persecuted-by-religion". A goal is satisfied if one of its subgoals is. A test is associated with each subgoal to determine which events contained in a knowledge-base can be used to make the point. In the example, the test associated with the subgoal "show-thinkers-persecuted-by-religion" determines whether an event involves both the creation of an idea system and the execution of the creator, and whether the idea system conflicts with some religious belief system. Once a goal is chosen by the audience, Terminal Time runs the associated tests to select the events. Each rhetorical goal also has an associated rhetorical plan which is designed to present events so that they create an argument supporting the goal. Rhetorical plans *spin* the events by selecting only the details that serve a rhetorical purpose. An example of a plan for the "show-religion-causes-war" goal (subgoal of "show religion is bad") is ([44], p. 39): *Describe the individual who called for the war, mentioning their religious belief — Describe the religious goal of the war — Describe some event happening during the war — Describe the outcome*.

After the events are "spun", Terminal Time puts them in the storyboard. A set of rules called rhetorical devices are then used to sequence the events in order to create narrative connections between them. Events that match a rule's precondition are sequenced before events that match the rule's postcondition. After the sequencing is done, a Natural Language Generation engine generates the narration. Up until this stage, Terminal Time operates purely on textual information. As a last step, video and audio tracks are selected to illustrate the story segment. This search is based on weighted keyword indexing on each video. The clips are shown together with the generated narration.

Terminal Time implements some aspects of the rhetorical form as defined in HLR 1. Terminal Time creates and presents arguments and can generate sequences using the propagandist point of view. The audience can choose (or strongly influence) the content of the presentation and the point of view, as specified in the SUBJECT-POINT OF VIEW [HLR 2] requirement. On the other hand, an important problem of Terminal Time is its extensibility. If a documentarist would want to add new footage to Terminal Time's media repository, describing the event depicted by the footage would not be enough to have it included in a documentary. The documentarist would need to check whether Terminal Time can "understand" the new event: for example, she should check whether there is an existing goal the new item could be used for, whether there is an existing test according to which the item could be selected, whether there is a rhetorical plan

able to put a spin on it, etc. The documentarist would then need to add each missing element to Terminal Time. Ideally, if the system already contained all the knowledge it could possibly need, there would be no problem. In reality, encoding all the potentially required knowledge is not an option for Terminal Time, since the system needs very specific information. A documentarist would probably need to add missing information to the system each time a new media item is added to the repository. If she chooses not to do so, new media material would only be used to represent storylines already encoded in the system, and not to create new ones. Therefore, Terminal Time is not MEDIA-DRIVEN [HLR 6].

For the automatic video generation we aim for, Terminal Time analysis demonstrates that we cannot use a knowledge-base containing domain-specific knowledge and domain-specific rules to interpret that knowledge, since this is against the MEDIA-DRIVEN [HLR 6] requirement. The documentarist must be able to enrich the generation capabilities of the system just by adding new annotated material, without (or with minimal) encoding of new knowledge.

3.2.1.4 Splicer

Terminal Time's complex presentation models require complex knowledge representation. This is not needed when the presentation models are simpler. **SPLICER** [58] allows the user to browse and resequence (i.e. change the montage of) video clips from a database containing congressional hearings on the Iran-Contra case. Splicer implements **montage rules** for automatically sequencing video clips. One rule, called **point-counterpoint**, is built into the system. This rule creates a sequence where the first clip makes a point which is argued by the following clips. Splicer's media repository contains 43 annotated clips, and each video clip is annotated with keywords describing the *speaker* (e.g. "Oliver North") and the *issues* addressed by the speaker (e.g. "aid for Contras"). Three relations (*hastopic*, *group-member* and *adversary*) are used to define 105 Prolog-like facts relating issues to topics, speakers to groups, and groups to other groups. Each speaker belongs to a *group* (e.g. "Oliver North" *group-member* "Administration" ; "Jack Brooks" *group-member* "Congress") and some groups are against each other (e.g., "Administration" *adversary* "Congress"). Furthermore, issues belong to general *topics* (e.g. "war in Nicaragua" *hastopic* "contra-issues", "deal with Khomeni" *hastopic* "Iran-issues").

The point-counterpoint rule is implemented with three Prolog-like predicates and works as follows:

1. A start clip is given as input to Splicer and added to the output video sequence. The issues the clip discusses are retrieved from the annotations, as well as the topic the issues address, the speaker and the group the speaker belongs to.

2. Splicer retrieves a clip that discusses issues belonging to the same topic as the start clip (i.e., same subject), where the speaker belongs to a group that is against the group of the start clip's speaker (i.e., opposing speakers).

3. The clip selected at step 2 becomes the new start clip and the process iterates again from step 1. The process stops when no more clips can be added to the output sequence which are not already present in it.

Splicer's sequencing method described above is based on the assumption that opinions presented in clips selected at step 1 form an argument of the type point-counterpoint

with opinions presented in clips selected at step 2, since the speakers belong to adversarial groups (relation *adversary of*). This point-counterpoint technique could be used to present opposing positions, one of the aspects of the rhetorical form (HLR 1). This mechanism continues to work when new clips are added to the repository, if the new clips are annotated with the speaker and the issues addressed by the speaker. Minimal modifications would be required if the speaker and/or her group are not in the knowledge-base, amounting to inserting the speaker in a group and, if the group does not yet exist, relating the group to the other existing groups (Splicer contains 7 groups). New video clips would therefore enrich the capability of the system to display arguments, without requiring the documentarist to specify how to present it. Splicer satisfies therefore the MEDIA-DRIVEN [HLR 6] requirement. An assumption behind the system, however, is that people belonging to adversarial groups have different opinions, which is not always true. In interview documentaries, especially in the case of a large number of interviewees, it is not always possible to classify every speaker as belonging to a group, and then define whether two groups are adversarial or not. People being interviewed are more likely to form groups depending on the topic being discussed (e.g., people for and against the war in Afghanistan), and people who agree on a particular topic might disagree on another one. Instead, Splicer assumes that people share the same opinions of the group they belong to for the whole range of possible topics. This simplification makes Splicer extensible with new clips or media-driven, but it does not provide the granularity we need for our approach[3].

3.2.1.5 ScholOnto

Splicer shows that focusing on the presentation form (and not on the content, as Terminal Time does) has the potential of making the system media-driven. Therefore, in order to use CUMULA's defeat actions for argument manipulation, we need a structure to relate arguments to each other without formally encoding their content, and at a greater level of detail than Splicer can offer.

Argument structures can be found in **scholarly hypertext**, which focuses on tracking, analyzing or debating concepts in literature from a scholarly perspective, and is, in particular, concerned with arguments as a way to support scholarly inquiries. An example of a scholarly hypertext application is **SCHOLONTO** [63], a tool to support researchers' efforts of contextualizing ideas in relation to existing literature. ScholOnto provides a representation schema that can be used to specify how a concept in a document relates to other concepts in documents. In the schema two entities are specified:

- **concepts** represent some content contained in a document, and are of type *Analysis*, *Approach*, *Problem*, *Software*, etc.
- **relations** between concepts: descriptive relations such as *Addresses*, *Uses/Applies*, *Analyzes* and argumentation relations such as *Proves*, *Refutes*, *Is-consistent-with*, *Is-inconsistent-with*.

When using ScholOnto, a researcher needs to determine the concepts described in a document and to relate each concept to concepts in other documents. For example, "Splicer" in the paper from Warren Sack is a concept of type "Software" and has relation "Addresses" with the concept "automatic video editing" of type "Approach". This

[3] Splicer is not restricted to generate point-counterpoint sequences, since users can define their own montage rules. The limitations we have pointed out do not necessarily apply to other models that could be implemented, although Splicer itself does not explain how to specify such models.

is encoded in the form *concept A - <concept A type / relation / concept B type> - concept B*, which is called a *claim*. In the example, the claim *Splicer <Software / Addresses / Approach > automatic video editing* would be associated to Warren Sack's paper. Different researchers can associate different claims to the same papers and these claims do not need to be in agreement. In this way, researchers can enrich documents with nodes (the concepts) and links (the relations) to form an evolving semantic network reflecting a research community's opinion about relevant literature. This semantic network is used as input to rule-based and visualization tools that help researchers in answering questions such as what is the intellectual lineage of ideas, what is the impact of an approach, or is an approach consistent with its theoretical foundations. ScholOnto's authors assign a *polarity* to the relations used to build a claim [64], indicating with a "+" relations that support a concept (e.g. *Addresses*, *Uses/Applies*, *Analyzes*, *Proves*, *Is-consistent-with*, etc.) and with a "-" relations that contradict a concept (e.g. *Refutes*, *Is-inconsistent-with*, etc.). This polarity is then used to find arguments for or against a particular paper, by following claims with positive or negative polarity, respectively.

ScholOnto does not try to capture the subject matter of a research field, since this is specific to each domain. Instead, the representational schema models the researchers' method of investigating literature, i.e. by making and discussing claims. The assumption is that this activity is common to all fields of research, regardless of the domain. ScholOnto is able to asses whether information is relevant, not by looking at the content, but by examining the relations to other documents: the actual content is transparent to the system and its examination is left to the researcher.

In ScholOnto's approach, annotations must be done manually. Even if automatic techniques could be run on large text corpora with little human effort, with the current state of the art it is impossible to automatically identify the relationships between documents required by ScholOnto. For instance, automatic techniques are able to usefully cluster documents on statistical indexes, but are unable to distinguish whether they are in disagreement, or form a coherent perspective with respect to some common problem, method, or theory ([63], p. 5).

ScholOnto's approach can potentially satisfy some aspects of the rhetorical form (as defined in HLR 1) and the SUBJECT-POINT OF VIEW [HLR 2] requirement, as we show in the following. A researcher can request ScholOnto to show documents that talk about a particular concept she is interested in, together with documents either supporting or counterarguing this concept (or both). ScholOnto's semantic network can then be traversed initially selecting documents that address the concept (relations *Addresses* or *Analyzes*), and then examining the claims associated with these documents. Claims containing relations with a positive polarity, such as *Proves* or *Is-consistent-with*, point to supporting documents, while claims containing relations with a negative polarity, such as *Refutes*, *Is-inconsistent-with*, point to counterarguing documents. These two presentation modes (i.e. either support or counterargue) correspond to showing only one part of the debate about a concept, as a propagandist would do. Showing both supporting and counterarguing documents would make ScholOnto behave as a binary communicator.

If we compare ScholOnto and CUMULA, we see that ScholOnto's argumentation relations (i.e. *Refutes*, *Is-inconsistent-with*, etc.) used in claims between documents correspond to CUMULA's defeat actions between arguments. ScholOnto offers more granularity in the semantics of the argumentation: for example, *Refutes* is a different type of counterargument from *Is-inconsistent-with*. CUMULA does not distinguish between different types of counterarguments, but between different strategies to counterargue. This difference in granularity does not hold when considering only the polarity

of the argumentation relations in ScholOnto. In this case claims composed by relations with a negative polarity are equivalent to a generic counterarguing action, analogously to CUMULA's defeat actions. If we substitute documents with arguments, the relations defined by ScholOnto can then motivate whether a defeat action can take place between two arguments. A claim data structure analogous to ScholOnto's can therefore be used to determine when CUMULA's actions can be applied between arguments.

We can now integrate the model of Toulmin with CUMULA's and ScholOnto's approaches. ScholOnto's claims can be applied between the different parts of an argument as defined by Toulmin, instead of between concepts in documents. ScholOnto's relations can then motivate whether CUMULA's defeat actions can be applied to one of the parts defined by the model of Toulmin. If a part *a* of an argument *A* is linked with a relation of negative polarity to a part *b* of an argument *B*, *a* counterargues *b*. *B* and *a* can be combined to form a new argument.

As explained in section 3.2.1.2, depending on what role the counterargued part *b* plays in the argument *B* (according to Toulmin), *B* and *a* can be combined with a *rebuttal* (if *b* is the claim of *B*) or an *undercutter* (if *b* is the data, the warrant or the backing of *B*). If there are more counterarguing parts for *B*, they can be combined with *parallel weakening* or *sequential weakening*, depending on whether the defeat action is directed to different parts of the counterargued argument *B* (parallel weakening) or to the same part of *B* (sequential weakening).

3.2.1.6 Automatic link generation

ScholOnto's arguments have a finer granularity than Splicer's, since ScholOnto relates each concept to any other concept, without requiring concepts to form groups. On the other hand, in Splicer a clip with a speaker belonging to a particular group is automatically related with a generic argumentation relation such as CONTRADICTS to all clips containing speakers belonging to adversarial groups. ScholOnto requires that each concept in a document has to be linked manually to all related concepts/documents. This is not feasible in our case, where new clips can be added to the repository over time. We cannot expect the annotator/documentarist to find and link all other possible concepts. Therefore, even though ScholOnto is potentially capable of providing several solutions for our video generation approach, it does not satisfy the MEDIA-DRIVEN [HLR 6] requirement.

This limitation can be removed if claims between concepts can be established automatically. Cleary and Bareiss [19] discuss the advantages of **AUTOMATIC LINK GENERATION** with respect to manual linking in the field of question-answering hypermedia systems. To define types for the link between documents, they use a set of eight "conversational associative categories" based on a simple theory of conversation. This theory argues that, at any point in a conversation, there are only a few general categories of follow-up statements that constitute a natural continuation rather than a topic shift. The eight categories are represented as binary alternatives under four broader classes: REFOCUSING: *Context*/*Specifics*; COMPARISON: *Analogies*/*Alternatives*; CAUSALITY: *Causes*/*Results*, and ADVICE: *Opportunities*/*Warnings*. The authors claim that manually creating these typed links is time-consuming and cannot cope well when documents in a repository are added or changed. Therefore, they present different approaches to automatic link generation, evaluating them according to their efficiency in creating typed links versus ease of use for an indexer to annotate the material.

These approaches are (in decreasing ease of use but increasing efficiency):

- **simple concept linking**: documents are annotated with keywords, and links between documents are proposed based on similarity between the keywords describing them.

- **elaborated concept linking**: documents are annotated with two sets of keywords, the first related to concepts mentioned in the document and the second to concepts elaborated in the document. Linking is similar to simple concept linking, except that documents that mention a concept are linked to documents that elaborate on it (and not the other way around).

- **point linking**: parts of documents (text snippets) are annotated with a structure that "captures" the point made by the snippet, of the form [*concept X*] [*relation modifiers*] [*relation*] [*concept Y*], e.g. [*concept:Simulations*] [*relation modifiers:Do Indeed*] [*relation:enable*] [*concept:Learning-by-doing*]. Rules defined by the authors are then used to establish the links and their type.

- **narrative linking**: each document is considered a narrative about deliberate behavior and annotated according to domain-independent slots (for example, *AgentRole*, *Goal*, *Plan*). In the same way as point linking, rules use this information to establish the links and their type.

Cleary and Bareiss claim that the first two approaches still need considerable human intervention to determine if the links created by the automatic process are correct. The third and the forth perform well, but the latter requires more time than manual linking, due to the complexity of the annotation schema.

The point linking technique, being able to produce automatically typed links with a reasonable amount of annotation effort, could be used to create links similar to ScholOnto's claims. In this way, ScholOnto's argument structure can be automatically created, making the approach media-driven, as required by HLR 6. The result of this combination is a domain independent argument structure built in a media-driven way. Furthermore, this structure is able to support the rhetorical form (specified in HLR 1), the SUBJECT-POINT OF VIEW [HLR 2] requirement, and can be integrated with the model of Toulmin and CUMULA's approach, as we saw in section 3.2.1.5.

3.2.2 Arguments based on pathos

Up until now we presented systems that operate only on the logos level. In the following we examine work that can provide insights on how to model the other two rhetorical techniques, pathos and ethos.

Pathos is related to emotions (section 2.2.4). Using pathos in arguments implies solving the following two problems:

- determine which elements in video cause which emotions,

- determine which emotions affect the viewer in such a way that the speaker's argument appears stronger/weaker.

We use a well-established cognitive model, the **OCC MODEL**[4] [52], to describe emotions and determine how they are elicited. In this model, emotions are divided into three general classes: reactions to **events**, i.e. being pleased or displeased because

[4]OCC is the acronym of the authors, Andrew Ortony, Gerald L. Clore and Allan Collins.

Emotion group	**Emotions**
WELL-BEING	*joy*: pleased about an event *distress*: displeased about an event
FORTUNE-OF-OTHERS	*happy-for*: pleased about an event desirable for another person *gloating*: pleased about an event undesirable for another person *resentment*: displeased about an event desirable for another person *sorry-for*: displeased about an event undesirable for another person
PROSPECT-BASED	*hope*: pleased about the prospect of a desirable event *fear*: displeased about the prospect of an undesirable event *satisfaction*: pleased about the confirmation of the prospect of a desirable event *fears-confirmed*: displeased about the confirmation of the prospect of an undesirable event *relief*: pleased about the disconfirmation of the prospect of an undesirable event *disappointment*: displeased about the disconfirmation of the prospect of a desirable event
ATTRIBUTION	*pride*: approving of one's own praiseworthy action *shame*: disapproving of one's blameworthy action *admiration*: approving of someone else's praiseworthy action *reproach*: disapproving of someone else's blameworthy action
ATTRACTION	*liking*: liking an appealing object *disliking*: disliking an unappealing object
WELL-BEING/ ATTRIBUTION	*gratitude*: admiration + joy *anger*: reproach + distress *gratification*: pride + joy *remorse*: shame + distress

Table 3.1: The OCC model

something happened, reactions to **agents**, i.e. approving or disapproving what someone did, and reactions to **objects**, i.e. liking or disliking an object (objects can also be persons). Events, agents and objects constitute the possible eliciting elements for emotions. An emotion is elicited as a reaction of appraising the eliciting element. According to the OCC model, events are appraised as desirable or undesirable with respect to one's own **goals**. One approves or disapproves the actions of an agent with respect to one's own **norms/standards**. Objects are appraised as appealing or unappealing with respect to one's own **tastes/attitudes**.

The three basic emotion classes are further specified in the model (table 3.1). Reactions to events are differentiated into three groups: the first group is called WELL-BEING, and consists of emotions elicited by events that affect oneself. Emotions in this group are *joy* (pleased about an event) and *distress* (displeased about an event). The second group is called FORTUNE-OF-OTHERS, and consists of emotions elicited by events that affect other people. Emotions in this group are *happy-for* (pleased about an event desirable for another person), *gloating* (pleased about an event undesirable

for another person), *resentment* (displeased about an event desirable for another person) and *pity* or *sorry-for* (displeased about an event undesirable for another person). The third group is called PROSPECT-BASED, and consists of emotions experienced in response to expected events and in response of the confirmation or disconfirmation of such events. When an event has not yet happened, the emotions are either *hope* (pleased about the prospect of a desirable event) or *fear* (displeased about the prospect of an undesirable event). When an event has happened, four different emotions can be elicited: *satisfaction* (pleased about the confirmation of the prospect of a desirable event), *fears-confirmed* (displeased about the confirmation of the prospect of an undesirable event), *relief* (pleased about the disconfirmation of the prospect of an undesirable event) and *disappointment* (displeased about the disconfirmation of the prospect of a desirable event).

The group of emotions elicited by approving or disapproving the actions of agents is called ATTRIBUTION. Emotions in this group can be differentiated depending on whether the agent is oneself or another person. In the first case, the emotions are *pride* (approving of one's own praiseworthy action) and *shame* (disapproving of one's blameworthy action). In the second case, the emotions are *admiration* (approving of someone else's praiseworthy action) and *reproach* (disapproving of someone else's blameworthy action).

The group of emotions elicited by liking or disliking objects is called ATTRACTION. Depending whether the object is appealing, the emotions can be generally *liking* (e.g. love, liking an appealing object) or *disliking* (e.g. hate, disliking an unappealing object).

The OCC model further defines one group of compound emotions, reactions to an event caused by the action of an agent. This group is called WELL-BEING/ATTRIBUTION and it consists of *gratitude* (admiration + joy, i.e. approving of someone else's praiseworthy action + pleased about a desirable event), *anger* (reproach + distress, i.e. disapproving of someone else's blameworthy action + displeased about an undesirable event), *gratification* (pride + joy, i.e. approving of one's own praiseworthy action + pleased about a desirable event) and *remorse* (shame + distress, i.e. disapproving of one's own blameworthy action + displeased about an undesirable event).

We now examine how the OCC model can be applied to interview documentaries, which mainly show interviewees answering questions. Our goal is to determine which emotions affect the viewer in such a way that she considers the speaker's argument stronger or weaker. According to the OCC model, the viewer is either pleased or displeased by an *event* that, in the case of interviews, is narrated by the interviewee. A reasonable assumption is that if the viewer is pleased by a narrated event that is desirable for herself (either something already happened, as with WELL-BEING emotions, or something that might happen, as with PROSPECT-BASED emotions) or for the interviewee (FORTUNE-OF-OTHERS emotions), she will consider the presented arguments stronger, or weaker in the case that she is displeased. This assumption would require a user model that describes which events are appraised positively or negatively by the viewer. The video should also be annotated with these events. This approach presents two problems:

- The user model needs to be updated when generating documentaries about different subjects, since the possible events depend on the domain of the subject (for example, events related to war are different from events related to social policies). This is against the MEDIA-DRIVEN [HLR 6] requirement.
- In interview documentaries about matter-of-opinion issues, the focus is on what

the interviewee thinks about an event more than on the event itself. Very often the same events are discussed by the interviewees (for example, the war in Afghanistan), but different opinions are expressed.

Similar conclusions apply to ATTRIBUTION emotions in response to an action of an agent and to WELL-BEING/ATTRIBUTION emotions in response to an event caused by the action of an agent. In interview documentaries about matter-of-opinion issues, the interviewees do not mostly talk about their actions, but about their opinions.

The third eliciting elements in the OCC model are *objects*. In interview documentaries, most of the "objects" are the interviewees themselves (an object can be a person in the OCC model). We can therefore assume that if the viewer likes the interviewee, she will be more inclined to think the interviewee's position is stronger. This requires establishing what the viewer considers as appealing, which is not depending on the domain of the documentary subject but on the physical appearance of the interviewee.

A second element for modeling pathos can be found looking at film theory, in particular at framing and camera position. Although, as Bordwell says ([12], p. 263), no absolute meaning can be assigned *in general* to qualities of framing, in the case of interview documentaries Rabiger states that a central gaze establishes a direct relationship between the audience and the interviewee ([54], pp. 182-184). This relationship is made stronger by a closer distance of the camera to the subject, i.e. by a closer framing. Therefore, we make the assumption that pathos can be determined by cinematic information on the clip, i.e. the framing distance and the gaze direction. An interviewee shot at a closer distance and with a central gaze has more pathos than interviewees shot at longer distances looking off-camera.

3.2.3 Arguments based on ethos

As we stated in the previous section, in interview documentaries about matter-of-opinion issues, the interviewees do not mostly talk about their actions, but about their opinions. A different approach is to consider implied actions. For example, from the way an interviewee talks, a viewer could imply that she has studied and is highly educated. As we saw in the OCC model, actions are appraised based on one's own norms and standards. If a viewer considers education as praiseworthy, she would feel *admiration* for the speaker, and she could be more inclined to think the speaker's position is strong. Analogous to education, other categories related to the viewer's norms and standards could elicit admiration or reproach: type of job, religion, race, gender, etc. The advantage in considering these kinds of stereotypical categories is that interviewees belong to them, regardless of the subject of the documentary. Such categories are domain-independent and can be reused for each documentary without modifications, as required by the MEDIA-DRIVEN [HLR 6] requirement. Recalling that ethos is an appeal to the reputation of the speaker, there is a link between the categories we mentioned and the ethos of a speaker. We assume, therefore, that the ethos of a speaker can be calculated based on the social categories the speaker belongs to and the viewer's attitude to each of these categories.

As one can see, modeling pathos and ethos is based on assumptions, since in literature there is no established relation between particular types of emotions and behavioral inclinations as believing/not believing somebody. According to Ortony [51], emotions only *constrain* the possible inclinations, since the variation per individual is still present.

3.2.4 Conclusions

In this section we examined previous work in order to implement the rhetorical form as specified in the PRESENTATION FORM [HLR 1] requirement. None of the presented systems satisfy this requirement, since they only model logos, but some approaches can be combined to provide a model capable of implementing the rhetorical form.

Arguments are the building blocks of the rhetorical form. In section 3.2.1.1 we discussed the model of Toulmin, a model that represents the structure of arguments, independently from their subject. In this model an argument is broken down in several functional components, namely the claim, the qualifier, the data, the backing, the warrant, the condition and the concession. Our goal is not only to represent static arguments, but also to create them when generating a documentary. We therefore investigated rules to manipulate arguments examining CUMULA (section 3.2.1.2), which provides a set of four defeat actions to counterargue a conclusion. These actions can be related to Toulmin's functional components, so that arguments can be represented statically with Toulmin and combined dynamically using CUMULA's actions.

Since CUMULA's defeat actions require a logic-based approach, we investigated with Terminal Time and Splicer whether logic-based approaches can be used in video generation. We found that Terminal Time implements expressive argument structure, capable of satisfying the rhetorical form in HLR 1 (except for the pathos and ethos part) and the SUBJECT-POINT OF VIEW [HLR 2] requirement, but it does not satisfy the MEDIA-DRIVEN [HLR 6] requirement. Terminal Time is by design not media-driven because it needs complex content information, such as a knowledge-base of historical facts, inferencing rules and rhetorical goals. The system potentially requires the documentarist to provide this information every time a media item is added to the repository. Unlike Terminal Time, Splicer generates video sequences using clips' relations to other clips. These relations are inferred automatically from each clip's annotations using a simple knowledge-base, and this makes Splicer media-driven. On the other hand, Splicer's arguments have insufficient granularity to be applied to general interview documentaries. The analysis of Terminal Time and Splicer reveals that approaches where a knowledge-base is required to describe or interpret clips' content cannot be media-driven, while automatically establishing the relations between clips from the annotations can potentially be media-driven.

We investigated therefore ScholOnto, a system that does not try to capture the subject matter of arguments, but the methods people use in creating them. In using ScholOnto, a researcher relates documents to each other using claims. Claims have a polarity, which can be positive (e.g. concept A *addresses* concept B) or negative (e.g. concept A *refutes* concept B). The resulting argument structure is a graph where concepts in documents are the nodes connected by claims to each other. This structure supports the rhetorical form defined in HLR 1 and the SUBJECT-POINT OF VIEW [HLR 2] requirement. On the other hand, ScholOnto requires the examination of all the documents contained in a repository to manually link all the relevant concepts with claims, which is against the MEDIA-DRIVEN [HLR 6] requirement. Manual linking can be avoided using an automatic linking approach as described by Cleary and Bareiss. They define a method that consists of capturing the content expressed in a document with a sentence-like structure. This structure is then manipulated using rules to create typed links between documents. ScholOnto's approach can be integrated with automatic linking in order to meet the media-driven requirement. The result of this combination is a domain independent argument structure built in a media-driven way. If we substitute documents with arguments, the relations defined by ScholOnto can

Figure 3.2: What is the content of this image: "five individuals in a camper" or the concept of "going on holiday"?

then motivate whether CUMULA's defeat actions can take place between two arguments. ScholOnto, CUMULA and automatic link generation can be combined together with the model of Toulmin to provide a structure for arguments (Toulmin) and a claim data structure (ScholOnto) created automatically (automatic link generation) to support argument manipulation (CUMULA).

In order to model pathos and ethos arguments, we examined a cognitive theory of emotions, the OCC model. Using this model for arguments is based on assumptions, since there are no well-defined relations between particular types of emotions and behavioral inclinations as believing/not believing somebody. We saw that pathos can be modeled based on whether the viewer likes or dislikes the interviewee. Pathos can also be modeled using film theory, namely the camera framing distance and the interviewee's gaze direction. Ethos can be modeled using social categories such as education level, gender and age.

If we compare the approach we took to modeling logos with modeling pathos and ethos, there is an important difference. Logos (as in Terminal Time, Splicer, ScholOnto and CUMULA) can be used to create arguments, while pathos and ethos are used to assess whether a single position (from an interviewee) appears convincing to the viewer. In other words, logos can be used to organize information in a documentary so that supporting/opposing positions can form arguments, and pathos and ethos can be used to assess the perceived strength of each position.

3.3 Video annotations

Automatic video generation systems use descriptions of the media items in order to make decisions about how to create a video sequence. These descriptions are called annotations, or metadata, i.e. data about the data; we use (and have already used in this chapter) the term annotations. In the previous section we concluded that automatically creating argumentation links between documents would satisfy the logos part of the PRESENTATION FORM [HLR 1], the SUBJECT-POINT OF VIEW [HLR 2] and the MEDIA-DRIVEN [HLR 6] requirements. Cleary and Bareiss use a particular annotation structure for their point linking technique, introduced in section 3.2.1.6. In

section 3.3.1 we explore different annotation structures (including that of Cleary and Bareiss's) to investigate which can support automatic linking with the least required annotation effort.

After having investigated the structure of annotations, we look at the content that the annotations must capture, to satisfy high-level requirements HLR 1, HLR 2 and HLR 6, and the CONTINUITY RULES [HLR 5] requirement. When dealing with video, two main questions must be answered: the first is how to segment the content in order to annotate it, because video can be annotated in units as small as a single frame up to the entire video. The second is at what level the content should be annotated, since video can be interpreted at different levels. For example, the frame in figure 3.2 can be analyzed according to how it represents something (e.g. the color histogram, gradient information, etc.) or according to the shapes and forms it represents, or with "five individuals in a camper" or with the concept "going on holidays". In general, annotations do not capture all possible information, because it would be unfeasible and useless: annotations are designed to support a particular task and contain only the information necessary to that task. We examine these issues in section 3.3.2.

3.3.1 Structure of annotations

In section 3.2.1 we concluded that a claim data structure similar to ScholOnto's, created with an automatic link generation approach, can be used to satisfy the logos part of the PRESENTATION FORM [HLR 1], the SUBJECT-POINT OF VIEW [HLR 2] and the MEDIA-DRIVEN [HLR 6] requirements. The goal of this section is to determine what annotation structure is needed for our approach so that an automatic link generation approach can generate a claim data structure. An annotation structure is composed of two parts:

- the structure of the **description** (e.g. the claim structure in ScholOnto),
- the structure of the **values** used to fill the description (e.g. the predefined vocabularies in ScholOnto).

In section 3.3.1.1 we introduce three different types of description structures (keyword, property and relation based) and four different types of value structures (free text, taxonomies, thesauri and ontologies) to represent the range of possible annotation structures.

We then investigate the properties of different annotation structures by discussing systems that use those structures. We group these systems according to the description structure they use: in section 3.3.1.2 systems that use keywords, in section 3.3.1.3 systems that use properties and in section 3.3.1.4 systems that use relations. For each of these groups we discuss whether the annotations would be suitable for our approach.

In this section we focus on annotation structure and not on annotation content. We do not thus restrict our analysis only to video generation systems, but we also consider other types of generation systems, such as presentation, abstract or hypertext generation systems. As a consequence, we do not focus on media specific properties. Here we consider a media item as a representation of a concept, regardless of its media type. Media specific issues will be dealt with in section 3.3.2.

3.3.1.1 Description and value structures

The generation systems we discussed in section 3.2.1 use description structures that are based either on keywords, properties or relations:

- In structure based on *keywords* (**K-annotations**), each item is associated with a list of terms (words) that represent the item's content. The association to the content is unspecified: for example, an annotation consisting of two keywords "*Rembrandt*, *painting*" can indicate that the annotated item represents a painting made by Rembrandt or a painting about Rembrandt. In this sense, *K-annotations* are ambiguous.

- In structures based on *properties* (**P-annotations**), items are annotated with property-value pairs, e.g. *subject-NightWatch*, *creator-Rembrandt*, *date-1642*. Categories allow the disambiguation of cases such as the one abo ve: using *P-annotations*, *Rembrandt* would be either the value of the property *creator* or of the property *subject*.

- In structure based on *relations* (**R-annotations**), items are annotated with property-value pairs as in *P-annotations*, only that some of these values are references to other annotations, e.g. [item *X represents Rembrandt*] *hasOffspring* [item *Y represents Titus*].

Applying this classification to the systems we examined in section 3.2, we see that Terminal Time uses *K-annotations*, Splicer *P-annotations* and ScholOnto *R-annotations*.

When annotating an item, keywords, properties and relations need to be assigned a value. In the case of *R-annotations*, values are a reference to other annotations, while for *K-annotations* and *P-annotations*, values can be chosen from four different types of sources ([6], [62]):

1. **free text** gives the annotator complete freedom to choose the word that better expresses the content. Terms have no relation to each other.

2. a **taxonomy** consists of terms and their hierarchical structure. Each term in a taxonomy is in one or more parent-child relationships to other terms in the taxonomy. A taxonomy does not provide associational relationships between the terms.

3. a **thesaurus** consists of terms and their relationships. Relationships within a thesaurus can be hierarchical (as in a taxonomy) and associational (e.g., term A *is related to* term B).

4. an **ontology** consists of concepts, which have hierarchical and associational relationships, as in a thesaurus. An ontology attempts to define concepts and show the relationships between concepts, whereas a thesaurus attempts to show the relationships between terms ([6], p. 8). Unlike the terms in a thesaurus, concepts in an ontology can have properties and formal constraints on how they can be used together.

These four types of values are organized in a form that becomes increasingly structured, from no structure in free text, up to properties, constraints, hierarchical relationships and associational relationships in ontologies. Using free text implies that there is no check on the semantics of the value: different words can mean the same (e.g. *Rembrandt* and *Rembrandt van Rijn*), as well as the same word can mean different things (e.g. *arms* of the body and *arms* as weapons). The more values are constrained by the structure, the less ambiguous and prone to misinterpretation a single value is, and

the more semantics are associated to it. We call taxonomies, thesauri and ontologies **controlled vocabularies** because values have some constraints because of the structure they belong to.

More semantics for the values comes at a cost. The annotation effort increases when more structure is required: free text is available with no effort; taxonomies, thesauri and ontologies might need to be built for the annotation task. In this case, the building effort increases together with the level of semantics that needs to be defined for each concept. Taxonomies are thus the easiest to build, followed by thesauri and then ontologies.

Having introduced these general concepts about annotations, in the next sections we present related work according to the types of description structure introduced above. We start with *K-annotations* used in ConTour [47] in section 3.3.1.2. We then continue in section 3.3.1.3 with systems that employ *P-annotations*, i.e. SemInf [41] and the Adaptive Abstract Builder [22]. We conclude by examining Disc [29] in section 3.3.1.4, which uses *R-annotations*.

3.3.1.2 *K-annotations*

CONTOUR [47] was developed to support evolving documentaries, i.e. documentaries that could incorporate new media items as soon as they were made. The underlying philosophy was that some stories keep evolving, and so should the documentaries describing them. The system was used to support an evolving documentary about an urban project in Boston[5].

ConTour has a twofold aim: for the author, to provide a framework for gathering content and making it available without having to specify explicitly how, and in what order, the user should view the material; for the user, to support visual navigation of the content.

ConTour allows the author to create and expand the repository by adding material to it. The author is required to attach keywords (called *descriptors*) to each media item. The goal of the descriptors is to capture the complete set of abstract ideas or elements relevant for the documentary story, e.g. names of people or places. Descriptors are created beforehand, with values belonging to the categories of character, time, location and theme. Referring to our classification in section 3.3.1.1, ConTour's value structure can be considered a simple version of a taxonomy that has four top classes to which all values belong.

ConTour allows the user to browse the content by displaying on the screen thumbnails of clips and pictures, plus texts representing the descriptors. Descriptors play the role of concepts in the documentary story. Clips represent particular places and people, interviews with residents and politicians, coverage of events and meetings. In ConTour's annotations, the descriptors are used to drive the presentation of the documentary. Any time a media item is being shown to the user, the associated descriptors become more prominent, or their *activation value* increases. In turn, a descriptor spreads its activation value to media items described by that descriptor. When a clip is finished, ConTour chooses as next clip the one associated to descriptors with the highest activation value, and so on. In general, the active descriptors constitute the context of the story being presented to the user, and the next clip to play is the one that best fits this context. The user can influence the activation values indirectly by choosing to view a particular clip or directly by clicking on a descriptor[6].

[5]The "Big Dig" project was aimed at relieving Boston, MA, from a huge traffic problem caused by an elevated six-lane highway, called the Central Artery, that ran through the center of downtown.

[6]ConTour allows also negative effect, i.e. when the selection of a particular descriptor *decrements* its

Keywords in ConTour relieve authors from the process of defining explicit relationships or links between units of content. Instead, the author connects media items only to keywords. By doing so, the author defines a potential connection between a media item and other media items that share that keyword. Since there are no explicit links between the clips, sequencing decisions are made during viewing, based on the implicit connections via the keywords. Deferring sequencing decisions in this way has as a consequence that the base of content is extensible. Every new media item is simply described by keywords, rather than hardwired to every other relevant media item in the system. In this way, the potentially exponentially-complex task of adding content is managed and requires a constant effort. In this sense ConTour's approach is truly media-driven as specified in the MEDIA-DRIVEN [HLR 6] requirement, and the key factor is that links are created automatically by the system, and not by the author.

On the other hand, by using keywords ConTour can only determine to what degree two media items are related, from unrelated (if they have no descriptors in common) to totally related (if they have the same descriptors). This relation cannot be further specified by the system: is one media item providing further information with respect to the other, or is it contradicting the information presented by the other? This limitation is inherent to keywords, as Cleary and Bareiss also demonstrated ([19], p. 35).

3.3.1.3 *P-annotations*

SEMINF [41] is a system that creates presentations using media items from annotated repositories. SemInf uses repositories from the Open Archives Initiative[7], which are annotated with the Dublin Core (DC) schema [25]. This schema is designed to be very simple in order to facilitate a widespread adoption, with little overhead in annotating the material. It was conceived by and for librarians, and it contains properties for classifying items in a library, e.g. *creator*, *date*, *description*, *format*, *title*.

SemInf's main concern is how to layout media items so that the viewer understands the semantics of the presentation. To determine this, SemInf infers relations between the media items it has to display. For example, if media item *X* is annotated with the property-value pair *DC.creator - Rembrandt* and media item *Y* is annotated with *DC.description - Rembrandt*, SemInf infers that *Y* represents the *creator* of *X*, or, in other words, the relation between the two media items is (person depicted in) *Y creates X*. In this way, a set of 10 relations are inferred of the type *X creates Y*, *X describes Y*, *X colleagueOf Y*, etc. These relations are then translated to spatial/temporal relations in the presentation, driving the layout of the items on the screen. For example, *X creates Y* is translated to *X spatialLeft Y*, causing *X* to be displayed on the left of *Y*.

SemInf shows that *P-annotations* make the process of inferring relations between annotated items possible, although the relations SemInf is able to infer are very simple. This is due to DC's simplicity, since DC was designed to find items in a (digital) library, more than to support presentations about those items. Furthermore, DC annotations use values from free text. As we mentioned in section 3.3.1.1, using free text different words can mean the same, as well as the same word can mean different things. This hinders the inferencing: for example, to determine that *X created Y*, SemInf checks

activation value. In this case the presentation of the material privileges different materials instead of related materials. The authors call the former *breadth-first* exploration, and the latter *depth-first* exploration.

[7] The Open Archives Initiative [36] is a community that has defined a framework, the Open Archives Metadata Harvesting Protocol, to facilitate the sharing of annotations. Using this protocol, data providers are able to make annotations about their collections available for harvesting through an HTTP-based protocol. Service providers are able to use these annotations to create value added services.

whether the condition *Y.creator == X.subject* holds. Ambiguity in the values causes relations to be created incorrectly.

From examining SemInf, we can draw two conclusions. The first is that *P-annotations* allow inferencing relations but not necessarily argumentation relations: the annotation schema must be designed to support arguments. The second is that inferencing needs annotations that use a controlled vocabulary of values, such as a taxonomy, a thesaurus or an ontology. The next system we present has, at least potentially, both properties.

The **ADAPTIVE ABSTRACT BUILDER (AAB)** [22] generates abstracts of video programs, such as symphonies or football matches, by selecting salient scenes. In using AAB, the user specifies preferences for content and a total duration for the abstract. The system then creates a selection of the interesting parts with the given duration.

The annotations the authors define (described in [20]) are modeled after Sowa's Conceptual Graphs [68], and take the form of "somebody (X) does something (Z) to somebody/something (Y)". The resulting structure is:

[ACTION : Z] -		[ACTION : Z] -
(AGENT) ⇒ [X]	or	(AGENT) ⇒ [X]
(OBJECT)⇒ [Y]		(RECIPIENT)⇒ [Y]

For example, a video sequence of the football player P_1 passing the ball to P_2 can be annotated with the Conceptual Graph *[ACTION:pass] - (AGENT)⇒ [Player:*P_1*], (RECIPIENT)⇒ [Player:*P_2*]*. This formal structure provides more descriptive precision than using keywords. With keywords, the same image would be annotated with three values, P_1, *pass,* P_2, giving no information on who performs the action and who is the recipient of the action. Conceptual Graphs allow a more fine-tuned search, for example, retrieve all sequences where *X* is the *AGENT* and *Z* is the *ACTION*. In AAB, each video sequence is annotated with the duration (which is used to calculate whether the generated abstract satisfies the time constraint specified by the user) and with one or more Conceptual Graphs describing the video content. AAB generates abstracts by assessing whether a particular sequence is relevant for the user. The authors use the above-mentioned annotation structure to describe both the video sequences and the user preferences. A matching mechanism is used to assess whether a video sequence is relevant for the user. This mechanism rates the similarity between a video sequence and the user preferences by counting how many equal terms with the same role appear in the annotations. AAB uses free text for the values *X, Y, Z*. In later work, the authors use an ontology [21] to improve the matching mechanisms by removing semantic ambiguities.

AAB focuses on retrieval of relevant video clips and does not infer relations between them, even though the annotation structure would make it possible. For example, AAB could define a rule according to which two video sequences A and B, the first annotated with player X passing the ball to player Y and the second sequence annotated with player Y scoring, should be linked with the relation *enables*, i.e. *A enables B*.

AAB's annotation structure is very similar to the one used by the point linking technique in **Cleary and Bareiss**' approach [19], introduced in section 3.2.1.6. This technique consists of annotating parts of documents (text snippets) with a formal structure that "captures" the point made by the snippet. Rules are then used to establish the links and their type. Points are inspired by the way experienced annotators refer to the content of documents when deciding whether to link them or not. The structure of a point is the following: [*concept X*] [*mode M*] [*sense S*] [*relation R*] [*concept*

Y]. Concepts are defined by the annotator according to the subject of the document, expanding a predefined concept taxonomy which is provided for guidance. The taxonomy includes top-level concepts such as *Attribute*, *Domain*, *Event*, *Object*, *State*. The mode, sense, and relation fields indicate how the concepts in a point relate to each other. According to the authors, there are only a small number of important ways in which speakers (and authors) relate concepts to each other. This is why fixed vocabularies are used for the three fields which deal with how concepts interrelate. The mode field allows an annotator to modify the meaning of a point, and can be *Do*, *Should* or *Can*. The sense field allows an indexer to confirm, negate or invert a point, and can be *Indeed* (leaves the meaning unchanged), *Not* (negates a point) or *Anti* (invert a point, i.e. the opposite of a point holds true). The point language provides a predefined taxonomy of relations, which is not intended to be extended by annotators. The relations it contains are intended to capture the main conceptual relationships which hold between concepts in the points people make, such as causal relations (e.g. enable, explain), interpretation relations (e.g. compare, is-better), temporal relations (e.g. precede), planning relations (e.g. has-function, has-limitation).

When generating a question-answering hypertext, related points (and the text they represent) are linked together by rules. These rules use the semantics defined by the relation, the sense and the mode, to determine whether to create a link and what type of link (Cleary and Bareiss's link types were introduced in section 3.2.1.6). Rules are based on the specific values defined in the taxonomies for the sense, mode and relation properties, while the particular concepts represent the variables. For example, a rule might state that if point A is "*concept:X mode:Do sense:Indeed relation:contain concept:Y*" and point B is "*concept:Z mode:Do sense:Indeed relation:contain concept:Y*", then the documents relative to point A and point B should be linked with link type *alternative*, since they discuss two different concepts containing the same third concept. In the example, point A and point B could be, respectively, "*Workshop Do Indeed contain Learning-by-doing*" and "*Simulation Do Indeed contain Learning-by-doing*".

Points could be used to create automatically typed links which are similar to ScholOnto's claims (described in section 3.2.1.5). For example, if a document A about concept x is annotated with the point "*concept:x relation:uses concept:y*" and a document B about concept z is annotated with the point "*concept:z relation:refutes concept:y*", these two documents could be linked by the claim A *Is-inconsistent-with* B. Points are similar to AAB's Conceptual Graphs. Both annotations have a fixed structure with fixed roles and use a controlled vocabulary of possible values for the annotations, in AAB to remove ambiguity and in Cleary and Bareiss to allow the definition of linking rules. We conclude then that annotation structures such as *points* or *conceptual graph*, together with a controlled vocabulary, serve both the task of describing a video sequence's content and the task of relating (or linking) video sequences to each other.

3.3.1.4 *R-annotations*

ScholOnto, introduced in section 3.2.1.5, is an example of a system using *R-annotations*. ScholOnto uses argumentation relations to make claims between documents.

Using *R-annotations* ScholOnto allows the researcher to specify the relations between concepts in different documents, since *R-annotations* can explicitly reference other annotations. Repositories annotated with *R-annotations* can be represented as a graph whose nodes reference the items (documents in ScholOnto's approach) and whose edges are the relations between them. We call this graph a *Semantic Graph*. In ScholOnto's approach, a semantic graph is created by linking documents with claims.

The *Semantic Graph* can then be traversed to generate presentations composed of the annotated items. This approach is adopted by **DISC** [29], a multimedia presentation generation system for the domain of cultural heritage. This system uses the annotated multimedia repository of the Rijksmuseum[8], to create multimedia presentations.

Disc uses *R-annotations* of the form [media item *X represents Rembrandt*] *hasOffspring* [media item *Y represents Titus*], [media item *X represents Rembrandt*] *hasTeacher* [media item *Z represents PeterLastman*]. The system uses the stereotypical structure of well-established narrative genres, such as biography, to organize the content for a presentation. The content is selected by rules applied to the semantic graph. When an annotation fulfills a rule, the corresponding item is included in the presentation. For example, a biography about an artist typically discuss the artist's teacher, if there was one. To generate a biography about Rembrandt, Disc executes a "teacher" rule that verifies whether the item annotated as representing Rembrandt also has a relation *hasTeacher*, i.e. *Rembrandt hasTeacher z*, where *z* is represented in media item *Z*. If this is the case, as it is in Rembrandt's case with Pieter Lastman, the media item *Z* representing Rembrandt's teacher is included in the presentation, in the section talking about Rembrandt's career.

In both Disc's and ScholOnto's cases, the semantic graph is traversed by rules to generate narratives in the former or present a scholarly argument about a topic in the latter. The semantic graph provides the story space or argument space within which the rules select narrative or rhetorical presentations. The rhetorical form we aim to implement for documentaries (PRESENTATION FORM [HLR 1]) can be based on a semantic graph, providing it contains argumentation links, as in ScholOnto's case. The drawback of manually creating a semantic graph is that all possible combinations of two items in the repository must be examined to check whether the concepts they represent should be related (section 3.2.1.5). This fact makes the initial annotation process, and subsequently each time a new element is added to the repository, particularly cumbersome. Therefore, even though *R-annotations* allow to manually create a data structure which can be used to satisfy the PRESENTATION FORM [HLR 1] and SUBJECT-POINT OF VIEW [HLR 2] requirements, they do not satisfy the MEDIA-DRIVEN [HLR 6] requirement.

3.3.1.5 Conclusions

In section 3.2.1 we saw that a claim data structure such as ScholOnto's, built in a media-driven way with automatic link generation, can be used to satisfy the logos part of the PRESENTATION FORM [HLR 1] requirement, the SUBJECT-POINT OF VIEW [HLR 2] requirement and the MEDIA-DRIVEN [HLR 6] requirement. The goal of this section was to determine what annotation structure should be used to obtain this claim data structure.

In section 3.3.1.4 we defined the claim data structure as a graph of annotations whose nodes correspond to the items in the repository and whose edges are argumentation relations between them. We called this graph a *Semantic Graph*. As Disc and ScholOnto show, this graph can be traversed using rules to generate narrative or argumentation presentation structures. A semantic graph can be given, as in the case of *R-annotations*, or it must be inferred when using other types of annotations. *R-annotations* cannot be used for repositories that can grow with new elements, because of the need to examine all existing items in order to asses whether they should be related

[8] `http://www.rijksmuseum.nl/`

to the new one. *R-annotations* are against the MEDIA-DRIVEN [HLR 6] requirement. *K-annotations*, used in ConTour, do not require the documentarist to specify how the items should be related. On the other hand, *K-annotations* only allow the creation of generic associational links between items, while we need argumentation links for our approach. *P-annotations*, used in AAB and Cleary and Bareiss's point linking technique, provide a means for describing content and are able to support an automatic link generation process, as long as the values are not ambiguous, as in SemInf's case. The value structure needs therefore to be a *controlled vocabulary*, i.e. a taxonomy, a thesaurus or an ontology.

For our approach we conclude that *P-annotations* of the form of *points* or *conceptual graph*, together with a controlled vocabulary, serve both the task of describing an item content and the task of supporting the creation of a *Semantic Graph* for the repository.

3.3.2 Content of annotations

In section 3.3.1 we looked at annotations from the point of view of the structure, with the goal of examining which properties of the structure are required for our approach. In this analysis we did not pay attention to the specific characteristics of the items described by the annotations, making no distinction between texts, images and videos. In this section, we analyze different strategies for annotating video content, dealing with the particular characteristics of this medium. With respect to the requirements used in the previous section, content annotations must also take CONTINUITY RULES [HLR 5] into account, since this requirement is media specific.

Video has two essential features that are relevant for the annotation problem: the first one is that video is a spatio-temporal medium, and video content can be segmented spatially and temporally. A spatial segmentation, as modeled, for example, in the Amsterdam Hypermedia Model ([34], pp. 13, 54) and in HyperCafe [60], would allow the annotation of regions within each frame, in order to describe different aspects that are represented in the video simultaneously. Although interesting, for the domain of video interviews this is not required, since the interview is usually the only aspect that needs to be annotated in a frame. Temporally, any segmentation is possible for the annotations, from annotating each single frame to treating the whole video as a unit (the method used in film archives, for example). Two different temporal segmentation strategies are possible (section 3.3.2.1): **clip-based annotations**, and **stream-based annotations**, introduced in [66].

The second feature of video content is that it can be annotated at five different levels (section 3.3.2.2): the perceptual level, the denotative level, the connotative level, the cinematic level and the subtextual level. We discuss each level of content representation by examining video generation systems that make use of that representation level.

3.3.2.1 Content segmentation

Video content can be temporally segmented into clips. A clip is a set of sequential frames. Clips are parts of a video that are obtained by cutting out "pieces" of it. An annotation can be associated to a clip, and each annotation refers to the whole clip and not to any of its potential subparts.

In a video generation approach, clips are used as the building blocks for the sequence to be generated. The granularity of a clip is a key issue. There are two potentially conflicting requirements. On the one hand, since annotations describe a clip's

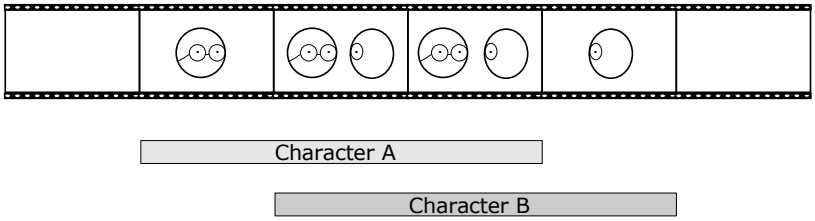

Figure 3.3: In stream-based annotations, annotations exist as independent layers. The two annotations relative to A and B segment the sequence in three parts, based on who is on the scene: A alone, A and B, and B alone

content, the clip must be short enough so that its content does not change, or otherwise the annotations will not apply to the whole clip. On the other hand, clips must be suitable to be dynamically edited together. If a clip is too short, it cannot establish sufficient context for the viewer to understand it. In this case the meaning of the clip is totally dependent on what comes before and after, an effect known as the **Kuleshov effect**. An example of this effect was demonstrated by the following experiment [53]. The following sequence was shown to an audience:

- a long take in close-up of the Russian actor Mozhukin's expressionlessly neutral face — cut — a bowl of steaming soup
- the same face of the actor — cut — a woman lying dead in a coffin
- the same face of the actor — cut — a child playing with a toy bear

When audience members were asked what they saw, they said, that the actor was hungry, then he was sad, then he was happy. The same exact image of the actor's face was used in each of the three short sequences.

If the clip is long enough, it can be self-contained and coherent on its own. This cannot be guaranteed in stream-based annotations, where the video is not cut but each annotation is connected to a particular portion of it by indicating a start point and an end point. The benefit of this approach is that annotations exist as independent overlapping layers (see fig. 3.3). The start and stop points for each annotation do not have to match any notion of a clip or shot boundary, since content can change inside a particular segment or be constant across segments. Stream-based annotations allow a scene to be represented more precisely than clip-based annotations, since the latter apply only to whole clips. For example, in a stream-based approach, a scene with two characters, where each character enters and exits the frame independently, would be annotated with two independent annotations. Each annotation would describe one character, and the start and end moment that character is present. This would indicate how many people are in the scene at any moment: two when the two annotations overlap, otherwise one or none. For this reason, stream-based systems are particularly well suited to the task of "low-level" editing, i.e. editing where a precise level of description is required to determine exactly where the material must be cut.

3.3.2.2 Levels of content

There are several levels for content representation in video. The **FRAMES** project [38], a dynamic video synthesis environment that uses archives of video data, defines five levels at which video can be described:

- the perceptual level involves inherent features of moving images such as line, shape, color, texture, and movement;
- the denotative level (diegetic level) involves elements such as agents (e.g., characters), objects (visual and auditory), actions, and events, all within a particular cultural context;
- the connotative level involves metaphorical, analogical, and associative meanings that attach themselves to events and objects (or to relationships between objects) outside of films, i.e., in culture;
- the cinematic level accounts for the use of formal film/video techniques to produce particular artistic/formal/expressive results;
- the subtextual level: more specialized meanings of symbols and signifiers. This level includes meanings of concern to specialized subcultures, such as those of film theorists, academic critics or Johnny Depp fans.

In the following we discuss all these levels, with the exception of the subtextual level, since it provides a level of specialization we do not need because we do not focus on any particular subculture.

3.3.2.3 The perceptual level

Annotating video content at the perceptual level is used by the **FRAMES** project [39] to support the generation of video sequences according to the **abstract form** ([12], p. 147). Using this form, the film draws the audience's attention to abstract visual and audio qualities of the thing depicted, such as shape, color, rhythm. Objects represented in abstract films are selected for their visual and audio qualities and not for their function in the picture. To generate abstract sequences in FRAMES, single clips are selected based on their similarity with respect to properties at the perceptual level, such as color, shape and texture. Even though it represents a possible way of structuring video sequences, the abstract form is not suited for documentaries, which usually try to convey information more than to elicit viewer reactions to the visual properties of the film. Therefore, this level of content representation does not apply to our approach.

3.3.2.4 The denotative level

Two examples of video generation systems that use denotative content representation are **MEDIA STREAMS** [23] and **AUTEUR** [48]. Media Streams is a system for the representation, retrieval and composition of video data. Media Streams's goal is to enable the reuse of video by supporting the composition of sequences different from the original one. Davis calls this multiuse or **repurposing**. Auteur implements an automated film editing process to realize humorous sequences. The system generates visual jokes of the so-called "slapstick" style, e.g. slipping on a banana-peel.

Both systems use stream-based annotations. Stream-based annotations enable the system to determine what content is represented in the video and when it is represented.

On the other hand, the validity of this representation can be hindered by the Kuleshov effect, which shows that the semantics of video is determined by what comes before and what comes after. Therefore, an application using stream-based annotations must ([23], p. 101):

- determine which aspects of video are invariant and sequence-independent and which aspects are variable and sequence-dependent,
- capture in the annotations only invariant and sequence-independent aspects.

Invariant and sequence-independent aspects represent the **denotative level** of content, while variable and sequence-dependent ones represent the **connotative level** of content. For a stream-based approach, content representation of video should be as objective as possible, in order to allow different connotative meanings to arise according to the context in which the material is presented ([48], p. 75). Content should not be annotated according to the particular semantics in a given stream, but rather using physically-based descriptions. For example, a shot of two people shaking hands should not be annotated as "greeting" or as "agreeing", but rather according to the motion of objects and people in space, e.g. "two people shaking hands". In fact the position of the shot in a sequence could determine the meaning: the shot could mean greeting at the beginning of a sequence or agreeing at the end.

Media Streams focuses on non-verbal action-centric video, with a very minimal audio representation. Media Streams's denotative annotations describe characters, objects, characters and objects actions, relative positions, screen positions, locations, times and weather.

Repurposing in Media Streams means the user can query for specific video content. A query for a video sequence will not only search for that particular sequence but also compose a sequence out of parts from various videos. For example, the user might ask for a sequence of a hammer hitting a nail, and in case this sequence is not present in the repository, Media Streams would look for a sequence of a falling hammer, and a sequence of a nail, and compose them together. Repurposing is equivalent to composing video sequences as a way of satisfying requests to find them, which Davis calls RETRIEVAL BY COMPOSITION. The repurposing offered by Media Streams is user-driven, in the sense that the user, and not the system, determines the meaning of the composed sequence. Media Streams does not try to infer any relations between video sequences from their denotative annotations. In the example of the hammer and the nail, the system is not aware of the relation between these two sequences retrieved to satisfy the query[9]. Instead, in our approach we are required by PRESENTATION FORM [HLR 1] to present supporting/opposing positions in relation to each other. This user-driven composition is not applicable to our case, since the user specifies her request in terms of the subject and point of view of the documentary according to the SUBJECT-POINT OF VIEW [HLR 2] requirement, and not in terms of content of the single video sequences.

Unlike Media Streams, Auteur seeks to infer relations between video sequences to be able to generate a joke. For example, in the case of a shot annotated with *Paul* whose

[9] A similar paradigm of user-driven video creation is implemented by **TALKTV** (`http://www.media.mit.edu/explain/projects/talkTV/`), which allows viewers to search through the Closed Captions of digitized broadcasts for quotes and to extract them. The system returns video clips containing the specified phrase. The user can then use the small video clip as a message, for example as a greeting to others via email or chat programs, but she can also edit the video clips together to create scenes. For example, she can have somebody ask a question and somebody else answer.

action is *to walk*, Auteur's knowledge base contains descriptions of properties related to the action *walking*, for example the location of the actor. Possible locations where the shot can take place are evaluated by comparing the expected objects in a location according to the knowledge-base (e.g. a bench in a park) and the actual objects in the shot. Once the location is identified, conclusions can be drawn about the intention or goal of the actor, for example "Paul walks in a park for pleasure" or "Paul walks through a park in a hurry, taking the park as a shortcut". Auteur considers the meaning of a shot to be determined not only by the shot itself, but also by what comes before and after (what we described as the Kuleshov effect). This implies that each interpretation of a shot is tentative until it can be supported by further shots. In the example above, if the system is not able to retrieve shots supporting the concept of *hurry*, it needs to resort to another interpretation of the shot, for example "Paul walks in a park for pleasure".

Auteur is therefore able to manipulate the content on a connotative level using denotative annotations. An analogous approach could theoretically be used for video documentaries. On the other hand, in a media-driven approach the content is not known beforehand. It is therefore impossible to determine in advance the knowledge required to infer from the denotative aspects of the video its connotative meaning[10]. Analogously to Terminal Time [44] (section 3.2.1.3), Auteur's approach is against MEDIA-DRIVEN [HLR 6].

Furthermore, neither Media Streams nor Auteur take into account audio, which in interview documentaries is predominantly speech. A denotative approach for interview documentaries should determine what the invariant and sequence-independent aspects in speech are: phonemes, words or sentences? Moreover, an approach of this type would require encoding knowledge to compose those aspects in a meaningful way (for example, natural language processing to form sentences from words). To edit sequences with speech would then require defining audio continuity rules, analogously to the visual continuity rules we introduced in section 2.3.4. We are not aware of any previous work dealing with these issues.

We conclude therefore that we cannot (only) use denotative annotations for our approach, since we need to manipulate content on the connotative level. This also implies that we cannot be stream-based, since stream-based annotations need a denotative approach.

3.3.2.5 The connotative level

The majority of the video generation systems we examined in this chapter use connotative annotations: Terminal Time and Splicer in sections 3.2.1.3 and 3.2.1.4, ConTour in section 3.3.1.2, and AAB in section 3.3.1.3. All these systems are also necessarily clip-based. The granularity of the clips is decided a priori, based on the genre of the generated video: interviews for Splicer and ConTour, historical events for Terminal Time, scenes from a football match for AAB.

Clip-based systems can use connotative annotations and are not restricted to capturing only invariant and sequence-independent aspects of video and audio. This is because clips are "self-contained", i.e. their semantics is not totally dependent on their position in a sequence. The advantage of connotative annotations is that content, if required, can be described at a higher level than the description of what can be seen or heard in the video. This is particularly useful when dealing with speech, as in our case,

[10] We do not take into account initiatives as the Cyc project [37] or the LifeNet project(`http://csc.media.mit.edu/LifeNetHome.htm`), which try to capture world-knowledge or common-sense knowledge.

since annotations do not need to capture all the words, but can encode shorter and more abstract summaries.

3.3.2.6 The cinematic level

One of the problems Media Streams and Auteur address is to compose the video material so that the resulting sequence obeys the continuity rules we introduced in section 2.3.4. When two different shots are retrieved to satisfy a user query, Media Streams checks whether the cut from one to the other breaks the continuity rules. If it does, the combination is rejected and another is tried. Auteur checks if the chosen shots can be acceptably joined. Characteristics that must be examined in the shot are the direction of the actor's movements, movements of the relevant body parts, the location, and then shot properties such as framing distance and camera angles. If the join breaks the continuity rules, an alternative shot must be chosen, which fulfills the same narrative role, or the story needs to be revised.

To implement continuity rules, both systems use annotations at the cinematic level, annotating cinematic properties of the video material: shot properties (photographic aspects, framing and camera movement, section 2.3.2) and cinematic transitions (fades, dissolves and wipes, section 2.3.3).

Both systems focus on two types of continuity (section 2.3.4):

- continuity of content, i.e. continuity of actor, role, location and action. Auteur and Media Streams focus on sequences where an action is carried out, often by a character, such as walking in the park or falling from a cliff. Video interviews do not show actions but people talking.

- spatial and graphical continuity, the latter including also photographic continuity (i.e. continuity of the photographic aspects of the shot, such as same film stock used, same speed of motion and same focal length for the lens). In Auteur and Media Streams the viewer should not be aware that the produced video sequence has been created from a repository of shots. The final result should look as the footage was shot and edited to satisfy the user request, just as in traditional films. Video interviews can contain archive material or material shot at different times and places.

As we discussed in the previous chapter, interview documentaries only require continuity of role from the first bullet point and do not require photographic continuity from the second bullet point. Therefore, continuity rules are less strict in the domain of video interviews and in order to satisfy the CONTINUITY RULES [HLR 5] requirement, our annotations do not need to contain all the cinematic properties Media Streams and Auteur have, but just a subset of them.

3.3.2.7 Conclusions

In section 3.3.2 we discussed content segmentation and level of content for annotations. We examined two possible content segmentation strategies: *stream-based* and *clip-based*. We also introduced four levels of content representation for video: the perceptual level, the denotative level, the connotative level and the cinematic level.

We determined that, because we need to deal with audio and video, we need to represent content on a connotative level, and that we cannot infer connotative information

from denotative annotations, as in Auteur's approach. In the case of interview documentaries, where speech is predominant, the use of connotative annotations allows the interpretation or summarization of the sentences on whatever level of abstraction might be needed. On the contrary, denotative annotations would require to be objective and refrain as much as possible from encoding interpretations and/or summaries.

Since stream-based annotations need to be denotative, while clip-based annotations can be (and usually are) connotative, we conclude that we need clip-based annotations. When using clip-based annotations, the granularity of each clip is a key issue. Each piece must be self-contained and coherent on its own. At the same time it must be possible to dynamically edit it together with other related clips to form an argument.

Media Streams and Auteur use cinematic annotations to support continuity editing. Their requirements are more strict than we need for our approach, since their target, Hollywood-style films, has stronger continuity constraints than interview documentaries. Media Streams and Auteur can therefore provide guidance for continuity rules in the domain of video interviews.

3.4 Elements for the categorical and narrative forms

A documentary can be structured according to the categorical or the narrative form. In the PRESENTATION FORM [HLR 1] requirement we stated that narrative or categories can be used to shape the high-level structure of a generated documentary (i.e. the macro-level). In this section we look at existing generation systems to investigate how categories and narrative have been used to shape the presentation of material. Our goal is to select a form that we can apply at the macro-level, which also satisfies the other high-level requirements we set. We discuss the categorical form first, since it is the easiest. This presentation form is alternative to having a story, and a story is desirable for a documentary (section 2.2.1). We therefore investigate whether we can use more complex forms, by examining three types of narrative, in order of complexity: **associative narrative**, **template-based narrative** and **story-based narrative**. Each step is potentially able to provide a better narrative, at the cost of more knowledge required.

3.4.1 Categorical form

Categories are groupings that individuals or societies create to organize their knowledge of the world (section 2.2.2). **TOPIA** [57] makes use of categories to generate multimedia presentations from annotated repositories.

Topia uses *P-annotations* and the categories correspond to the properties in the annotations used to describe each media item. For example, annotations for museum artifacts may contain a property named *creator*. Several media items representing paintings could have this property with value *Rembrandt*. All these paintings belong to the category "paintings made by Rembrandt". The system then uses a clustering technique on the other properties contained in the annotations to further structure the items into a hierarchy of categories. This hierarchical structure is used to present the items to the viewer. Since the clustering technique does not depend on particular properties or particular values, Topia can use any repository using *P-annotations* to generate presentations.

Topia shows that the categorical form can be used without the need for additional knowledge, because categories can be based on the properties of the annotated material. This form can therefore be used in our approach.

3.4.2 Associative narrative

In using associative narrative, a system assembles sequences of clips which are associated (in some form) to each other and/or to the viewer's interests. The specific nature of the association is transparent to the system, which only measures how much two clips are associated with each other and edit the sequence accordingly. Very often this form of narrative is used by *K-annotation* systems, and the degree to which two clips are associated is calculated using the number of overlapping keywords. Two examples of this are ConTour [47] (section 3.3.1.2) and the **KORSAKOW SYSTEM**[11], a nonlinear and interactive documentary generation system. The system arranges the screen so that it shows one main film part and up to three follow-up clips in a preview window. The user can decide which one to follow. As in ConTour, follow-up clips are chosen by the system looking at the keywords with which the clips have been annotated, where the greater the number of overlapping keywords the stronger the relation between clips is.

An associative narrative can also be generated using *P-annotation* systems, which use more complex rules to determine the degree of association, as in Lev Manovich's **SOFT CINEMA**[12], a system that edits movies in real time by selecting media items from a database. Analogously to the Korsakow System, Soft Cinema presents the screen divided into several regions where different media are played. During playback, individual media items are assigned to the regions according to rules. These editing rules are written by an author[13] and select and sequence clips according to the clip content and shot properties. A rule can sequence clips based on whether their content annotations are analogous or different (e.g. select clips that are annotated with *location:interior of a pub*) and/or on whether their shot properties are analogous or different (e.g. select clips that are annotated with *camera movement: pan left*).

Soft Cinema and the Korsakow System both show media in a way that stimulates the viewer to make sense of what she is seeing. In this scenario the viewer's role is a very active one, continuously trying to construct meaning out of what she sees, which allows both systems to implement only simple editing rules and leave the viewer free in her meaning-making endeavor. An associative narrative does not require the definition and the exploitation of complex relations between media items and could be implemented using the properties of the annotated material. On the other hand, associative narrative is best suited for artistic purposes, and Bordwell ([12], p. 154) describes how this type of narrative (called associational form) is used in experimental films. These films draw on a poetic series of transitions between groups of images that may not have immediate logical connections, but, because of the juxtaposition, invite the viewer to find some.

3.4.3 Template-based narrative

As early approaches to video generation showed (see for example the first experiments in Train of Thought [32]), sequencing video clips according to associative narrative did not produce engaging stories. Since there is no model of what a coherent and engaging story is, one can only hope that the story is coherent and engaging. To overcome this limitation, video generation systems looked at narrative studies with the goal of building a story model. Such models can be roughly divided into two main approaches:

[11] `http://www.korsakow.com/ksy/index.html`

[12] `http://www.softcinema.net/`

[13] Soft Cinema can be considered an authoring tool for editing movies, while the role of the viewers is purely passive and consists in watching the edited movies.

trying to capture the **structure of stories** and trying to capture the **logic of stories**. In this section we examine the first approach, leaving the second one for the following section.

Modeling the structure of stories is based on the assumption that stories are composed of an ordered list of different phases. For example, **AGENT STORIES** [15] is an authoring environment for metalinear narratives, i.e. narratives composed of small related story pieces designed to be arranged in different ways, to tell different stories from different points of view. Agent Stories uses a story structure based on Branigan's work [13], composed of *introduction*, *character introduction*, *conflict*, *negotiation*, *resolution*, *diversion* and *ending*. Each phase has particular properties, for example the introduction presents the subject of the story and its environment.

An interesting feature of Agent Stories is that the system provides **feedback functionality** to support the author. This consists of an environment where the writer can apply different story styles to see which stories are generated. An automatic generation approach should provide some feedback to the author (the documentarist in our case) about how the final sequences look, so that the author can revise the process in case the quality of the generation is unsatisfactory.

Two main techniques are used to place items in a story structure: the first is to annotate each item with the particular story phase for which it is suited, while the second is to establish rules that determine in which story phase an item should be placed.

The first approach is used by, for example, Agent Stories and **TRAIN OF THOUGHT** [32], a story engine for multi-threaded video stories. Train of Thought presents the user with multiple versions of the same love story, selecting and sequencing video clips contained in an annotated database. Assigning a narrative role to media items has two drawbacks: the first is that it can be difficult to asses the role a media item should play, since it might depend on the story. The second is that the use of the clip can potentially be limited to play only the role the author determined at authoring time.

Disc [29] (section 3.3.1.4) uses rules to determine in which story phase an item should be placed. Rules in Disc are based on the concept of actants from Greimas [30]. Actants are stereotypical characters or roles that recur in each narrative of a particular genre; Greimas analyzed fairy tales where the recurrent characters are, for example, the *subject* (the hero of the story) or the *opponent* (an obstacle to the hero's mission). Based on this idea, Disc allows the author of a presentation to specify the possible characters for each genre, such as biography.

Using rules allows more flexibility in the generation of narratives, since the role an item plays need not be determined beforehand at authoring time. The system can determine at run-time which items fulfill which role in the presentation. On the other hand, rules in Disc use domain concepts and characters specific to a genre (biography in Disc's case). In general, (non-trivial) rules need to be based on a domain model and a story model. The applicability of this approach to our case is therefore analogous to the applicability of a story-based narrative.

3.4.4 Story-based narrative

Story-based approaches try to capture the **logic of stories**. This logic is encoded in a story plan, executed by some planner. One of the components of a story plan is a story structure such as the ones used in template-based narratives (section 3.4.3). The difference is that the story planner composes this structure so that the resulting story is **coherent**, i.e. every clip is chosen according to an intent. Therefore, at each moment in

the story the choice of which clip to insert depends on the previous clips. Characters' actions should be in line with their own previous actions [70]. Related work in the previous section practically ignored this direct dependency.

Terminal Time [44] (section 3.2.1.3) is an example of a story-based video generation system. Terminal Time uses a rhetorical plan as the story plan. Another example is Nack's Auteur [48] (section 3.3.2.4). Auteur uses a model of its domain, i.e. *humor*. This model provides the strategies to achieve humor in a story, by encoding concepts such as inappropriateness, paradox, dissimilarity or unexpected outcome of actions. Finally, **IDIC** [59] generates video sequences according to a story plan by selecting appropriate segments from an archive of annotated clips from trailers of "Start Trek: The Next Generation", a space adventure story. According to the authors, the montage of video sequences could be done using inferencing rules based on common-sense knowledge. This would require sufficient common-sense knowledge to be encoded to cover all possible montage sequences.

Story-based narrative approaches need a model of the domain (e.g. rhetorical goals in Terminal Time, humor in Auteur and space adventure stories in IDIC), and a knowledge-base representing common-sense knowledge (in Auteur and IDIC), or more specific knowledge (rhetoric in Terminal Time). Since in our case the repository can grow with new items, we cannot foresee what knowledge will be required to reason on items which have not yet been added. Building a knowledge-base is not feasible in our approach, which is required to be media-driven (HLR 6).

3.4.5 Conclusions

The goal of this section was to determine whether the categorical or the narrative form can be used to provide a *macro-level* structure for our automatic video generation approach.

The categorical form (section 3.4.1) does not require to encode additional knowledge because categories can be based on the properties of the annotated material, as we saw with Topia. This form can therefore be used in our approach.

Associative narrative (section 3.4.2) could be used at the macro level, since it does not require the definition and the exploitation of relations between media items and could be implemented using the properties of the annotated material. On the other hand, this type of narrative is more suited for artistic films than for documentaries about matter-of-opinion issues, and so we choose not to use it.

Template-based narratives (section 3.4.3) can give a structure to the generated stories and could be used in our approach to organize the documentary on the macro level. On the other hand, there is no suitable way to assign items to the different phases other than annotating the role each item should play in the documentary. This requires the documentarist to determine what role each clip could play, without knowing the stories that the system can generate. Such stories can then increase in number when new items are added to the repository (as required by the MEDIA-DRIVEN [HLR 6] requirement). Therefore, we conclude that template-based narratives cannot be used in our case.

Story-based narratives (section 3.4.4) provide coherence to generated sequences, but require additional components: a model of the domain, a story planner and a knowledge base representing common or domain-specific knowledge. Such approaches are feasible for a closed repository, but unfeasible when the repository can grow with clips whose content is unknown beforehand. Given that a closed repository does not satisfy the MEDIA-DRIVEN [HLR 6] requirement, story-based narratives cannot be used in our case.

A feedback functionality is essential in automatic generation systems (section 3.4.3), since the author releases control to the system over the final result and cannot always envision what the viewer will see. In the approaches we saw (with the exception of Agent Stories), authors have to take the role of viewers to check the outcome. The author, however, needs to get some indication of the stories the system will generate without having to view them all.

3.5 Low-level requirements

Based on our analysis of related work, we now derive low-level requirements in order to implement an automatic video generation model for documentaries. We divide this model into two parts: an **annotation schema** needed to describe video footage and a **generation process** that manipulates the annotations to produce the final documentary. In section 3.5.1 we define requirements for the annotation schema, and in section 3.5.2 for the generation process.

3.5.1 Requirements for the annotation schema

3.5.1.1 Logos Argument requirement

According to the MEDIA-DRIVEN [HLR 6] requirement, the structure of arguments in the rhetorical form (PRESENTATION FORM [HLR 1]) must be modeled in a general (i.e. domain-independent) way. In order to do so, argumentation theory provides the model of Toulmin for arguments built using logos.

LOW-LEVEL ANNOTATION REQUIREMENT 1 *(*LOGOS ARGUMENT*) The annotation schema must capture the form of arguments, which is independent from the subject matter, using the model of Toulmin.*

3.5.1.2 Pathos Argument requirement

Besides logos arguments, also pathos and ethos arguments must be modeled (HLR 1). Two elements can be used to evaluate the pathos of an interviewee: the first is motivated by the OCC model and is based on whether the user likes (finds appealing) the interviewee. The second is motivated by film theory and is based on cinematic characteristics of the clip, i.e. the framing distance and the gaze direction (section 3.2.2).

LOW-LEVEL ANNOTATION REQUIREMENT 2 *(*PATHOS ARGUMENT*) The pathos value of an interviewee can be determined on the basis of:*

- *whether the interviewee is appealing to the viewer.*
- *the framing distance of the video clip and the gaze direction of the interviewee*

3.5.1.3 Ethos Argument requirement

Arguments using ethos can be selected and evaluated using the OCC model with respect to the social categories the interviewee belongs to and the user attitude towards those categories (section 3.2.3).

LOW-LEVEL ANNOTATION REQUIREMENT 3 *(*ETHOS ARGUMENT*) The ethos value of an interviewee can be determined on the basis of the social categories the interviewee belongs to and a user profile determining how important for the user these categories are.*

3.5.1.4 Annotation Structure requirement

In section 3.3.1.4 we saw that a semantic graph where the nodes represent the media items and the edges are argumentation relations (such as in ScholOnto's approach) can be used to present material according to the rhetorical form (PRESENTATION FORM [HLR 1]). This graph must be generated automatically because of the MEDIA-DRIVEN [HLR 6] requirement. The investigation of annotation structures in section 3.3.1 revealed that *P-annotations*, i.e. annotations based on properties, together with a controlled vocabulary, can support the automatic generation of a semantic graph. Particularly, fixed sentence-like structures as in AAB's Conceptual Graph and Cleary and Bareiss's points (section 3.3.1.3) can describe the content of a clip or document AND support the generation of a semantic graph.

LOW-LEVEL ANNOTATION REQUIREMENT 4 *(*ANNOTATION STRUCTURE*) Annotations must use a structure composed of **P-annotations** and a **controlled vocabulary**. In particular, media items must be described with a fixed sentence-like structure of the form <**subject**> <**modifier**> <**action**> <**object**>.*

3.5.1.5 Annotation Content requirement

In the case of interviews where speech is important, connotative annotations offer clear advantages with respect to pure denotative annotations. We cannot adopt a stream-based approach since it would require us to use only denotative annotations, ultimately violating the MEDIA-DRIVEN [HLR 6] requirement (section 3.3.2). Therefore, we adopt a clip-based approach with denotative and connotative annotations. We also need to encode the cinematic level of content to support the CONTINUITY RULES [HLR 5] requirement.

LOW-LEVEL ANNOTATION REQUIREMENT 5 *(*ANNOTATION CONTENT*) Annotation must be clip-based and must model content on the denotative, connotative and cinematic level.*

3.5.2 Requirements for the generation process

3.5.2.1 Graph Generation requirement

As we mentioned when motivating the ANNOTATION STRUCTURE [LLR 4] requirement, a semantic graph such as the one used by ScholOnto can be used to present material according to the rhetorical form (PRESENTATION FORM [HLR 1]). However, in a media-driven approach (as specified in the MEDIA-DRIVEN [HLR 6] requirement), this graph must be automatically generated from the annotations and not manually created as in ScholOnto's approach.

LOW-LEVEL PROCESS REQUIREMENT 6 *(*GRAPH GENERATION*) The generation process must create, using inferencing from the annotations, a semantic graph where the nodes correspond to the media items and the edges are argumentation relations*

having one of two possible polarities, positive (i.e. supports) or negative (i.e. contradicts).

3.5.2.2 Argument Composition requirement

In section 3.2.1.2 we saw that arguments can be composed together using CUMULA's four defeat actions, i.e. *rebuttals*, *undercutters*, *sequential weakening* and *parallel weakening*. These actions provide means for attacking an argument.

LOW-LEVEL PROCESS REQUIREMENT 7 *(*ARGUMENT COMPOSITION*) The generation process can compose arguments together using actions such as **rebuttals**, **undercutters**, **sequential weakening** and **parallel weakening**.*

3.5.2.3 Quality Feedback requirement

The system should provide the documentarist with an indication of the quality of the documentaries it can potentially generate (section 3.4.3).

LOW-LEVEL PROCESS REQUIREMENT 8 *(*QUALITY FEEDBACK*) The generation process must provide feedback to the documentarist about the quality of the documentaries it can generate.*

3.5.2.4 Categorical Form requirement

From the analysis of categorical and narrative forms in section 3.4 we concluded that, while the rhetorical form must be used at the micro-level, the categorical form must be used at the macro-level.

LOW-LEVEL PROCESS REQUIREMENT 9 *(*CATEGORICAL FORM*) The generated documentary must use the categorical form to assemble content on the macro-level.*

3.6 Summary

The analysis of our target domain in the previous chapter resulted in the definition of high-level requirements. These requirements are on a more general level than that needed to guide the implementation of a generation model on a computer. To bridge this gap, we looked in this chapter at related work with two goals: to position our work with respect to previous research and to determine which existing approaches can be used for our purposes.

We first examined how previous work can provide a model for the rhetorical form (section 3.2). From this analysis we saw that no existing approach can satisfy all the high-level requirements we set in the previous chapter, but many research efforts have tackled issues we also have to deal with. No system has dealt with modeling pathos and/or ethos for the generation of arguments, as needed for the PRESENTATION FORM [HLR 1] requirement. Our solution was to look at a cognitive theory of emotions, the OCC model in section 3.2.2. Such a theory provides guidance to assess an interviewee's ethos based on the social categories she belongs to. Pathos can be assessed based on whether the interviewee is appealing to the viewer. Another way to assess pathos is using two shot characteristics, the framing distance and the gaze direction.

On the contrary, logos has been modeled in several existing approaches. In order to model arguments according to their form and not to the subject matter, we adopted

the model of Toulmin for representing arguments (section 3.2.1.1) and CUMULA's four defeat actions to compose arguments together (section 3.2.1.2). We then studied argument composition by discussing generation systems that have used arguments, i.e. Terminal Time and Splicer in sections 3.2.1.3 and 3.2.1.4, and ScholOnto in section 3.2.1.5. We adopted ScholOnto's data structure, which does not capture the subject matter of a field, since this is specific to each domain, but models the method used to create arguments. This data structure is a semantic graph representing the items in the repository. In ScholOnto this graph is generated manually, which does not satisfy the MEDIA-DRIVEN [HLR 6] requirement. In our approach this graph must be created automatically using an automatic link generation technique (section 3.2.1.6).

We then examined how to annotate video material (section 3.3). The goal of automatically creating a semantic graph drove the investigation of an annotation structure able to support this task (section 3.3.1). We found that annotations based on properties and on a controlled vocabulary for the values can support the process of graph creation (section 3.3.1.3). We then examined the problem of annotating video content (section 3.3.2). Video annotations differ with respect to segmentation strategies (clip-based versus stream-based approaches) and level of content representation (perceptual, denotative, connotative and cinematic). For our approach, where speech is predominant, denotative annotations cannot be used, being against the MEDIA-DRIVEN [HLR 6] and the SUBJECT-POINT OF VIEW [HLR 2] requirements. We need connotative annotations and therefore we have to adopt a clip-based approach. We also need cinematic annotations to support the CONTINUITY RULES [HLR 5] requirement.

The PRESENTATION FORM [HLR 1] requirement specifies that the rhetorical form must be used at the micro-level. We therefore looked at what presentation forms can be used for the macro-level. We chose the categorical form, since other forms are not suited for documentaries or break the MEDIA-DRIVEN [HLR 6] requirement. As a last point, we observed that when using an automatic video generation process, the documentarist cannot foresee all the possible video sequences generated by the system. Author feedback is essential. The author needs some indications about the stories the system will generate, without having to view all potential outcomes.

The analysis performed in this chapter led us to formulate low-level requirements for an automatic video generation model. Two components are specified by these requirements, an annotation schema (implemented in chapter 4 as the answer to *Research Question Annotation Schema [2]*) and a generation process (implemented in chapter 5 as the answer to *Research Question Generation Process [3]*).

Chapter 4

The annotation schema

In the previous chapter we defined the automatic video generation model we aim for as composed of two components: an annotation schema and a generation process. In this chapter we define the annotation schema. The annotation schema models positions, arguments and statements contained in video to implement the rhetorical form at the micro level, and categories, such as question asked, to implement the categorical form at the macro level. Furthermore, cinematic properties of video, such as framing distance, are modeled to support continuity editing. In the next chapter we define the other component of the model, the generation process, which uses the information encoded in the annotation schema to generate documentaries. The definition of an annotation schema in this chapter provides an answer to *Research Question Annotation Schema [2]*. This chapter is based on [10] and [11].

4.1 Introduction

In this chapter we specify how video content must be encoded in an annotation schema to capture the information needed for our automatic video generation approach. This annotation schema, together with the generation process we introduce in the next chapter, satisfies all the high-level requirements we set, although this can only be fully understood when we explain how the generation process uses the information encoded in the annotations.

The design of the schema is mainly driven by the need to implement the two presentation forms specified in HLR 1, namely the rhetorical form and the categorical form, as well as the need to support continuity editing as specified in HLR 5.

Two low-level requirements, the ANNOTATION CONTENT [LLR 5] requirement and the ANNOTATION STRUCTURE [LLR 4] requirement, explicitly specify what properties the annotations must have. According to the former, annotations need to be clip-based and capture three levels of content: the denotative, connotative and cinematic levels (section 3.3.2.2). The latter specifies that we need to use *P-annotations* and a controlled vocabulary, and describe clips with a fixed sentence-like structure.

In order to use our approach, the video material needs to be examined and annotated. This is typically done by the documentarist, but it could also be done by someone else. We therefore use in the discussion the term “annotator”.

The structure of the chapter is as follows. The first two sections satisfy the PRESENTATION FORM [HLR 1] requirement: in section 4.2 we examine what information

Media type	Information conveyed	Perception channel
image	Non Verbal	visual
video	Non Verbal	visual
writing	Verbal	visual
noise	Non Verbal	auditory
music	Non Verbal	auditory
speech	Verbal	auditory

Table 4.1: Relation between media types, information conveyed and perception channel

is needed to implement the rhetorical form, while section 4.3 is dedicated to the categorical form. In section 4.4 we examine the cinematic properties of video to support continuity editing as specified in the CONTINUITY RULES [HLR 5] requirement. In section 4.5 we summarize the properties of our schema and we show how it is used to annotate video material.

4.2 Rhetorical form annotations

The goal of this section is to define which content contained in video must be annotated, and how it must be annotated, to implement the rhetorical form as specified in the PRESENTATION FORM [HLR 1] requirement. The elements of this form are points of view, positions and logos, pathos and ethos arguments. Modeling positions and arguments allows to also represent points of view, as we will show in chapter 5. Therefore, points of view do not need to be explicitly modeled. How arguments must be modeled is further specified by the low-level requirements LOGOS ARGUMENT [LLR 1], PATHOS ARGUMENT [LLR 2] and ETHOS ARGUMENT [LLR 3]. The ANNOTATION CONTENT [LLR 5] requirement specifies the levels of content representation we can use, namely the denotative, connotative and cinematic levels.

Content in video is conveyed by the video track and in the audio track, which are processed by the visual and auditory perception channels, respectively. Metz identifies six media types in video ([69]):

- **visual channel** (video track):
  - **image** (photographic image)
  - **video** (moving photographic image)
  - **writing** (credits, intertitles, subtitles, written materials in a shot)
- the **auditory channel** (audio track):
  - **noise** (recorded noises)
  - **music** (recorded musical sound)
  - **speech** (recorded phonetic sound).

These media types can contain two types of information:

- Verbal: information conveyed by language (e.g. speech, writing)

- Non Verbal: all other information (e.g. noise, music, video).

In Table 4.1 we summarize the relation between types of information, perception channels and media types. In order to model the elements of the rhetorical form, we need to determine which media types contain the content relevant for arguments and positions. Having determined where the relevant information is, we have to define how to model it. For this task, ANNOTATION STRUCTURE [LLR 4] specifies that we must use *P-annotations* and a controlled vocabulary.

We first model arguments based on logos (section 4.2.1), pathos (section 4.2.2) and ethos (section 4.2.3). We then model positions in section 4.2.4.

4.2.1 Modeling logos

The **logos** technique appeals to logic or reason (section 2.2.4). Arguments using logos are based on factual data and on the conclusions that can be drawn from it. These conclusions should be accepted by an audience because they sound rational. Logic and rationality require a certain degree of abstraction, and are expressed using language, which can be of any type, e.g. natural language or symbolic language. For these reasons we model logos by modeling *verbal* information. Verbal information is present in speech in the auditory channel and writing in the visual channel, therefore in order to model logos we need to look at speech and writing.

In interview documentaries, most of the verbal information is conveyed by speech, i.e. by the interviewees' answers to questions. The LOGOS ARGUMENT [LLR 1] requirement specifies to use the Toulmin model to represent arguments. Using Toulmin, the argument an interviewee makes during an interview can be decomposed into the functional parts defined by the model. Each part must then be encoded separately, together with the role it plays in the argument. The ANNOTATION STRUCTURE [LLR 4] requirement states that the annotation of the clips must have a sentence-like structure and use a controlled vocabulary. We use statements to encode the different parts of an argument decomposed according to the Toulmin model (section 4.2.1.1) and a thesaurus for the controlled vocabulary (section 4.2.1.2). We then present the problem of clips granularity in the context of video generation and in particular of argument composition (section 4.2.1.3), and we show how the Toulmin model, by encoding an argument structure, can provide a context for clip annotations (section 4.2.1.4).

4.2.1.1 Statements

A **statement** is a short sentence that captures the sense of what the speaker says, such as "War is not effective", or "Diplomacy cannot be used". A statement can summarize the actual words used by the interviewee while expressing her position. For example, the transcript "I am never a fan of military actions, in the big picture I do not think they are ever a good thing" can be summarized by the statement "*Military actions are not effective*" or "*Military actions are not good*". Statements (called point in Cleary and Bareiss's approach[1], section 3.3.1.3) can capture and summarize the content of a clip. Statements do not capture all the semantics contained in the original sentences. This is not a limitation, since we only need to encode sufficient information to represent how arguments can be built, analogous to the approaches adopted by ScholOnto [63] (section 3.2.1.5) and Splicer [58] (section 3.2.1.4). If using only one statement cannot fully capture the content of a clip, multiple statements can be used.

[1] We adopt the term *statement* because it is more commonly used for interviews.

subject	*war*	*diplomacy*	*military actions*
war	*Id*	*Opposite*	*Similar*
diplomacy	*Opposite*	*Id*	*Opposite*
military actions	*Similar*	*Opposite*	*Id*

Table 4.2: Example of thesaurus terms and relations between them for the subject part of the statement.

modifier	*once*	*never*	*always*
once	*Id*	*Opposite*	
never	*Opposite*	*Id*	*Opposite*
always		*Opposite*	*Id*

Table 4.3: Example of thesaurus terms and relations between them for the modifier part of the statement. Not all terms need to be related, as in the case of *once* and *always*.

According to the ANNOTATION STRUCTURE [LLR 4] requirement, statements must be encoded in a fixed sentence-like structure, using a controlled vocabulary for the values of the properties. We model statements using a three-part structure: a **subject**, a **modifier** and a **predicate**. The subject (s) represents the subject of the statement, the predicate (p) qualifies the subject and the modifier (m) modifies the relation between the subject and the predicate. A statement is not required to have a modifier (*no mod*), whereas the subject and the predicate are required. The statement "*They are using two billion dollar bombs on ten dollar tents*", for example, is encoded as s:Bombing m:not p:effective.

The choice of a three-part structure results from a trade-off between expressiveness (how well a statement represents what is actually said) and computational complexity (how processor-intensive inferencing on these statements is). Using more than three parts would increase the expressiveness but also the computational complexity, as it will be shown when discussing the generation process in section 5.3.2. In [9] we tried with a four-part structure, but we found that with three parts we can describe the clip content with a degree of detail sufficient to represent arguments. AAB [20] (section 3.3.1.3) also uses a three-part structure, while Cleary and Bareiss's point structure has five parts (section 3.3.1.3).

predicate	*effective*	*waste*	*useless*
effective	*Id*	*Opposite*	
waste	*Opposite*	*Id*	*Similar*
useless		*Similar*	*Id*

Table 4.4: Example of thesaurus terms and relations between them for the predicate part of the statement. Not all terms need to be related, as in the case of *useless* and *effective*.

4.2.1.2 Thesaurus

As specified in ANNOTATION STRUCTURE [LLR 4], the terms used for the subject, predicate and modifier must belong to a controlled vocabulary, i.e. the value structure (section 3.3.1.1) must be either a taxonomy, a thesaurus or an ontology. Using a controlled vocabulary allows inferencing of relations between the statements and the corresponding video clips (section 3.3.1.3), as required by GRAPH GENERATION [LLR 6]. Two conflicting interests are at stake in choosing the value structure: inferencing is facilitated by constrained structures, such as an ontology, while an annotator's effort is reduced by having a loose structure, such as free text (section 3.3.1.1). The ideal compromise is when annotating requires the least effort while still supporting the inferencing process. The choice of the controlled vocabulary is therefore dependent on the inferencing mechanism.

In our case, the GRAPH GENERATION [LLR 6] requirement specifies that the inferencing must relate video clips (i.e. the corresponding statements) using argumentation relations. Relations between statements can be inferred from the relations between the terms used in these statements[2]. We need therefore a controlled vocabulary to have relations between terms. A taxonomy provides only a hierarchy, but no relations between terms, and it is therefore not suited for our purpose. An ontology can provide relations between terms, as well as properties and formal constraints on how these terms can be used together (section 3.3.1.1). Since we only need the relations between terms, an ontology would require an unnecessary modeling effort. In our approach we use therefore a thesaurus. The main relations in a thesaurus are ([72]):

- ***Generalization*** (Broader Term, hypernym) is a more general term, e.g. "media" is a generalization of "television". This relation is the inverse of *Specialization*.
- ***Specialization*** (Narrower Term, hyponym) is a more specific term, e.g. "Americans" is a specialization of "people". This relation is the inverse of *Generalization*.
- ***Similar*** (Related Term) between two terms that have the same or similar meaning, e.g. "war" is similar to "military actions". This relation holds between synonyms and near-synonyms and it is symmetric.

Some thesauri such as Wordnet [45] have also an ***Opposite*** (antonym) relation, between two different words of opposite meaning, e.g. "war" is opposite of "peace". This relation is symmetric. To make notations easier to display, we also introduce an identity relation ***Id*** between each term and itself, e.g. "people" *Id* "people". This relation is symmetric.

These relations the between terms must support the process of inferring the argumentation relations between the statements, as specified by GRAPH GENERATION [LLR 6]. According to this requirement, argumentation relations must be either supports or contradicts. As we show in the next chapter, the relations *Similar*, *Generalization* and *Specialization* lead to a positive relation (i.e. supports), while *Opposite* leads to a negative one (i.e. contradicts).

A thesaurus allows the definition of a more fine-grained measurement of whether terms are related than yes or no, by assigning a degree of reliability that decreases

[2]This can only be fully understood when we present the inferencing mechanism in the next chapter. For the moment the reader must assume that the relation between two terms can be used to infer the relation between two statements that contain these terms.

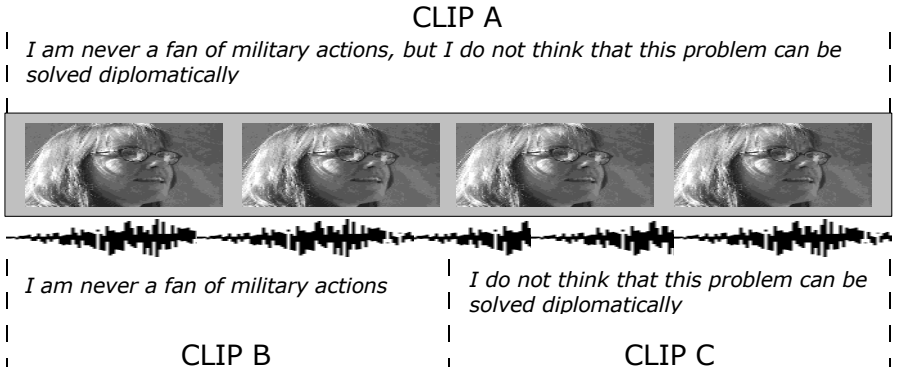

Figure 4.1: Clip A contains the complete interviewee's answer containing two statements, while Clip B and Clip C segment the answer in two parts of one statement each

as more steps are used to relate two terms. For example, if A is related to B, which is related to C, which is related to D, the relation between A and D is less reliable than the relation between A and B. The more steps needed to relate two terms, the less reliable the relation is. A threshold can also be established, beyond which the terms, even though connected by a chain of relations, are not considered to be related anymore. This is equivalent to saying that the property of being related is not transitive.

When annotating the content of a clip with a statement, an annotator is required to use terms belonging to the thesaurus as values for each of the three parts of a statement. Because subject, modifier and predicate each have a different role in the statement, terms used for one part of the statement are not related to terms used for another part of the statement. The thesaurus thus can be considered as composed of three independent thesauri, one for each part of the statement. An example of each thesaurus is represented in Tables 4.2, 4.3 and 4.4 on page 68.

The thesaurus can be built by an annotator in parallel with annotating media items, by inserting the terms she uses to compose the statements. Existing thesauri such as Wordnet can provide an aid or a starting point. An annotator also needs to relate each term she uses to the other terms in the thesaurus, using the four thesaurus relations.

4.2.1.3 Clips granularity

As specified in the ANNOTATION CONTENT [LLR 5] requirement, video content must be segmented into clips to be annotated. A clip is self-contained and coherent on its own, and viewers are able to understand what a clip is about even when the clip is shown within different sequences. This enables the reuse of the clip in different contexts (section 3.3.2.1). In the case of interview documentaries, the length of the clips must be determined as a trade-off between how easy it is to reuse the clip and how representative the clip is to the interviewee's intentions. Longer clips are more self-contained because they establish more context, but for the same reason more difficult to reuse in another context. Longer clips are thus more difficult to use for building different arguments. A finer granularity allows more options for building arguments but risks misrepresenting the interviewee's position. We discuss this issue with an example (see fig. 4.1): consider a video interview stating the following: "I am never a fan of military actions, but I do not think that this problem can be solved diplomatically". If

this video interview is annotated as a single clip (clip A), it can be used in an argument *for* military actions. If, instead, it is also segmented into the following two clips (clip B and clip C): "*I am never a fan of military actions*" and "*I do not think that this problem can be solved diplomatically*", clip B can be used in an argument *against* military actions, while clip C can be used in an argument *for* military actions.

A finer granularity offers thus more options to build arguments. On the other hand, clip A represents the position of the interviewee, clip C is still true to the position and clip B gives the wrong impression. Therefore, a side effect of a finer granularity is that clips can be taken out of context and misrepresent what was intended[3]. As specified in the CONTEXT [HLR 3] requirement, an automatic generation approach needs to encode context information to present a clip in order to avoid unintentional misunderstandings, as it would happen if only clip B would be shown, instead of clip A, to represent the interviewee's position.

To determine the clip granularity, we choose the smallest size for a clip such that the annotator can still assess that the associated statement (defined in section 4.2.1.1) applies to what the interviewee has stated. This requires that the clip contains at least one sentence from the interviewee. A clip where an interviewee replies to questions with a short answer such as "No" or "Yes" is therefore too short. These clips do not offer sufficient context to convey what the intention of the interviewee is ("no" or "yes" to what?). Their meaning depends entirely on what is shown before or after them. Furthermore, speech must be properly segmented so that the clip does not sound strange to the viewer. This requires starting and ending the clip at appropriate points of the interviewee's answer, respecting word boundaries as well as the intended meaning of the sentences. For example, in fig. 4.1 neither Clip B nor Clip C contains the "but" in between the two sentences, since a clip starting or ending with "but" would give the viewer the impression that a part of the answer was left out by mistake. Nevertheless, the semantics associated with "but" cannot be lost: although they are contained in Clip A, they also need to be encoded in the context information associated with Clip B and Clip C. Context information can be provided by determining the role each of the interviewee's statements plays in building the argument.

4.2.1.4 Argument structure

Analyzing the example in fig. 4.1, it is clear that not all sentences an interviewee says have the same weight when expressing her position. How important a part of an argument is can be determined by an argument model.

The LOGOS ARGUMENT [LLR 1] requirement specifies that an argument should be encoded using the Toulmin model [71] (section 3.2.1.1). An advantage of the Toulmin model is that it can be extended if it is unable to capture a particular argument type. In our model, we do not use the qualifier, since its function is already encoded in the statement's modifier. This modification is also used by other approaches ([7, 50]). The adapted Toulmin model is represented in the upper left corner of fig. 4.2.

Using the Toulmin model to annotate the example in fig. 4.1, we see that the statement expressed in clip B is a concession, which only expresses a concern of the speaker, while clip C is the claim, the point the speaker wants to make. Therefore, Toulmin can explain why presenting Clip B (i.e. the concession in the interviewee's argument) is misrepresenting the real position of the interviewee. Context information

[3] As we discussed in section 2.2.7, misrepresenting someone's position is a risk that must be taken into account also by human documentarists, not only by an automatic generation system.

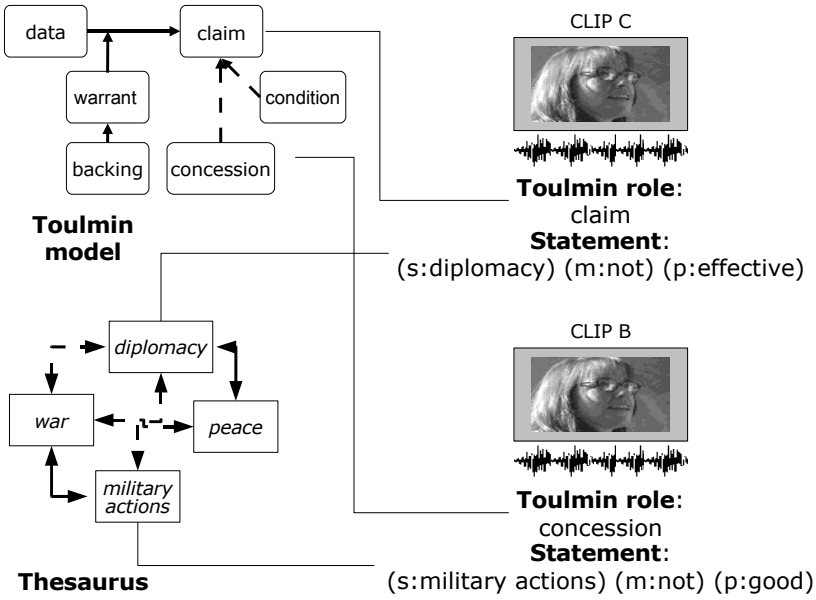

Figure 4.2: Annotating the two clips of fig. 4.1 according to the adapted Toulmin model (without the modifier) and the three-part statements. A thesaurus is used to provide the terms for the statements. Dashed lines in the thesaurus indicate relation *Opposite*, while continuous lines indicate relation *Similar*.

is therefore provided by encoding Toulmin's role together with the statement expressed in a clip.

In fig. 4.2 we show how the clips introduced in fig. 4.1 are annotated with the Toulmin model and the statements. Statements use terms from a thesaurus for the subject, modifier and predicate parts (in fig. 4.2 only terms for the subject part are shown).

4.2.2 Modeling pathos

The **pathos** technique gives strength to an argument because of the emotions the speaker has elicited in the viewer (section 3.2.2). In theory, an emotional message can be conveyed using each of the media types identified in Table 4.1. Emotions can be elicited by verbal information (e.g. an emotional speech) as well as non-verbal information (e.g. emotional images such as flags or children). The PATHOS ARGUMENT [LLR 2] requirement limits this possibility to modeling non-verbal information, taking two factors into account: first, whether the interviewee is appealing to the viewer and second, the framing distance of the video clip and the gaze direction of the interviewee.

There are no standards to define how appealing an interviewee looks to a viewer. One possibility is to define characteristics of the interviewee's aspect and a user profile that can be used to evaluate whether these characteristics are appealing to the viewer. Although possible, in order to simplify the annotation effort we choose not to model this factor, leaving it to future work.

Framing and gaze information is contained in the video media type. The framing property can assume the values extreme close-up, close-up, medium close-up, medium,

Figure 4.3: The first clip has framing medium and gaze right, the second framing medium close-up and gaze right-center, the third framing close-up and gaze right-center.

medium long, long and extreme long (section 2.3.2). For the gaze property we define the following values: left, left-center, center, right, right-center, mixed.

We model the pathos of an interviewee as increasing when the framing becomes closer and the gaze is more central. We assign a numerical value to each combination of Gaze and Framing. This value ranges from 0 when framing is equal to extreme long to 6 when framing is equal to extreme close-up and gaze to center (see Table 4.5). The maximum value for pathos is chosen to be equal to maximum value for ethos (section 4.2.3). We assume framing distances greater or equal to medium-long (when the subject is framed from the knees upwards) too far from the viewer to have a pathos effect. For the other framing distances, the more the gaze is directed to the viewer (and therefore to the center), the more the pathos of the interviewee increases. A "mixed" gaze, i.e. when the interviewee changes gaze direction in the clip, is considered not to be directed at the viewer, as a left or right gaze. For example, in fig. 4.3 the first clip's pathos value is 0.5, the second clip's pathos value is 2 and the third clip's pathos value is 4. In the case the framing distance changes during the clip, the pathos value is the average of the pathos values associated to the start and end framing.

Pathos can only assume positive values. This is because we only model whether the viewer feels close and connected with the interviewee (section 3.2.2), but we do not model whether this feeling influences the viewer positively or negatively. In our model the interviewee can either not affect the viewer (i.e. no pathos, with a value of zero), or affect her in a positive way.

We are aware that the values we assign to the different framing and gaze combinations are arbitrary and based on our interpretation of the theory we chose, i.e. film theory. Ideally, the documentarist can change them to reflect how important she thinks a particular combination should be. The same applies also to the values we use for ethos in section 4.2.3.

4.2.3 Modeling ethos

The **ethos** technique gives strength to an argument based on the speaker's authority and trustworthiness as perceived by the viewer. According to the ETHOS ARGUMENT [LLR 3] requirement, the ethos value of an interviewee can be determined on the basis of the social categories the interviewee belongs to and a user profile determining how important for the user these categories are. The social categories the interviewee belongs to must be determined on the basis of how the interviewee looks and sounds. Ethos information can be obtained therefore from the video, image and speech media

	Gaze		
Framing	mixed, left, right	left-center, right-center	center
extreme close-up	5	5.5	6
close-up	3.5	4	4.5
medium close-up	2	2.5	3
medium	0.5	1	1.5
medium long	0	0	0
long	0	0	0
extreme long	0	0	0

Table 4.5: Pathos values for the different framing/gaze combinations.

types. The social categories we use are:

- **age** with values teenager, young, middle-aged and old;
- **education** with values low education, medium education, high education;
- **employment** with values student, low-income job, middle-income job, high-income job, retired, unemployed;
- **race** with values White, Afro American, American Indian, Asian, and Hispanic;
- **religion** with values Muslim, Christian, Atheist;
- **gender** with values male and female.

In our approach we can annotate content on the denotative or connotative level, as specified in the ANNOTATION CONTENT [LLR 5] requirement. In the ethos case, connotative annotations are more appropriate because they can be used to describe the speaker directly in terms of the social categories. In the annotations, each interviewee is classified in terms of the above categories, for example "young - lowly-educated - low-income job - White - Christian - male".

To assess the ethos value, we use a user profile that contains weights for each of the values in the above mentioned list (from -1 to 1, with -1=negative influence, 0=neutral and 1=positive influence) reflecting how that particular social category influences the viewer in determining the authority of the speaker. The ethos rating of an interviewee is calculated by summing all weights corresponding to the specific viewer's values. For example, if the user profile contains the following values: (Christian=1, male=1, Afro American=-1, Hispanic=-1, White=1, high-income job=1, low-income job=-1, teenager=-1, young=-1), and the interviewee is classified as in the example above, then the ethos value is 1. In this way an ethos value is associated to each interviewee. At the moment we use stereotypical user profiles, such as "Pacifist" or "Racist".

Ethos can range from -6 (the viewer is negatively influenced by each social category the interviewee belongs to) to 6 (the viewer is positively influenced by each social category the interviewee belongs to). Unlike pathos, ethos can also have negative values, because we want to model the case the viewer dislikes particular social categories, so that she thinks an argument is wrong because of the interviewee who proposes it.

4.2.4 Modeling positions

To implement the rhetorical form as defined in the PRESENTATION FORM [HLR 1] requirement, the different positions of the interviewees need to be modeled. Each of these positions is supported by arguments. In theory, an interviewee's position with respect to a certain subject could be inferred knowing the interviewee's arguments. For example, an interviewee who says about the war in Afghanistan that "I am never a fan of military actions, but I do not think that this problem can be solved diplomatically" has a position *for* the war in that country. This kind of inferencing requires encoding domain knowledge in our model, an option that we ruled out in chapter 3 since it is against the MEDIA-DRIVEN [HLR 6] requirement. We therefore do not infer an interviewee's position, but we encode it explicitly in the annotations, using a connotative level of content representation (section 3.3.2.2), as in the case of ethos.

We model positions using two values, the subject, which is a controversial issue such as "*war in Afghanistan*", and the interviewee's attitude with respect to the subject, which can be "*for*", "*against*" and "*neutral*". The annotator is free to choose the subject, while the interviewee's attitude can only be one of the above-mentioned values.

4.3 Categorical form annotations

According to the CATEGORICAL FORM [LLR 9] requirement, the model should use categories in order to organize the information in the documentary at the macro-level (section 3.4). In this section we introduce the categories that can be used in documentaries about matter-of-opinion issues.

Interview categories

The first two categories concern characteristics of the interview, namely the **question** asked and the **interviewee identity** answering the question. These two categories allow the selection of clips based on the question asked (e.g. show all answers to the question "What do you think of the war in Afghanistan?") or on the person being interviewed (e.g. show all answers given by this interviewee). Furthermore, the interviewee can also be specified using the **social categories** introduced in section 4.2.3. This allows the selection of clips based on the social categories of the interviewee (e.g. show all answers given by Afro-Americans). A third category related to the interview is the **position** expressed, which is also needed for the rhetorical form (section 4.2.4). This category allows the selection of clips with a particular position (e.g. show all interviews that are for the war in Afghanistan).

Questions, interviewees or positions represent the possible subjects the generated documentary can have, and are therefore selected by the viewer (SUBJECT-POINT OF VIEW [HLR 2] requirement), as we show in the next chapter.

Location categories

We use the categories **environment**, which describes where the scene took place, such as *shop*, *parking lot*, *street*, and **geographical location**, which is where the scene took place, i.e. Cleveland (OH). These categories allow the selection of clips shot in a particular location or a particular environment, e.g. show all interviews that take place in a parking lot in Stamford (CT).

Temporal categories

We use the categories **time of the day**, which can assume the values *morning*, *afternoon*, *evening* and *night*, and **date**, which represents the date when the clip was shot. These categories allow to select clips shot on a particular date or, e.g., at night.

All these categories can be combined to select clips, such as selecting all clips where a particular question is asked to an Afro-American on a street at night.

4.4 Modeling cinematic content

Video has its visual language and its means of communicating content. Continuity editing provides rules to assemble sequences so that the viewer can follow the film and not feel disoriented (section 2.3.4). We specified that our approach needs to implement these rules with the CONTINUITY RULES [HLR 5] requirement. To encode the necessary information for continuity rules, we specified in the ANNOTATION CONTENT [LLR 5] requirement that our annotations need to capture cinematic content (section 3.3.2.2).

In this section we discuss how the continuity rules apply to video interviews and which properties of video need to be modeled to support the implementation of these rules.

Gaze continuity

According to the *180° system*, the interviewee cannot look one way in one shot and the opposite way in another, because this would give the impression she changed her position. Therefore, the annotation schema must encode the interviewee's gaze direction, i.e. the direction the interviewee is looking at. We model this with the **gaze property**, which can assume the values left, left-center, center, right, right-center and mixed. This property is also used to calculate the pathos value of a video clip (section 4.2.2).

Framing continuity

The *framing continuity* rule requires that the framing does not change too much from one clip to the following one. To prevent this, framing information must be encoded in the annotations. We model this with the property **framing**, which can assume the values extreme close-up, close-up, medium close-up, medium, medium long, long, extreme long (2.3.2). Since camera movements such as zooms and pans change the framing distance within the duration of a clip, the framing property has two values, a start value at the beginning of the clip and an end value at the end of the clip. The framing property is also used to calculate the pathos value of a video clip (section 4.2.2).

Interviewee Identity

Temporal continuity requires avoiding *jump cuts*. This rule applied to video interviews requires that two clips showing the same person cannot be joined with a hard cut, since the person would appear to suddenly jump to a different position in the frame. To prevent this, the clip must be annotated with the **interviewee identity**. This property is also used in the categorical form (section 4.3).

Rhetorical technique	image	video	writing	noise	music	speech
Logos	no	no	**yes**	no	no	**yes**
Pathos	yes	**yes**	yes	yes	yes	yes
Ethos	**yes**	**yes**	no	no	no	**yes**

Table 4.6: Relevant media types for each rhetorical technique. "yes" means the particular media type potentially contains information relevant to model the corresponding rhetorical technique. **yes** means we use information contained in the particular media type to model the corresponding rhetorical technique.

Camera movement continuity

The *camera movement* rule prevents the viewer from getting sea-sick, when a camera movement in one direction, for example a pan left, is followed by a camera movement in the opposite direction, for example a pan right. To avoid this, the annotations must encode the camera movement of the clip. We model this with the **camera movement** property, which can assume the values none, pan left, pan right, shaking, tilt down, tilt up, zoom in, and zoom out (section 2.3.2).

4.5 Conclusions

In this chapter we described the different parts of the annotation schema, which answers *Research Question Annotation Schema [2].*

We modeled the components that are needed to implement the rhetorical form, namely positions and arguments according to logos, pathos and ethos (section 4.2). We modeled logos arguments using statements and the Toulmin model. We modeled pathos and ethos as a value which can be assigned to each video clip. The relations between rhetorical techniques and media types are summarized in Table 4.6, where one can see that the model we defined in the previous chapter does not use all the information potentially relevant for pathos (section 3.2.2). Further research is therefore needed to provide theories motivating an interpretation of all the relevant information for rhetorical purposes.

In fig. 4.4 we show how the components of the rhetorical form relate to each other and to the video they describe. A *position* annotates a video clip containing a part of an interview. The position can be further decomposed into different *arguments* used to express it, two in this case. Each of the arguments can be decomposed according to the Toulmin model into different functional parts, three in this case: the claim, the data and the concession. Each of these parts is modeled with a *statement*, a *pathos* and an *ethos* value. Each position, argument and statement is associated with a part of the clip. In fig. 4.4 the annotations are associated with 9 different clips: one corresponds to the entire clip associated with the position (a-r), two clips are associated with the two arguments composing the position (a-h, i-r), and six clips are associated with six statements (b-c, d-e, f-h, l-m, n-o, p-q). Each one of these clips can be selected when generating a documentary. As can be seen in the figure, arguments and statements may also apply only to a subpart of the interviewee's answer or, in other words, not all the interviewee's answer need to be decomposed in arguments and statements. It is the annotator's choice whether to annotate a particular piece of footage. Considering

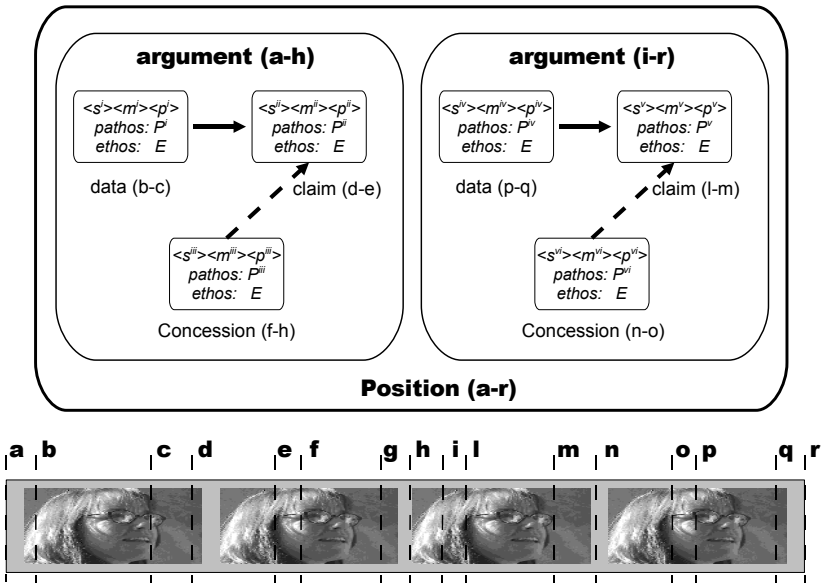

Figure 4.4: The relation between positions, arguments and statements in the rhetorical form. Each element relates to a portion of the video footage. Not all parts in the Toulmin model need to be present in an argument. The ethos value is constant, since the interviewee is the same.

that what is not annotated cannot be used to generate arguments, the annotator should annotate as much as possible. Not doing so would invalidate the benefits of the current approach, since the user would not have access to all the footage, thus retaining the limitations of traditional documentaries.

We modeled the categories needed for the categorical form (section 4.3), which are interview categories (question, interviewee identity, social categories of the interviewee, position), location categories (environment, geographical location), and temporal categories (time of the day, date). We then modeled the properties of clips needed to support the continuity rules (section 4.4), which are gaze for gaze continuity, framing for framing continuity, interviewee identity for avoiding jump cuts, and camera movement for the camera movement continuity. The complete list of properties for a clip is reported in Table 4.7.

We now show how an interview is annotated. Firstly, the **interviewee identity** and the social categories she belongs to are annotated. This information is used for the categorical form, to determine the ethos of the interviewee and to avoid jump cuts (sections 4.3, 4.2.3 and 4.4, respectively).

Secondly, the **question** asked is annotated. The duration of the interviewee's answer defines the clip to which the question is associated. This allows to present all answers to a particular question (section 4.3). If the interviewee's answer expresses a particular **position**, this is also annotated (section 4.2.4).

At this point the annotator analyzes the content of the answer, determining the **arguments** used by the interviewee to support her position, with the Toulmin model (section 4.2.1.4). Each part of the argument is decomposed into **statements** (section 4.2.1.1). Thus, at the finest level of granularity we have video clips annotated with

Property	**values**
position	"*subject - [for, against, neutral]*"
Toulmin role	claim, data, warrant, backing, concession, condition
subject	terms from thesaurus
modifier	terms from thesaurus
predicate	terms from thesaurus
question	free text e.g. "What do you think of the war in Afghanistan?"
interviewee identity	identifier e.g. lawyer in Harvard
social categories	age, education, employment, race, religion, gender
environment	e.g. parking lot
geographical location	e.g. Cleveland (OH)
time of the day	morning, afternoon, evening, night
date	e.g. 9-11-2001
gaze direction	left, left center, center, right, right center, mixed
framing	extreme close-up, close-up, medium close-up, medium, medium long, long, extreme long
camera movement	none, pan left, pan right, shaking, tilt down, tilt up, zoom in, zoom out

Table 4.7: List of all the properties in a clip annotation, with possible values

statements, which are subparts of the clips representing the interviewee's arguments, which in turn are subparts of the interviewee's answer. Each term used to annotate the statements belongs to a thesaurus, which the annotator builds while annotating by inserting terms and relating them to each other (section 4.2.1.2). Each video clip is also annotated with the categories for the categorical form (section 4.3) and the cinematic properties for continuity editing (section 4.4). The latter also include the properties to calculate the pathos (section 4.2.2).

The annotation schema we define in this chapter is used by the second component of our model, the generation process we explain in the next chapter.

Chapter 5

The generation process

This chapter completes the specification of our video documentary generation model, the first component of which, the annotation schema, is defined in the previous chapter. In the first part of the chapter we define a process capable of generating documentaries according to the requirements we set in chapters 2 and 3. This generation process uses the annotation schema defined in chapter 4 to establish the rhetorical relations among the media items in the repository and to compose and edit matter-of-opinion documentaries. The definition of this generation process provides an answer to *Research Question Generation Process [3].* In the second part of the chapter we define a method to evaluate the performances of the generation process, in order to provide the documentarist with feedback about the correctness of the annotations. These feedback methods provide an answer to *Research Question Author Support [4].* This chapter is based on [10] and [11].

5.1 Introduction

This chapter specifies the generation process, which, together with the annotation schema, forms the generation model specified by the Low-level requirements in chapter 3. The output of the generation process is a viewer-requested documentary, which satisfies the High-level requirements set in chapter 2.

The chapter consists of two parts, which answer *Research Question Generation Process [3]* and *Research Question Author Support [4].* The first part (section 5.2) specifies the generation process, discussing the rhetorical data structure it creates, how it uses this structure to compose and edit arguments, and how it assembles arguments to form documentaries. These phases use the information captured in the annotation schema and are driven by the Low-level Process requirements GRAPH GENERATION [LLR 6], ARGUMENT COMPOSITION [LLR 7] and CATEGORICAL FORM [LLR 9]. In the second part (section 5.3) we define a method to check the performance of the generation process in creating the rhetorical data structure. The quality of this data structure is dependent on the annotations provided by the documentarist, and influences the generation process's capability to generate documentaries. Our model is therefore able to provide some feedback to the documentarist about the documentaries it is able to generate, as specified by the QUALITY FEEDBACK [LLR 8] requirement.

In this chapter we use the word relation in two different contexts: among terms in the thesaurus and among statements in the repository. To avoid confusion, we use the

word **relation** only for terms in the thesaurus, while we say that between two statements there can be a **link**. Furthermore, we assume that a repository contains media items as well as the associated annotations.

5.2 Generating a documentary

In this section we specify a process capable of generating documentaries about matter-of-opinion issues, as specified by the high-level requirements in chapter 2. The definition of this process answers *Research Question Generation Process [3].*

In section 5.2.1 we explain how the process creates the data structure that is used to generate documentaries according to the rhetorical form, i.e. the semantic graph. This phase is divided into two steps because this allows the definition of a more detailed feedback mechanism, as we explain in the second part of this chapter (section 5.3).

Arguments are the building blocks of the rhetorical form. In section 5.2.2 we discuss how an argument can be composed based on the dynamically created semantic graph. We define a method to create an argument from an initial interview clip, and we explain how, depending on the viewer-selected point of view, the argument is composed.

Once the material is selected, it needs to be edited in a video sequence to be presented to the viewer. In section 5.2.3 we describe how editing must serve the goal of representing the rhetoric of the content and comply to the continuity rules dictated by film theory.

Being able to compose an argument is necessary to implement the rhetorical form, but does not provide sufficient material to create a documentary. An argument can be used to create a scene, which corresponds to the *micro-level*. In a documentary, more scenes can be organized together on the *macro-level* using the categorical form. In section 5.2.4 we discuss how to integrate the rhetorical form and the categorical form with each other to generate documentaries that go beyond a single argument.

5.2.1 Creating the story space

In this section we explain how the generation process creates the story space, i.e. the data structure that is able to support the generation of documentaries according to the rhetorical form. This structure corresponds to what we called the *Semantic Graph* (sections 3.2.1.5 and 3.3.1.4), which in general is a graph whose nodes represent the concepts to be presented and the edges the relations among them. In our case the GRAPH GENERATION [LLR 6] requirement specifies the characteristics of this semantic graph: the nodes are annotations associated with the media items in the repository, and the edges are argumentation relations of two possible polarities, either positive (i.e. supports) or negative (i.e. contradicts). The generation process uses these argumentation relations to build arguments. Given the MEDIA-DRIVEN [HLR 6] requirement, the semantic graph needs to be dynamically created each time an annotated media item is added to the repository.

The GRAPH GENERATION [LLR 6] requirement further specifies that the relations among the nodes in the semantic graph must be inferred from the annotations. In chapter 4 we define the statement as a formal representation of clip content (section 4.2.1.1). We therefore use statements to determine which clips should be related to each other, i.e. among which nodes in the semantic graph an edge should exist. In the semantic graph, each node corresponds to a statement, which in turn represents a video clip.

	bomb.	*war*	*peace*	*dipl.*	*mil. act.*	*econ. aid*
bombing	*Id*	*Gen.*				*Opp.*
war	*Spec.*	*Id*	*Opp.*	*Opp.*	*Sim.*	*Opp.*
peace		*Opp.*	*Id*			
diplomacy		*Opp.*		*Id*	*Opp.*	
military-actions		*Sim.*		*Opp.*	*Id*	
economic-aid	*Opp.*	*Opp.*				*Id*

Table 5.1: Example of terms and relations between terms contained in the thesaurus for the subject part of the statement.

	no mod	*not*	*never*	*possibly*	*once*
no mod	*Id*	*Opposite*	*Opposite*		
not	*Opposite*	*Id*	*Similar*		
never	*Opposite*	*Similar*	*Id*		*Opposite*
possibly				*Id*	*Similar*
once			*Opposite*	*Similar*	*Id*

Table 5.2: Example of terms and relations between terms contained in the thesaurus for the modifier part of the statement. The *no mod* has a positive meaning, being opposite to *not* and *never*.

	effective	*waste*	*useless*
effective	*Id*	*Opposite*	
waste	*Opposite*	*Id*	*Similar*
useless		*Similar*	*Id*

Table 5.3: Example of terms and relations between terms contained in the thesaurus for the predicate part of the statement.

Therefore, nodes represent video clips as well. Using statements also implies that the semantic graph is created using only logos information. This is because pathos and ethos can provide a measure on how convincing a clip seems, but they do not provide any information on what argumentation relation should link two clips.

The process that creates the semantic graph is composed of two sub-processes, the first generating new statements and the second linking them. The rationale behind this choice is that decomposing the process into two steps allows to better pinpoint how annotations contribute to the graph creation. This information can be used to improve the quality of the generated semantic graph, as we show in section 5.3. We describe these two sub-processes in detail in the following two sections. During the discussion we use as examples terms and relations from Table 5.1, Table 5.2 and Table 5.3.

5.2.1.1 Generating new statements

The aim of the first sub-process is to generate, for each existing statement s_n that annotates a video clip in the repository, the set of all possible related statements R^{s_n}, regardless of whether these generated statements correspond to a video clip in the repository. The end result is a semantic graph containing all potential relations (edges) among nodes, where nodes are represented by statements. Some of these nodes do not correspond to any media item in the repository. The second sub-process then selects only the nodes that correspond to video clips in the repository. The second sub-process

Statements			Transformations	Link
subject	**modifier**	**predicate**		
war	*not*	*effective*	*Generalization* s	SUPPORTS
economic-aid	*not*	*effective*	*Opposite* s	CONTRADICTS
bombing	*not*	*waste*	*Opposite* p	CONTRADICTS
war	*no mod*	*effective*	*Generalization* s - *Opposite* m	CONTRADICTS
peace	*not*	*effective*	*Generalization* s - *Opposite* s	CONTRADICTS
bombing	*no mod*	*waste*	*Opposite* p - *Opposite* m	SUPPORTS
war	*not*	*waste*	*Opposite* p - *Generalization* s	CONTRADICTS
military-actions	*no mod*	*effective*	*Generalization* s - *Opposite* m - *Similar* s	CONTRADICTS
economic-aid	*no mod*	*effective*	*Generalization* s - *Opposite* s - *Opposite* m	SUPPORTS
bombing	*no mod*	*useless*	*Opposite* p - *Opposite* m - *Similar* p	SUPPORTS
war	*not*	*useless*	*Opposite* p - *Generalization* s - *Similar* p	CONTRADICTS

Table 5.4: Example of statements generated from s:bombing m:not m:effective using transformations on subject, modifier and predicate with terms and relations from Table 5.1, 5.2 and 5.3 (*Similar* s means apply relation *Similar* to the subject, and so on). In the last column, the type of link to the original statement in terms of the argumentation relations SUPPORTS and CONTRADICTS.

checks for each generated statement $s_g \in R^{s_n}$, whether s_g is equal to an existing statement s_m. If this is the case, s_m is related to the initial statement s_n and the second sub-process links s_n and s_m together.

The input to the first sub-process is the set of statements contained in the repository (existing statements). New statements are generated by replacing the terms in the existing statements with related terms contained in the thesaurus. The rationale for this is that the relation between two terms in the thesaurus can be used to infer the relation between two statements that contain these terms, as we show in section 5.2.1.2. We describe now how to generate new statements from an existing one. For each existing statement, the first sub-process retrieves the subject, modifier and predicate. Each new statement is generated by replacing either the subject, the modifier or the predicate of the original statement with a related term. The thesaurus defines whether two terms are related, and with which relation: either *Similar*, *Opposite*, *Generalization* or *Specialization* (section 4.2.1.2). At this stage, each new statement is similar to the original one with the exception of one term, i.e. either the subject, the modifier or the predicate. For example, let us assume that the original statement is s:bombing m:not p:effective. The

term *bombing* is *Opposite* to the term *economic-aid* and *Generalization* to the term *war* in the thesaurus (Table 5.4). The process thus is able to generates the following two new statements: s:war m:not p:effective and s:economic-aid m:not p:effective. Replacing one term constitutes one round of transformations. The same process is applied again to the generated statements. At each transformation round, the difference from the original statement increases: at the n-th round the new statements have been obtained by replacing n times terms from the original statement. Each term used as the subject, the modifier or the predicate in a generated statement is related through one or more relations to the term in the corresponding part of the original statement.

Table 5.4 shows examples of new statements that can be generated from the statement s:bombing m:not m:effective, using the thesaurus shown in Table 5.1, Table 5.2 and Table 5.3, with up to three rounds of transformations. Transformations are represented in the second column with two terms (e.g., *Opposite s*), the first being the name of relation in the thesaurus, which relates the replaced term to its replacement (in the example *Opposite*), and the second which statement part has been replaced (in the example the *subject*). The statement in the fifth row, for example, s:peace m:not m:effective, has been generated from the statement s:bombing m:not m:effective using two rounds of transformations. First, by replacing the subject *bombing* with *war* (since *bombing Generalization war* in Table 5.1), giving s:war m:not m:effective (first row in Table 5.4). Then, replacing again the subject *war* with *peace* (since *war Opposite peace* in Table 5.1), giving s:peace m:not m:effective. In this particular example, the subject of the statement has been replaced twice, but at each round, any of the statement parts can be replaced.

Since generated statements are composed of terms that are related to the terms of the original statement, the statements are also related. The generated statements can be considered as the "semantic neighborhood" of the original statement. Generated statements represent the semantic "mutations" of the original statements based on the relations provided in the thesaurus. Not all the "mutations" exist as annotations in the repository.

5.2.1.2 Linking statements

The goal of the second sub-process is to establish which existing statements should be linked together and how. As we stated earlier, the first sub-process has generated, for each existing statement s_n, the set of all related statements R^{s_n}. If a generated statement s_g from this set is equal to an existing statement s_m, s_m is related to the initial statement s_n. Therefore, s_n and s_m must be linked. The end result of this phase is the semantic graph as specified in the GRAPH GENERATION [LLR 6] requirement.

To verify whether a generated statement is equal to an existing statement, this sub-process searches for it among the annotations in the repository. A generated statement that is found in the repository generates a hit. A generated statement can generate no hits if there is no media item annotated with that statement, or one or more hits if one or more media items are annotated with the same statement.

Once it is established that two existing statements should be linked, the link type must be determined. According to the GRAPH GENERATION [LLR 6] requirement, the link type must be either SUPPORTS or CONTRADICTS. We assign the link type based on the transformations used by the first sub-process to get from the original statement to the generated one. To map from transformations applied to either SUPPORTS or CONTRADICTS links, we use the following criterion: if the statement is derived using no or an even number of *Opposite* relations, we assume that the link is SUPPORTS, other-

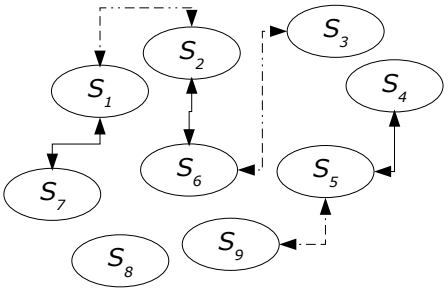

Figure 5.1: The generated semantic graph, with statements as nodes and edges of type SUPPORTS (continuous line) or CONTRADICTS (dashed line). When two statements are not related to each other, there is no link between them.

wise the link is CONTRADICTS. For example, the statement s:economic-aid p:effective (9th row in Table 5.4) has been derived from s:bombing m:not m:effective with three transformations: *Generalization* s, *Opposite* s, *Opposite* m. We therefore conclude that statement s:bombing m:not m:effective SUPPORTS s:economic-aid p:effective.

This inferencing method is not based on strict logic and cannot guarantee to produce always meaningful links. Intuitively, the more transformations used to derive a statement, the less we can rely on the conclusion. In order to avoid invalid links, we limit the number of transformations to three[1].

The end result of this phase is a semantic graph where the nodes are the statements and the edges are either SUPPORTS or CONTRADICTS links (see fig. 5.1). Since each statement is associated to a media item, the corresponding media items are also linked by either SUPPORTS or CONTRADICTS links. The generation process can thus use this data structure to support the composition of arguments.

5.2.2 The micro-level: arguments

In this section we explain how the story space we built in the previous section, together with the annotation schema defined in chapter 4, can be used to compose arguments and express a point of view with them. According to requirement PRESENTATION FORM [HLR 1], arguments must be composed based on logos, pathos and ethos. Arguments form the building blocks of the rhetorical form, which we use to organize material on the micro-level, or scene level (section 2.4).

We start in section 5.2.2.1 by focusing on how to compose a single argument. A single argument is based on a single interview segment, complemented by statements contained in other interview segments. The composition is based on the relations between the interview's statements and other statements contained in the semantic graph.

When requesting a documentary, the viewer can choose between the propagandist or the binary communicator point of view (SUBJECT-POINT OF VIEW [HLR 2] requirement). The way an argument is composed influences the point of view expressed. Logos is used to build the semantic graph (section 5.2.1). In section 5.2.2.2 we use pathos and ethos to assess the strength of an argument and drive an argument composition, so that the requested point of view can be represented in the generated documentary.

[1] We discuss this further in section 5.3.2.

In section 5.2.2.3 we show how the techniques we explain can be used to compose an argument that expresses a point of view.

5.2.2.1 Composing an argument

In order to compose statements into an argument, there must be a relation between them that motivates the composition. If they are related, two statements either support each other or they contradict each other. Requirement ARGUMENT COMPOSITION [LLR 7] specifies four defeat actions which can be used to contradict the statements of an argument decomposed according to the Toulmin model: *rebuttals*, *undercutters*, *sequential weakening* and *parallel weakening* (section 3.2.1.2). These actions only provide the means to counterargue an argument. When presenting a position, a documentarist sometimes wants to support the arguments presented by an interviewee with the statements presented by other interviewees. Therefore, analogously to the defeat actions, we define the following support actions:

- **non-rebuttals** state a conclusion similar to the given one;
- **non-undercutters** support the connection existing between a reason and a conclusion;
- **parallel and sequential strengthening** provide more than one reason for the conclusion.

The difference between parallel and sequential strengthening is analogous to the difference between parallel and sequential weakening: parallel strengthening is applied when more reasons or conclusions support different Toulmin parts of the arguments, while sequential strengthening implies that more reasons or conclusions support the same part of the argument. Defeat and support actions provide means for composing arguments. We discuss now how these means can be based on the generated semantic graph (section 5.2.1) to counterargue or support an argument to be presented. A *rebuttal* states a conclusion opposite to the given one or, using the Toulmin model, a statement that contradicts the claim. Since in an argument each part is encoded as a statement (section 4.2.1.4), rebuttals for a particular argument are all the statements that have a CONTRADICTS link in the semantic graph to the statement representing the argument's claim. Analogously, *non-rebuttals* are all statements that have a SUPPORTS link to the statement representing the argument's claim. *Undercutters* are directed to the connection existing between a reason and a conclusion. In the Toulmin model, the data, warrant and the backing support the claim, while the concessions and the conditions counterargue it. An undercutter can be applied to an argument in two ways: either contradicting a part that supports the claim, or supporting a part that counterargues the claim. Undercutters are therefore all the statements that have a CONTRADICTS link to any of the statements supporting the claim, plus all the statements that have a SUPPORTS link to the statements counterarguing it. Analogously, *non-undercutters* are all the statements that have a SUPPORTS link to any of the statements supporting the claim plus all the statements that have a CONTRADICTS link to the statements counterarguing it. *Parallel weakening* consists of applying undercutters and rebuttals to different parts of the argument, while *sequential weakening* consists of applying either rebuttals or undercutters to the same part of the argument. Analogously, *parallel strengthening* consists of applying non-undercutters and non-rebuttals to different parts of the

argument, while *sequential strengthening* consists of applying either non-rebuttals or non-undercutters to the same part of the argument.

Whether an argument needs to be supported or counterargued depends on the point of view the viewer has chosen, as we explain in the following section.

5.2.2.2 Expressing a point of view

The PRESENTATION FORM [HLR 1] requirement specifies that a documentary must be generated according to the propagandist or the binary communicator point of view, and the SUBJECT-POINT OF VIEW [HLR 2] requirement specifies that the viewer decides which point of view must be used. In this section we discuss how a point of view can be expressed when composing an argument.

Typically, interviewees have different positions with respect to particular subjects, i.e. they are either for, against or neutral about them. A propagandist tries to present material so that one of these positions is emphasized[2]. To do so, a propagandist can select only interviews that support a particular position. She can also include in the documentary interviews that express an opposing position, but select and present them so that her position looks more convincing. A binary communicator tries to present material so that both the for and the against positions appear equally strong.

Considering that opposing positions can be expressed in an argument, a position looks as strong or stronger than an opposing one when its statements are as strong as or stronger than the opposing position's statements. Disciplines such as argumentation theory define criteria to assess who the winner is of an argument between opposing parties. These are based on formal logic and on formal rules of turn-taking in disputing an argument [17], and are not applicable in our case because of the MEDIA-DRIVEN [HLR 6] requirement (as discussed in section 3.2.1). In our model, we choose to evaluate a position's strength in an argument by considering the total pathos and ethos value (sections 4.2.2 and 4.2.3, respectively) of all the interviewees taking part.

When composing arguments, an initial video clip is selected together with clips that either support or contradict the statements contained in it. To assess the strength of the position expressed by the initial clip, the generation process sums the pathos and ethos value of the initial clip's interviewee with the pathos and ethos values of each interviewee supporting (in the sense specified in section 5.2.2.1) the initial clip. The strength of the opposing position is given by the sum of the pathos and ethos values of all the interviewees counterarguing the initial clip[3]. Parallel weakening and sequential weakening can be used to find more clips counterarguing an initial clip, so that the pathos and ethos values of the opposing position can increase. Conversely, parallel strengthening and sequential strengthening can be used to increase the strength of the position expressed in the initial clip. Considering that ethos can have negative values, some clips may need to be omitted since the ethos value would decrease instead of increase.

When presenting the propagandist point of view, the generation process retrieves clips and composes arguments so that the ratio between the pathos and ethos values of the two positions is above a certain threshold T_{bias}:

$$\frac{P_s + E_s}{P_w + E_w} \geq T_{bias} \tag{5.1}$$

[2] Regardless as to whether it makes sense for a propagandist to present a neutral position as stronger, this is possible in our approach.

[3] A trivial case of the propagandist point of view is when only supporting positions are presented and only supporting clips are retrieved. In this case there is no need to calculate pathos and ethos values.

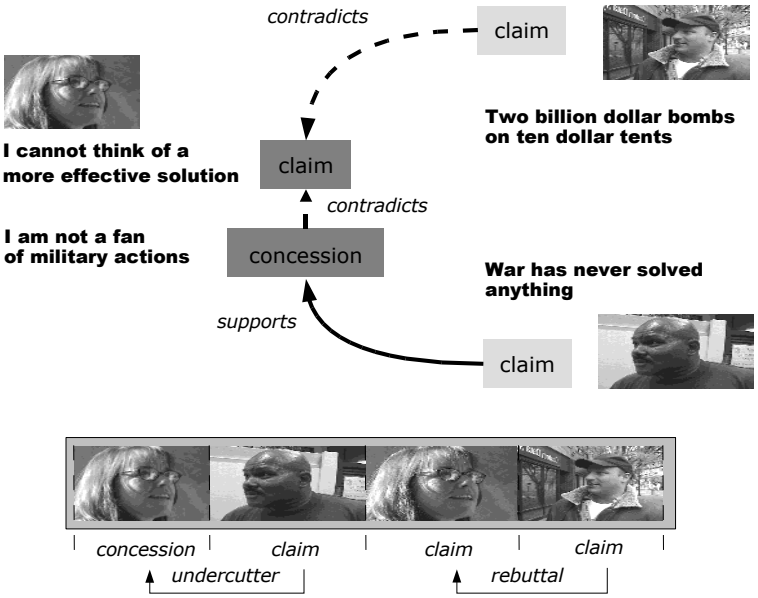

Figure 5.2: Assembling counterarguing arguments about the war in Afghanistan. Above, the argument structure, below, a possible video sequence representing the argument

with P_s, E_s, P_w and E_w the pathos and ethos values of the stronger and weaker argument, respectively, and $T_{bias} \gg 1$. When presenting the binary communicator point of view, the generation process retrieves clips and assembles arguments so that the pathos and ethos value of the first position (P_{arg1} and E_{arg1}) is similar to the pathos and ethos value of the second position (P_{arg2} and E_{arg2}):

$$\frac{P_{arg1} + E_{arg1}}{P_{arg2} + E_{arg2}} \approx 1 \tag{5.2}$$

5.2.2.3 An example of generating a point of view with an argument

We now show with a simplified example how to compose an argument and express a point of view with it. We suppose the viewer has determined the selection of a clip from an interview to start with, by choosing the position "*war in Afghanistan - For*" (section 4.2.4). Fig. 5.2 shows an interview in favor of the war in Afghanistan (the woman on the left) saying: "*I am never a fan of military actions, but in the current situation I cannot think of a more effective solution*"). This argument is composed of a claim ("*in the current situation I cannot think of a more effective solution*") and a concession ("*I am not a fan of military actions*"). Each part of this argument is encoded with a statement: the claim with *s:war p:no mod p:solution* and the concession with *s:military-actions m:never p:effective*. We suppose further that the viewer has chosen a propagandist point of view *against* the war in Afghanistan. The generation process tries to compose an argument so that the position "*war in Afghanistan - For*" appears weaker than the opposing position expressed by the initial clip. To do so, the generation process needs to counterargue the argument expressed in the interview by retrieving its

rebuttals and undercutters, as explained in section 5.2.2.1. In the figure, the generation process selects therefore an undercutter directed at the concession and a rebuttal (which by definition can only be directed at the claim). The undercutter is the statement *s:war m:not p:solution*. This statement has a CONTRADICTS link to the statement *s:war p:no mod p:solution*, since *no mod* has relation *Opposite* with *not* (see Table 5.2 on p. 83). The clip associated to this undercutter shows the man on the lower right saying "*War has never solved anything*". The rebuttal is the statement *s:military-actions p:no mod p:waste*. This statement has a SUPPORTS to the statement *s:military-actions m:never p:effective* since the modifier *m:never* is *Opposite* to *no mod* and the predicate *effective* is *Opposite* to *waste*. The rebuttal is associated to the clip of the man on the upper right saying "*They are using two billion dollar bombs on ten dollar tents*".

An argument composed in this way represents the viewer requested point of view only when equation 5.1 is satisfied. From fig. 5.2, the initial clip (top left) has framing = close-up and gaze = right, while the two counterarguing clips have framing = medium close-up and gaze = left. According to Table 4.5, the initial clip has the highest pathos value (3.5), but the pathos value of the counterargument is the sum of two individual pathos values (2 + 2 = 4). The ethos values of the two opposing positions depend on the user model. If we assume that the viewer values education, and the woman has a high education while the two men have a low education, the ethos value is 1 for the initial clip and 0 for the counter position. The total pathos/ethos value of the position "*war in Afghanistan - For*" is therefore 4.5, while the total pathos/ethos value of the position "*war in Afghanistan - Against*" is 4. The ratio between these values is close to one, and according to equation 5.2 the positions have similar strength. The selected clips express therefore a binary communicator point of view, which can be composed in the following video sequence (lower part of fig. 5.2): woman saying "*I am not a fan of military actions*"; lower man saying "*war has never solved anything*"; woman saying "*in the current situation I cannot think of a more effective solution*"; upper man saying "*two billion dollar bombs on ten dollar tents*". To represent a propagandist point of view against the war in Afghanistan, the process needs to select more clips counterarguing the initial clip so that the ratio between pathos/ethos values becomes $\gg 1$ (i.e. $\geq T_{bias}$).

5.2.3 Editing an argument

Once the video clips forming an argument have been selected, they need to be edited into a sequence to be presented to the viewer. In the editing phase, clips are ordered in a linear sequence and joined together using either cuts or transitions (section 2.3.3). Ordering requires that the initial structure, consisting of the interview segment (modeled with the structure of Toulmin) and the corresponding supporting and/or counterarguing clips, is transformed into a linear sequence (see also fig. 5.3on p. 91). We first examine how editing can be driven by the PRESENTATION FORM [HLR 1] requirement and then by the continuity rules specified in the CONTINUITY RULES [HLR 5] requirement.

5.2.3.1 Rhetoric-driven editing

In the rhetorical form, different positions and arguments need to be presented in relation to each other (requirement PRESENTATION FORM [HLR 1]). Having selected the clips to form an argument, the generation process needs to edit a sequence that shows the argument structure and the relation between arguments.

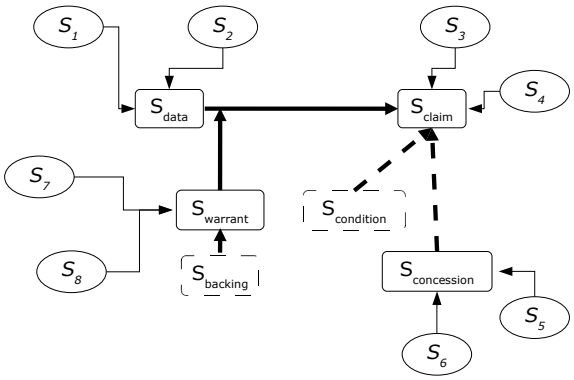

Figure 5.3: An initial interview complemented with supporting or counterarguing statements (statements are indicated with $S_{subscript}$). A dashed line means that that part of the argument is not present. Each statement corresponds to a clip. This structure must be linearized to be presented.

In sequencing the selected clips, the initial interview segment provides a structure for **argument editing**. In our approach we choose to display the clips belonging to this interview segment in the order in which they are recorded. In doing this we assume that the order in which the interviewee expresses her statements is important. Two options are then possible to display the supporting/counterarguing clips, so that positions are shown in relation to each other: either after the interview segment or after each statement of the interview they refer to. Referring to fig. 5.3, assume that the order in which the interview segment is recorded is $S_{data} - S_{warrant} - S_{claim} - S_{concession}$. According to the first of the above mentioned options, the generated sequence shows first the interview segment $S_{data} - S_{warrant} - S_{claim} - S_{concession}$ and then $S_1 - S_2 - S_7 - S_8 - S_3 - S_4 - S_5 - S_6$. The order of the supporting/counterarguing clips follows the order of the interview segment's clip they refer to, thus S_1 and S_2 are displayed first because they refer to S_{data}, and so on. The rationale for this choice of presenting the material is that the argument created by the supporting/counterarguing clips mirrors the structure of the argument presented in the interview. The structure of the argument as used by the interviewee is thus reused to compose a supporting or countering argument. According to the second of the above-mentioned options, the generated sequence is as follows: $S_{data} - S_1 - S_2 - S_{warrant} - S_7 - S_8 - S_{claim} - S_3 - S_4 - S_{concession} - S_5 - S_6$. This structure gives priority to the relations between clips rather than to the structure of the argument. Related clips are shown next to each other, so that their relation is more apparent than with the first option. In our approach we use both methods, allowing the viewer to determine which one she wants.

To enforce the idea of an argument being discussed between two parties, we use a technique called **shot-reverse shot** (or gaze matching), which consists of showing a character looking off-screen at another character, and then showing the other character looking "back" at the first character ([12] p. 314). Since the characters are shown facing in opposite directions, the viewer subconsciously assumes that they are looking at each other. Applied to interviews, if the interviewee shown in the first interview segment is looking to the right, the generation process presents supporting clips with interviewees

also looking to the right, and counterarguing clips with interviewees looking to the left. The result of the montage is a sequence of clips where the interviewees seem to look at each other, thereby creating the impression they are addressing each other. The technique does not influence the ordering of the clips as discussed above, and can therefore be used together with it. To support shot-reverse shot, the video footage is rendered in reverse, so that for each interviewee there are two clips with different gaze direction (unless the interviewee is looking to the center, in which case shot-reverse shot cannot be applied).

5.2.3.2 Continuity editing

The generated video sequence must satisfy the CONTINUITY RULES [HLR 5] requirement. The continuity rules we defined (section 4.4) are:

- the *gaze continuity* rule,
- the *framing continuity* rule,
- the *interviewee identity* rule,
- the *camera movement continuity* rule.

The **gaze continuity** rule implies that the interviewee does not look one way in one shot and the opposite way in another, giving the impression she changed her position. This rule can be enforced by listing the values for the gaze properties in the order *left*, *left-center*, *center*, *right-center* and *right* and ensuring that adjacent clips showing the same interviewee have the same or consecutive gaze values[4].

If clip X is followed by clip Y, the **framing continuity** rule specifies that the difference in framing between the end of clip X and the start of clip Y must be small. For example, a close-up cannot be followed by a long shot. To enforce this rule, we adopt an analogous approach to Auteur ([48], p. 123) by listing the values in the order *extreme close-up*, *close-up*, *medium close-up*, *medium*, *medium long*, *long*, *extreme long* and ensuring that adjacent clips have the same or consecutive framing values. Since a clip has a start and an end framing (section 4.4), the rules apply between the end framing of one clip and the start framing of the following one.

The **interviewee identity continuity** rule requires that two clips showing the same interviewee cannot be joined with a hard cut, since the viewer would not be able to understand why objects in the clip seem to jump to a new position or, in our case, that the interviewee's face suddenly changes expression and her head is suddenly in a slightly different position. To avoid a jump cut, one possibility is to enforce that

$$interviewee\{X\} \neq interviewee\{Y\}$$

for any two consecutive clips X,Y. Another solution is to insert a transition, which signals that time has elapsed:

$$if(interviewee\{X\} == interviewee\{Y\})$$
$$then\ insert\ transition\ between\ X\ and\ Y$$

[4]If the gaze is not constant within the clip (indicated in the annotation with the value mixed), the rule cannot be applied because we cannot enforce any gaze continuity.

for any two consecutive clips X,Y. We prefer this last solution since it does not constrain the sequencing of clips. Another way of avoiding a jump cut is to join two clips with different framings, as long as the framing continuity rule is satisfied, e.g. a close-up and a medium close-up shot ([54] p. 186):

> "*Because of the bold change of image size between the two frames [...] minor mismatches will go unnoticed by the audience, especially because the eye does not register the first three frames of a new image.* "

We therefore adapt the above rule to:

$$\begin{array}{l} if((interviewee\{X\} == interviewee\{Y\}) \quad AND \\ (framing_{end}\{X\} == framing_{start}\{Y\})) \\ then\ insert\ transition\ between\ X\ and\ Y \end{array}$$

for any two consecutive clips X,Y.

The **camera movement continuity** rule prevents the viewer from getting sea-sick, when a pan in one direction is followed by a pan in the opposite direction, or a zoom in is followed by a zoom out. These rules can be expressed as a list of if-then rules, such as:

$$\begin{array}{rcl} if(camera\ movement\{X\} & == & pan\ left) \\ then\ camera\ movement\{Y\} & \neq & pan\ right \end{array}$$

for any two consecutive clips X,Y. Analogous rules hold for tilt (*tilt down*, *tilt up*) and zoom (*zoom in*, *zoom out*). Camera movement can also have value None, for which no constraints apply, and Shaking, in which case the clip should not be used. If it has to be used because there is no alternative, then to avoid sea-sickness as much as possible, the camera movement of the adjacent clip should be None.

Compared to other automatic video editing systems such as Media Streams [23] and Auteur [48] (section 3.3.2), the continuity rules we use are simpler, since the domain of video documentaries has weaker requirements on cinematic quality with respect to action-based videos (sections 2.3.4 and 3.3.2.6). Nevertheless, with the exception of the interviewee identity continuity rule, the continuity rules constrain the possible choices in sequencing clips. This can lead to conflicts between rhetoric-driven and continuity-driven editing, such as when there is only one suitable video clip to counter-argue an interview segment, but it is not possible to include this clip without breaking some continuity rule. In these cases there is no clear solution: better rhetorical content but worse cinematic experience or worse rhetorical content but better cinematic experience? In our approach we leave the choice to the documentarist, by requiring her to specify which rules should be applied and in which order[5]. In section 6.4.2 we show an implementation of this mechanism.

5.2.4 The macro-level: the categorical/rhetorical form

A video sequence showing one argument does not provide sufficient material for a documentary. Documentaries usually present different topics. For example, **Interview**

[5]The order determines the priority, because if rule B is applied after rule A, rule A makes editing decisions which rule B cannot change. Rule B can only change what has not yet been established by rule A.

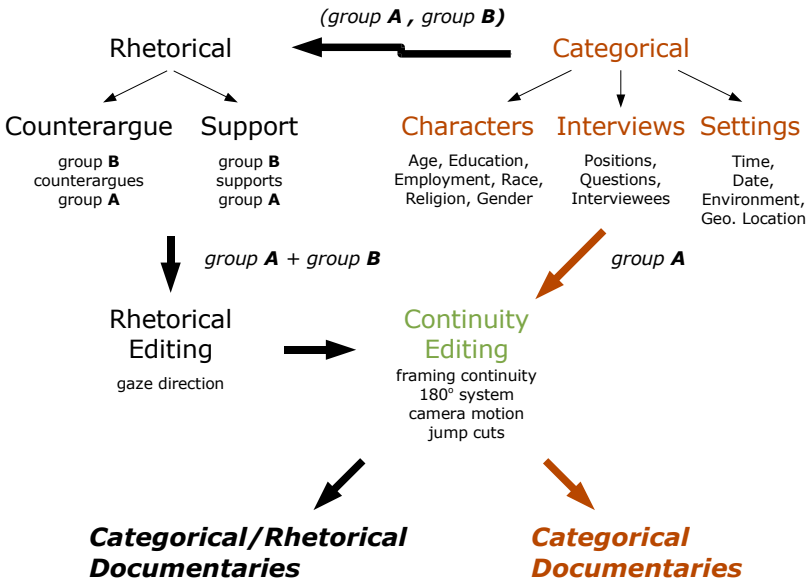

Figure 5.4: The role of the categorical form and the rhetorical form in creating documentaries: dark arrows lead to rhetorical/categorical documentaries, while light arrows lead to categorical documentaries.

with America, is about the war in Afghanistan, anthrax, media coverage and social integration in multicultural societies. Up until now we explained how the generation model can create short sequences consisting of a single argument, such as the one represented in figure 5.2. According to the PRESENTATION FORM [HLR 1] requirement, this represents the micro-level, or scene level (section 2.2.1), for which we are required to use the rhetorical form. To assemble longer sequences (the macro level), we specified with the CATEGORICAL FORM [LLR 9] requirement that the generated documentary should use categorical narrative (section 2.2.2). Categories are used to organize the material in the presentation in an analogous way to how individuals or societies organize their knowledge of the world. In this section we discuss how the rhetorical and the categorical presentation forms can be integrated to produce categorical/rhetorical documentaries. In addition to the PRESENTATION FORM [HLR 1] requirement, this type of documentary also satisfy requirements SUBJECT-POINT OF VIEW [HLR 2] and MONTAGE TECHNIQUE [HLR 4].

5.2.4.1 Categorical documentaries

The relevant categories for interviews documentaries are (section 4.3):

- **Question** asked to the interviewee.
- **Position** expressed by the interviewee.
- **Interviewee** being interviewed.
- **Social categories** the interviewee belongs to, i.e. age, education, employment, race, religion, gender.

- **Time of the day** the clip was shot, e.g. morning.
- **Date** when the clip was shot.
- **Environment** where the clip takes place, e.g. shop.
- **Geographical location** where the clip takes place, e.g. Cleveland (OH).

Categorical documentaries can be generated based on any combination of these categories. For example, clips could be selected based on the question asked and the social categories of the interviewee, for a categorical documentary about low-income, lowly-educated people being interviewed on topics related to the war in Afghanistan. This process is described on the right hand side of fig. 5.4: first the categories are selected by the viewer, and then the corresponding clips are edited using continuity rules (section 5.2.3.2) to form a categorical documentary. The resulting documentary would look like a list of interviews belonging to the selected categories. This presentation form does not satisfy the PRESENTATION FORM [HLR 1] requirement, because positions are not presented in relation to each other, interviewees' statements do not form arguments and the viewer cannot express the point of view the documentary should have. To satisfy the requirement, the categorical form needs to be integrated with the rhetorical form.

5.2.4.2 Categorical/rhetorical documentaries

Categorical documentaries are not suited for matter-of-opinion subjects (section 2.2.3), for which rhetorical documentaries are used. The PRESENTATION FORM [HLR 1] requirement specifies that the rhetorical form must be used at the micro-level. Therefore, with respect to the process of generating a categorical documentary, we add an argument-building step (section 5.2.2.1) between the choice of the categories and the continuity editing.

Using the rhetorical form requires the viewer to select two different groups, *A* and *B*, belonging to different categories. The clips corresponding to these two groups form the content of the documentary. In the rhetorical step (top left of fig. 5.4), arguments are formed with the selected material, so that positions, arguments or statements expressed by group *A* are either supported or counterargued (or both) by statements expressed by group *B*. The selected material is then edited according to rhetorical and continuity editing to produce a documentary, which we call categorical/rhetorical since the selection of the content is based at the macro-level on categories, and the material is organized in arguments at the micro-level. For example, a viewer might request a documentary about the war in Afghanistan, showing how differently educated people think about it. In order to generate a categorical documentary, she would select questions related to the subject "war in Afghanistan", and the categories she wants to see, for example people with low education (*group A*) and high education (*group B*). To generate a rhetorical/categorical documentary, she would need to specify whether the documentary should present the material to show that *A* and *B* are in agreement, in disagreement, or both. This decision would determine whether the system create arguments between *group A*'s and *group B*'s clips so that they support or counterargue (or both) each other, using the techniques explained in section 5.2.2.1. Imagining the viewer would choose that *group B* should counterargue *group A*, the engine would examine the statements contained in clips belonging to *group A* and counterargue them with statements contained in clips from *group B*. The viewer should also express what

point of view should prevail, by specifying that *group A* or *group B* should look more convincing (propagandist), or that they should look equally convincing (binary communicator). The generation process would then bias the presentation of the material by applying the techniques explained in section 5.2.2.2. Imagining she chooses the binary communicator, the generation process selects the content so that the total ethos and pathos value of *group A* clips and *group B* clips satisfies equation 5.2. The type of categorical/rhetorical documentaries that can be generated in this way are able to represent, for example, agreement or disagreement on particular issues within or across races, genders, social classes, age classes, etc. Some of the categories the viewer can select (the Social categories in the above list) are also meaningful for the ethos value of a clip. The viewer's choice of social categories can therefore limit the potential for biasing the argument using ethos. In this case, the expression of a point of view is a best effort attempt, i.e. the process tries to satisfy the viewer request with the available material.

Categorical/rhetorical documentaries implement the presentation forms specified in the PRESENTATION FORM [HLR 1] requirement. Two other requirements are also satisfied by this type of documentary: the SUBJECT-POINT OF VIEW [HLR 2] and MONTAGE TECHNIQUE [HLR 4] requirements. The former is satisfied because the viewer can choose the subject (i.e. the categories) and the point of view of the documentary. The latter specifies that interviews should be presented with the vox populi technique. This montage technique consists of either asking a number of people the same questions, and then presenting the replies together in sequence, or focusing the documentary on one or few people, and show how the statements of these main characters relate to mass opinion (section 2.2.6). The first type of vox populi montage is used when the viewer requests a documentary (either categorical or categorical/rhetorical), using questions as the category to organize material. The second type is used in categorical/rhetorical documentaries, where a category is chosen as the main character (*group A*), and the other category (*group B*) represents the opinion of the mass, which can support and/or counterargue the first group's statements. In section 6.3 we show how we implement categorical/rhetorical documentaries.

5.2.5 Summary

In section 5.2 we designed a process capable of generating documentaries about matter-of-opinion issues, as specified by the high-level requirements in chapter 2. The definition of this process answers *Research Question Generation Process [3].*

The process first generates dynamically the semantic graph, which represents the story space of all the documentaries that can be generated, i.e. the data structure that relates the media items with each other using argumentation relations (GRAPH GENERATION [LLR 6]). This structure supports the generation of documentaries according to the rhetorical form. The graph creation is divided into two phases (section 5.2.1): firstly, all potential relations between statements are calculated, while in the second phase only the relations between statements associated with media items are selected. These relations are then mapped to either SUPPORTS or CONTRADICTS links.

The semantic graph is then used to compose an argument, by complementing an initial interview segment with supporting or counterarguing video clips (section 5.2.2). An argument is formed using defeat and support actions, such as rebuttals and undercutters, which make use of the SUPPORTS and CONTRADICTS links in the semantic graph. Arguments are the building blocks of the rhetorical form, and are composed of supporting or counterarguing clips. When two sets of clips express two opposing posi-

tions and counterargue each other, the total pathos/ethos value of each set determines which position appears more convincing. The generation process uses this information to select clips in order to express the point of view requested by the viewer. For the propagandist point of view, the total pathos/ethos value of one position must be greater than the total pathos/ethos value of the other, while for the binary communicator the pathos/ethos values of the two positions must be comparable.

Having selected suitable material to compose an argument, the video clips must be edited to form a video sequence to be presented to the viewer (section 5.2.3). Rhetorical editing is driven by the need to represent the relation between arguments and clips (argument editing and shot-reverse shot). Continuity editing seeks to produce a video form that is appealing and not disorienting for the viewer, using continuity rules (gaze continuity, framing continuity, interviewee identity continuity, camera movement continuity).

Arguments represent the micro-level, or scene level, for a documentary. A documentary is composed of multiple scenes, and the categorical form provides a way of going beyond a single argument and organizing the presentation of the material on the macro-level (section 5.2.4). Categories can be used to drive the selection of the material, which is composed into arguments on the micro-level, to form a categorical/rhetorical documentary.

5.3 Author support

When using an automatic video generation approach, the documentarist releases control over the final result to the generation system. Other than through annotation, the documentarist has no influence on the run-time generation of the video sequences. She cannot control the quality of what the viewer will see, unless she checks every possible documentary viewers might request, which, for a large repository, is not feasible. A documentarist needs an indication about the documentaries the system can generate without having to view them all. Feedback about the potential quality of the documentaries is therefore essential in automatic generation systems, as we specified in the QUALITY FEEDBACK [LLR 8] requirement.

In our approach, the documentarist creates the thesaurus and annotates the material, as described in chapter 4. At this point, she does not have any insight as to whether the specified terms and relations in the thesaurus describe the media items sufficiently, so that the generation process is able to make the content available to the viewer. Specific problems are whether the clips contained in the repository are actually used by the generation process, or whether a recently added clip will make a difference to the generated documentaries. These questions are especially relevant, considering that our approach aims at making all the relevant material contained in the repository available to the viewer, avoiding potential information loss caused by a final version.

Arguments are the building blocks of the type of documentaries we aim to generate. To compose arguments, the generation process needs to find statements to support or contradict the statements in the initial video clip (section 5.2.2). This is directly related to how well linked the statements are in the semantic graph. If this graph is manually created, as in the approaches adopted by Disc [29] and ScholOnto [63] (section 3.3.1.4), the documentarist would have an idea about how connected the statements in the repository are. In our case, the graph is automatically generated (section 5.2.1) and depends on the quality of the annotations and the thesaurus. The documentarist, however, has no insight on how her annotations will influence the semantic graph. She

0	1-4	5-8	9-12	13
54	47	4	4	9

Table 5.5: Number of statements in the IWA repository having number of links in x-y range (13 is the maximum number of links)

Min	Max	Average
1	1610	203,1

Table 5.6: Result for the generation performance index

may forget to establish relationships between terms in the thesaurus or use a set of terms that does not appropriately express the content of the media items. For example, while working on the IWA repository, it turned out that the generation process was producing insufficient arguments given the number of statements in the annotations. An analysis of the graph creation process showed that 54 statements, of 118 present in the repository, were not linked in the semantic graph, and were effectively lost for the purpose of composing arguments (see Table 5.5). When examining the statements that were not linked, we found that some of them were correctly not linked because their semantics was not contained elsewhere in the repository (e.g. the semantic content "*Israel is a secure country*" was only present in one statement). Other statements, though, should have been linked, since other similar statements were present in the repository.

In order to help the documentarist to verify whether the thesaurus and the annotations are correct, in section 5.3.1 we define three performance indexes. These indexes measure whether the graph creation process produces sufficient links between statements and point out which annotations possibly cause problems. The indexes can also be used to suggest possible relations between the terms so that the graph creation process is optimized.

When generating new statements during the process of graph creation, we limit the number of transformation rounds to three (section 5.2.1). In section 5.3.2 we show that this choice represents a trade-off between the number of created links and computational complexity. The performance indexes can be used to support the documentarist in deciding the optimal number of transformation rounds in her case.

All the methods we introduce in this section have been tested on the Interview with America repository.

5.3.1 Correcting the annotations

In this section we introduce a method to measure the performance of the two steps in the graph creation process (section 5.2.1), with the purpose of helping the documentarist to correct the annotations when the process produces insufficient links. In order to correct the annotations, two operations are possible: re-annotate with different statements the media items that are insufficiently linked in the semantic graph and redefine or add relations between terms in the thesaurus. Changing one statement can solve the problem of one media clip not having sufficient links. Adding or modifying a relation between two terms in the thesaurus can have a greater impact, since this change potentially influences all the statements that contain one of the two terms. Changing relations can therefore improve the performance more than changing statements. Moreover, chang-

Min (%)	Max (%)	Average (%)	Best Percentage
0 (0%)	17 (5,3%)	3,3 (0,9%)	15 (6,6%)

Table 5.7: Results for the linking performance index, as value and as percentage of the generated statements

ing statements can also potentially require redoing many hours of annotating work. Therefore, even when examining why some statements are not sufficiently linked, we consider whether we can draw some conclusions to improve the relations.

The two phases of the graph creation process allow the definition of two different types of performance indexes: the first type measures the potential of the process to generate links (the first phase, section 5.2.1.1), while the second type measure its precision (the second phase, section 5.2.1.2).

The **generation performance index** in section 5.3.1.1 belongs to the first type and can give an indication of whether the terms in the thesaurus are sufficiently related with each other. The **linking performance index** in section 5.3.1.2 belongs to the second type and gives an indication of whether these relations are correct, i.e. whether they capture the semantics of the repository. The **thesaurus performance index** in section 5.3.1.3 also belongs to the second type and determines which relations do not create sufficient links in the semantic graph. This information can help the documentarist in modifying the relations in the thesaurus.

5.3.1.1 Generation performance index

The generation performance index measures the number of statements generated from each statement by the first phase of the graph generation (section 5.2.1.1). This index is calculated for each statement, and gives an indication of the probability of a statement to be linked to others in the repository: the lower the number of generated statements (and thus the index), the less likely it is to find clips annotated with one of them in the repository, and vice-versa. For the sake of clarity, if we apply this index to the example in Table 5.4 on page 84, assuming that the table presents all statements that can be generated from *bombing not effective*, the index for this statement would result in 11 (the number of generated statements).

The generation of new statements depends on the relations defined in the thesaurus for the initial statement's terms (section 5.2.1.1). The more relations the initial terms have to other terms in the thesaurus, the more substitutions can be done, and the more statements can be generated. The generation performance index therefore gives an indication about how well each statement's terms are related to other terms in the thesaurus. A low value informs the documentarist that terms in the statement have few relations in the thesaurus. This can be intentional, or because the documentarist might have overseen potential relations for the terms. The index thus gives an indication to review the relations between the terms used in the statements with a low generation performance value.

This index was applied to the IWA repository and the results are shown in Table 5.6[6]. The most important data in the table is the minimum number of statements generated (column 1), which is 1. This value indicates that there is at least one statement containing terms that are insufficiently related in the thesaurus. The statement

[6]For the sake of clarity, all the tables show summaries, while the indexes are calculated per statement.

with index = 1 was *daily life partially changed*. This was also one of the statements with zero links in Table 5.5. This was clearly a mistake, since the repository contained other statements that should have been linked to it, such as *daily life changed* and *daily life normal*. When we verified the terms in the thesaurus, we found that relations for the term *partially* and *changed* were missing and we added them.

Statements with zero links are not necessarily the result of a mistake. Repositories (especially small ones such as IWA, which contains at the moment 118 statements) are likely to have semantically isolated statements. Nevertheless, this index suggests to the documentarist where to look in the annotations for possible problems hindering the semantic graph creation.

5.3.1.2 Linking performance index

Generating sufficient statements from an initial one only indicates a high probability of creating links. Generated statements contribute to creating links in the semantic graph only if they correspond to existing statements (i.e. to statements that are used to annotate clips in the repository, section 5.2.1.2). The linking performance index measures the percentage of the generated statements which correspond to existing statements, and is also calculated for each statement. For example, in Table 5.4 on page 84, 11 statements are generated. If we assume that 5 of these generated statements correspond to existing statements in the repository, this index would result in the value 5/11 = 45%.

In Table 5.7 we describe the result of applying this index to the IWA repository[7]. We report the lowest number of links created for a statement (column 1), the highest number of links created for a statement (column 2), the average link created per statement (column 3) and the best percentage links created/statements generated (column 4). The minimum value, namely 0, was expected since we knew that 54 statements did not have any link, as shown in Table 5.5. A low positive value for this index is not in itself a problem. This can be because many statements are generated and the repository contains only a few (see for example Table 5.6, 1610 potentially related statements are generated while the repository contains only 118). Nevertheless, a low value can indicate that despite a good performance of the first phase of graph creation (section 5.2.1.1), as measured by the generation performance index, the process performs poorly in the second phase (section 5.2.1.2). This could be an indication that the first phase generates "nonsense", i.e. statements that, from the point of view of the repository semantics, do not make sense and do not correspond to any existing statements. A high number of generated statements (the first performance index) indicates that the terms are well connected in the thesaurus, but a low linking performance index indicates that these relations are not able to capture the semantics of the material contained in the repository. This index suggests to the documentarist that there might be a gap between the semantics of the thesaurus and the annotations in the repository. The documentarist might consider revising the terms (and their relations) used for the statements with low values. The maximum value in Table 5.7 gives an indication of what a high percentage value can be: 5,3% when finding the most existing statements (or 6,6% when considering the best ratio between existing statements and generated statements). Assuming the documentarist correctly entered the relations for the statements with the best performance, these values provide an indication of the upper limit

[7]We discuss here only the meaning of the minimum and maximum values, since at the moment we do not have any golden standard against which to compare average and best percentages, which we report for completeness.

Term 1	Relation	Term 2	Hit Score	Miss Score
people	*Specialization*	*normal people*	8	3802
I	*Generalization*	*people*	10	3084
no mod	*Similar*	*best*	3	2292
not	*Similar*	*never*	8	2253
people	*Specialization*	*rich people*	0	2110
fearful	*Similar*	*attentive*	8	2049
no mod	*Similar*	*can*	0	1925
Americans	*Generalization*	*people*	5	1837
no mod	*Similar*	*always*	11	1755
war	*Generalization*	*violence*	1	1460

Table 5.8: Worst 10 relations based on "miss" score

Term 1	Relation	Term 2	Hit/Miss
only	*Opposite*	*not only*	16,6%
economic-aid	*Opposite*	*bombing*	16,0%
not only	*Opposite*	*only*	15.8%
waste	*Opposite*	*effective*	14,0%
diplomacy	*Opposite*	*war*	11,9%
economic-aid	*Opposite*	*war*	10,8%
not best	*Opposite*	*no mod*	10,5%
ground forces	*Opposite*	*bombing*	9,9%
only	*Similar*	*not only*	8,9%
not only	*Similar*	*only*	8,3%

Table 5.9: Best 10 relations based on $\frac{hit}{miss}$ ratio

for this index.

5.3.1.3 Thesaurus performance index

The performance indexes introduced so far indicate which statements do not contribute to generating a well-linked semantic graph. In this section, we introduce an index to measure the performance of the relations in the thesaurus, the thesaurus performance index. To calculate it, we keep track of the relations used to generate each statement. If a generated statement corresponds to an existing statement, the relations used to generate it cause a **hit**, otherwise the relations cause a **miss**. The hit and miss scores form the basis of this performance index.

For example, the fifth statement in Table 5.4, *peace not effective*, has been generated using two relations: relation *Generalization* between *bombing* and *war*, and then relation *Opposite* between *war* and *peace*. Each relation gets 1 point on the hit score if *peace not effective* is also an existing statement, 1 point on the miss score otherwise.

Calculating this index for the IWA thesaurus, we found that out of 199 relations, 101 had a hit score of zero, i.e. they were never able to generate a hit. Relations with a zero hit score do not contribute to create the semantic graph: the same links would be generated even if these relations would not be present in the thesaurus. The documentarist should consider eliminating them (at least as long as the content of the

Term 1	Relation	Term 2	Hit Score	Miss Score
war	*Specialization*	*bombing*	45	970
bombing	*Generalization*	*war*	39	746
no mod	*Opposite*	*not*	32	876
not	*Opposite*	*no mod*	30	684
military-actions	*Similar*	*war*	22	363
waste	*Opposite*	*effective*	21	150
war	*Similar*	*military-actions*	21	527
effective	*Opposite*	*waste*	21	1561
not best	*Opposite*	*no mod*	12	114
bombing	*Opposite*	*ground forces*	1	1460

Table 5.10: Best 10 relations based on "hit" score

Term 1	Term 2	Hit Score	Miss Score
USA	*war*	91	11228
bombing	*war*	52	6416
I	*war*	52	6416
normal people	*war*	52	6416
terrorists	*war*	52	6416
current world	*war*	52	6416
American media	*war*	39	4812
Muslims	*war*	39	4812
casualties	*war*	39	4812
daily life	*war*	39	4812

Table 5.11: Best 10 **suggested** relations for the subject part of the statement based on "hit" score. The method suggests only that there should be a relation, but not which relation.

repository does not change) or modifying them.

Table 5.8 shows the relations with the highest miss score. Among them, the relations with a zero hit score could be deleted as they only consume computer resources. For the others, the documentarist should consider whether they describe a few, but valuable, semantic options, which are simply kept, or if they are misconstructed and should be modified.

For example, the first and the fifth rows in Table 5.8 indicate a semantic distinction between *normal people* and *rich people*, both *Specialization* from *people*. This distinction, though, generates a large number of miss scores (almost 6000) in comparison to 8 hits. This semantics apparently is weakly represented in the repository as far as the annotations are concerned. In fact the statement s:people p:threatened generates 1342 statements, of which 4 corresponded to existing statements (0,3%). The documentarist can now see the relations causing the low number of hits and thus address the problem.

Table 5.9 shows the best 10 relations according to the ratio $\frac{hit}{miss}$. Relations with a high value for this ratio capture the semantics of the repository and should not be changed, unless the documentarist makes a conscious decision to change the semantics of the repository. For completeness, the best 10 relations according to the hit score are shown in Table 5.10.

The linking performance index can provide further guidance to the documentarist, by suggesting how to relate terms in the thesaurus. Imagine the documentarist has annotated the media items with terms from a thesaurus, which she built, but has not related the terms yet. Or she has already done that, but she wants an indication whether she has captured the semantics in the repository. The linking performance index can be used to suggest the best relations from the point of view of the efficiency in generating hits.

To achieve this, the generation process assumes that each term in the thesaurus is related to each other term. The particular type of the relation is not important at this stage and the process assumes the relation is unspecified. What matters is that each term is related to all other terms, which means that:

- all the potential relations between terms are contained in the thesaurus,
- every term can be replaced by any other term. All the statements that can be formed with the terms in the thesaurus are therefore generated.

Calculating the linking performance index with a fully-related thesaurus provides an indication of the best potential relations between terms. Table 5.11 shows the result of applying this method to the IWA repository (apparently, *war* is present in many clips and forms a sort of focus of the annotated material). A potential relation between two terms does not necessarily mean that the terms can be related in the thesaurus: the relation needs to be equivalent to either *Similar*, *Opposite*, *Generalization* or *Specialization*. The suggestion mechanism is numerical and does not take semantics into account. Only a few of the suggested terms in Table 5.11 can be actually related using one of the thesaurus relations (e.g. *bombing Similar* or *Specialization war*). It is up to the documentarist to decide whether to accept the suggestion or not. If she does, she also has to decide which specific relation applies between the selected terms.

5.3.2 Fine-tuning the semantic graph creation process

In the first phase of the semantic graph creation, we use three rounds of transformations for each statement in order to generate new statements (section 5.2.1.1). This number is based on a trade-off, which we explain in this section. On one hand, the more rounds of transformations we apply, the more statements we generate and therefore the more links we can potentially create (section 5.3.1.1). On the other hand, increasing the number of rounds increases the time used by the generation process. When dealing with graphs, **computational complexity** is something that has to be investigated, since our approach might not scale to large repositories, even though the semantic graph is built only once (if the annotations do not change), rather then each time a viewer requests a documentary.

To estimate the time required, we recall that the graph is created in two steps: for each statement, the process generates all possible statements according to the relations in the thesaurus and then checks which of these statements correspond to existing statements (sections 5.2.1.1 and 5.2.1.2). The time required by both steps T is proportional to the number of generated statements S_g. An estimate of S_g is:

$$S_g = S\left(\frac{(P_s\overline{R_t})^{(N_r+1)} - 1}{(P_s\overline{R_t}) - 1}\right) \approx S(P_s\overline{R_t})^{N_r}$$

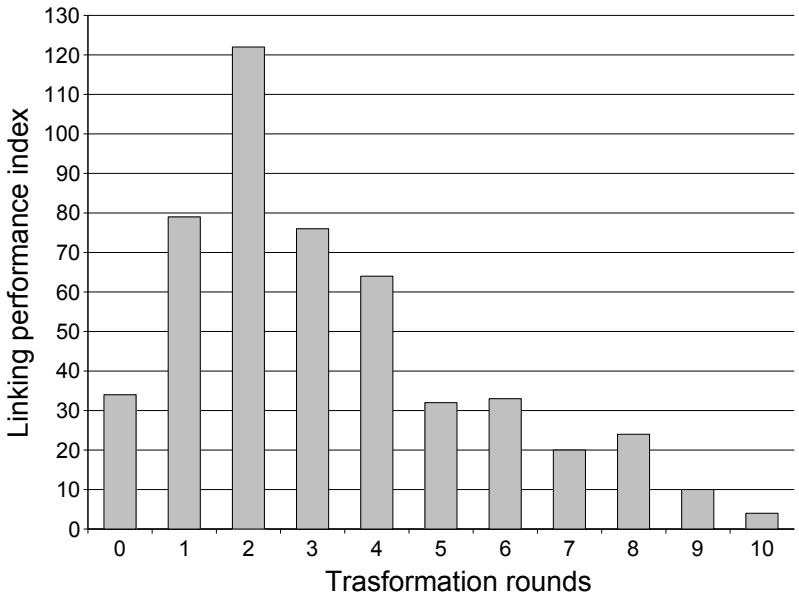

Figure 5.5: The increase in the linking performance index as a function of the number of transformation rounds (0 rounds means the statement is not transformed, i.e. the repository contains some equal statements)

where S is the number of statements in the repository, $\overline{R_t}$ is the average number of relations each term in a statement has to other terms in the thesaurus, N_r is the number of transformation rounds per statement, and P_s is the number of parts in a statement.

The estimate is derived by observing that at each step the statements generated from an existing statement increase by a factor $(P_s\overline{R_t})$. For example, if the number of parts in a statement is three ($P_s = 3$), and each of these three terms can be substituted with three other terms ($\overline{R_t} = 3$), 9 new statements can be generated, and so on[8]. The number of generated statements is then equal to the sum of a geometric series for which the following formula holds:

$$\sum_{k=0}^{n} q^k = \left(\frac{q^{k+1} - 1}{q - 1}\right)$$

which can be approximated to q^k when $q \gg 1$.

S_g grows linearly with the number of statements S, which makes our approach scalable to larger repositories. On the other hand, it grows polynomially with the average number of related terms $\overline{R_t}$ and parts of the statement P_s, while it grows exponentially with the number of transformation rounds N_r. To avoid the undesirable consequences on performance due to exponential growth, we need therefore to limit the number of transformation rounds. To establish a good trade-off between results obtained and time required to obtain them, we need an indication about how many statements are retrieved as a function of the number of transformation rounds.

[8]Actually the increase per transformation round is lower, since some of the related terms at the n^{th} transformation round have already been used at the $(n-1)^{th}$ round.

The linking performance index (section 5.3.1.2) can be used for this purpose. We present the result of applying this method to the IWA repository, by showing how the index increases as a function of the number of transformation rounds (fig. 5.5). The figure forms a Gaussian curve, starting with a small increase in the index with zero transformations (i.e. some statements are present more than once in the repository), growing to a peak with two transformations and then slowly decreasing. After a certain point it seems that the generated statements become semantically too far removed from the original content of the repository, as if too many manipulations have led to statements that make no sense or make no sense in the domain of the repository. The chart reported in fig. 5.5 provides an indication of when to stop. Even though each transformation results in more retrieved statements, limiting the transformations to 3 or 4 represents a good trade-off between results obtained and time required. For our repository, with three transformation rounds, the graph is created in about 90 seconds on a i686 athlon running Linux. The semantic graph must be calculated every time an item is added to the repository[9]. Depending on whether items are added frequently or not, the documentarist can choose a shorter graph creation time or more results. We chose a shorter graph creation time since we needed to test frequently the implementation of the model.

5.3.3 Summary

In section 5.3 we discussed the problem of providing the documentarist with an indication of the capability of our approach to generate documentaries, without having to check every possible viewer-requested documentary. Since our documentary generation is based on the capability to compose arguments, and this in turn is based on the relations between statements in the semantic graph, we chose to examine whether the graph contained sufficient relations. We found that this was not the case for the graph created using the IWA repository. Relations are automatically created by the generation process based on the annotations provided by the documentarist. The quality of the annotations influences the process of graph creation. We introduced therefore a method for the documentarist to verify and correct the annotations used to describe the media items in the repository, as specified by the QUALITY FEEDBACK [LLR 8] requirement and to provide an answer to *Research Question Author Support [4].*

The two phases of the graph creation process allow the definition of two different types of performance indexes: the first type measures the potential of the process to generate links in the first phase, while the second type measures its precision in the second phase. The generation performance index belongs to the first type (section 5.3.1.1), while the linking performance index (section 5.3.1.2) and the thesaurus performance index (section 5.3.1.3) belong to the second. These indexes point out possible problems in the annotations, focusing on the generated statements (the first two) and on the relations in the thesaurus (the last one). The thesaurus performance index can also be used to help the documentarist in crafting the thesaurus, by suggesting which terms should be related. The indexes were tested on the IWA repository, leading to improvements for the graph creation process.

Finally, we analyzed the computational complexity of the graph creation process to check whether our approach would scale to larger repositories and to study the trade-off between time required and results obtained (section 5.3.2). We found that com-

[9]The use of incremental update methods would reduce the time required to build the semantic graph. Nonetheless, these methods are also exponentially dependent on the number of transformation rounds.

putational time depends linearly on the number of statements in the repository (thus our approach is scalable), and exponentially on the number of transformation rounds. We then showed how the linking performance index can provide an indication for the trade-off between results obtained and computational time, by measuring the relation between number of links created and the number of transformation rounds needed.

Chapter 6

Implementation

In the previous two chapters we described the automatic generation model satisfying the requirements we set in chapters 2 and 3. In this chapter we present an implementation of this model, called Vox Populi. We describe how the viewer can request the system to generate a documentary and the tools the documentarist can use to author and maintain a documentary repository. We applied the system to the IWA repository. Vox Populi is also used in two other projects, which we discuss to show the general applicability of the model. We then provide a brief technical evaluation and some general considerations on automatic video generation approaches.

6.1 Introduction

In this chapter we describe **Vox Populi**[1], an implementation of the video generation model that we made to demonstrate the model's functionality, and to test the annotation schema (chapter 4) and the generation process (chapter 5). Testing has led to corrections to our model, as well as to added functionality. For example, we realized an author support mechanism is necessary because we needed to check why the system was not performing as expected (section 5.3).

The focus of this chapter is on the technical details of the model implementation (section 6.2) and on how the viewer and the documentarist can use the system. We explain how a viewer can request and view a documentary (section 6.3) and how a documentarist can author the repository from which documentaries are generated (section 6.4). Vox Populi has also been used outside the scope of our own work. Two projects have adopted it, in two different settings: one in the domain of art and the other in the domain of ambient environments (section 6.5). We then provide a technical evaluation of Vox Populi's graph creation process (section 6.6), since this process determines Vox Populi's performance by creating the story space from which all the documentaries can be generated. Finally, we report some conclusions on automatic video generation approaches derived from our experience as documentarists for the IWA project (section 6.7).

[1] We already used the same term to indicate a technique used in interview documentary (section 2.2.6). To avoid confusion, we refer to the system as Vox Populi and to the technique as vox populi, which is consistent with previous chapters.

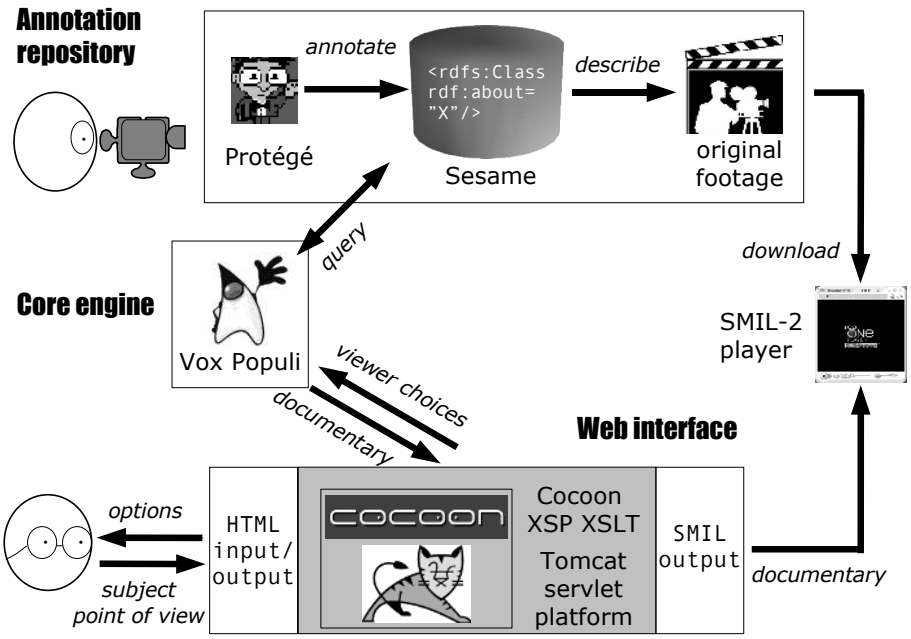

Figure 6.1: The general architecture of Vox Populi

6.2 Vox Populi's architecture

Vox Populi's general architecture is represented in fig. 6.1. From a functional point of view, three components can be distinguished: the web interface, the annotation repository and the core engine.

The web interface (shown in fig. 6.2) allows the viewer to input the subject and the point of view of the documentary. This interface is generated using Cocoon[2], a web development framework that provides active server pages functionality (XSP[3]) and an XSLT [18] engine with HTML serialization. Cocoon runs on top of Tomcat[4], a servlet container. XSP makes it possible to call the core engine and present the viewer with the options she can choose from. XSLT is used to serialize all the options produced by the core engine to HTML so that the Web interface can display them to the viewer. Once the viewer has specified her choices, XSP transmits them to the core engine and requests the generation of the documentary.

The core engine is implemented as a Java package which XSP can call from within the web environment. This package queries the annotation repository to retrieve the options to be presented to the viewer and creates the semantic graph needed to generate documentaries (section 5.2.1). Once the viewer has input her choices, the core engine generates the documentary as SMIL-2 format [74], which the viewer can watch with any SMIL-2 player. In our setting we use RealPlayer from RealNetworks. SMIL-2 output does not embed the media files but points to where they are stored. It is the task

[2] http://cocoon.apache.org/
[3] http://cocoon.apache.org/2.1/userdocs/xsp.html
[4] http://jakarta.apache.com/tomcat

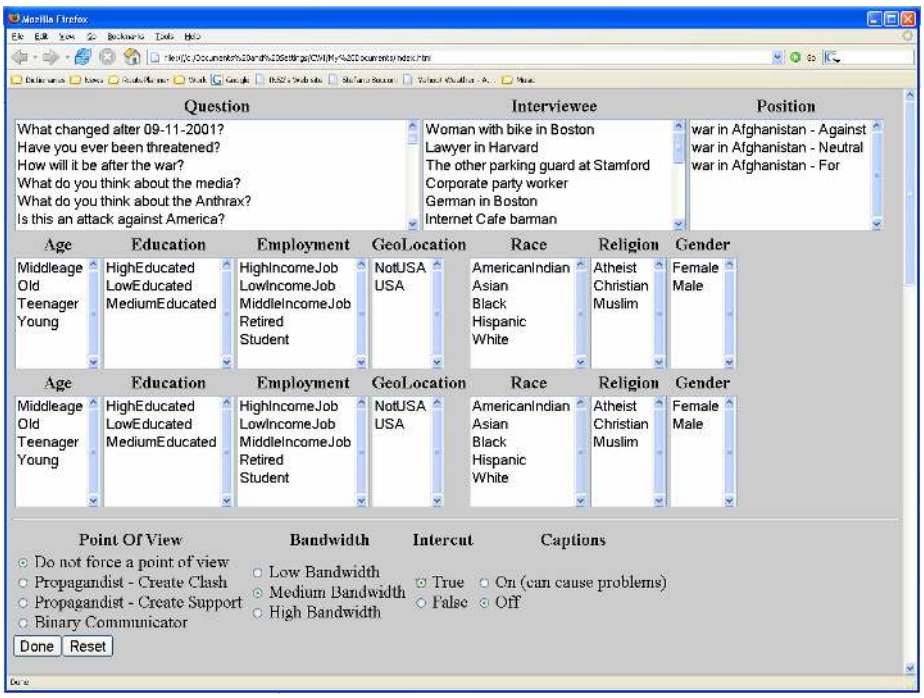

Figure 6.2: Vox Populi web interface

of the player to download them from their location.

Annotations are encoded in RDF(S) [75] and are stored using Sesame [14], an RDFS storage server which supports query-based access to the data. The documentarist can create all the required annotations using Protégé [31], a knowledge-base editor.

6.3 Vox Populi for the viewer

In this section we describe how the viewer can request a documentary using the Web interface (section 6.3.1) and we examine the properties of the generated SMIL output (section 6.3.2).

6.3.1 The web interface

Here we concentrate on Vox Populi's Web interface and how a viewer can interact with it, omitting the details of the XSP/XSLT code. Through this interface (see fig. 6.2), the viewer can request documentaries of the categorical and the categorical/rhetorical types (section 5.2.4), by specifying the subject and the point of view of the documentary. In order to generate a documentary, the viewer needs to select first the subject and the point of view.

6.3.1.1 Selecting the subject

At the top of the window, three select boxes contain the **Question** asked during the interviews, the **Interviewee** and the **Position** expressed in the material contained in the repository. The second and third row each present a list of select boxes corresponding to the social categories of the interviewees (defined in section 4.2.3). To choose the subject, the viewer can:

- select one or more **questions** in the Question select box. All the clips showing an interviewee who replies to this/these question(s) are retrieved. If the viewer selects also one or more interviewees, only clips showing that/those interviewee(s) are selected.
- select one or more **positions** in the Position select box. All the clips where this/these position(s) is/are expressed are retrieved. If the viewer selects also one or more interviewees, only clips showing that/those interviewee(s) are selected.
- select one or more **interviewees** in the Interviewee select box (interviewees are indicated with a small description, e.g. "Lawyer in Harvard"). All the clips where this/these interviewee(s) is/are shown are retrieved. If the viewer selects also one or more questions, only clips showing that/those questions(s) are selected. Otherwise, if the viewer selects also one or more positions, only clips showing that/those positions(s) are selected.
- select a **class of interviewees** by categories, by selecting in the second row the social categories the interviewees belong to. All the clips where this class of interviewees is shown are retrieved. If the viewer selects also one or more questions, only clips showing that/those questions(s) are selected. Otherwise, if the viewer selects also one or more positions, only clips showing that/those positions(s) are selected.

The viewer can combine two selections, with the exception of questions asked and positions expressed, which cannot be selected independently, since positions are expressed as a result of questions being asked to interviewees.

6.3.1.2 Selecting the point of view

After the content for the documentary has been specified, the viewer can select the **Point of View** she wants the documentary to have, using the radio-buttons on the bottom left of fig. 6.2. If the viewer wishes to see a documentary of the categorical type (section 5.2.4.1), based on the categories she has selected, she must choose *Do not force a point of view*. The documentary will show a sequence of the chosen clips in no particular order, edited with the continuity rules defined in section 5.2.3.2.

If, however, the viewer wishes to see a documentary of the categorical/rhetorical type (section 5.2.4.2), which presents arguments, she needs to select at least another group that will support or counterargue the statements expressed by the group she has already selected. This initial selection represents *group A* and a new selection is needed for *group B*, and possibly for *group C*, depending of the chosen point of view. A categorical/rhetorical documentary can have three different points of view:

- **Propagandist - Create clash** causes Vox Populi to present *group A* being counterargued by *group B*. *Group A* is selected as explained in section 6.3.1.1, and

can be either a single interviewee or a group. *Group B* is always a class of interviewees, which can be selected from the classes in the third row of select boxes. To create the counterarguing effect, the system presents only matching clips, i.e. it presents a clip from *group A* only if it can find a counterarguing clip from within *group B*. Furthermore, the system tries to present *group B* as more convincing than *group A*, using the techniques introduced in section 5.2.2.2, i.e. ethos, pathos and parallel weakening. The calculation of the ethos value is currently based on a default user profile, which defines high education as important for the viewer.

- **Propagandist - Create support** causes Vox Populi to present *group A* being supported by *group B*. Analogously to the *Create clash* case, *group A* is selected as explained in section 6.3.1.1, and *group B* is selected from the classes in the third row of select boxes. The system presents only matching clips, i.e. it presents a clip from *group A* only if it can find a supporting clip from within *group B*.

- **Binary Communicator** causes Vox Populi to present opposing positions to balance the strengths of the two opposing parties (section 2.2.3). In our implementation, we use this point of view to complement the position of a main character with statements of other interviewees. Therefore, the subject of the documentary must be selected so that there are three groups: the viewer selects the main characters in the first row (*group A*), while the second and the third groups (*group B* and *group C*) are selected in the second and third rows, respectively. The generated documentary presents the main character (*group A*), followed by *group B* supporting it and then *group C* producing the counterargument. Since a binary communicator tries to give equal importance to both sides, Vox Populi tries to select and present *group B* as convincing as *group C*, using the techniques introduced in section 5.2.2.2.

The interface presents three more options the viewer can select:

- **Bandwidth** of the connection, Low, Medium and High. This option causes Vox Populi to use video clips rendered at 20 Kbps, 100 Kbps and 225 Kbps, respectively.

- **Intercut**, on or off (section 5.2.3.1). If Intercut is off, Vox Populi presents *group A*'s video clips entirely, and then *group B*'s (and *group C*'s if a third group is required). If Intercut is on, *group A*'s clips are intercut whenever a statement can be matched (i.e. either supported or counterargued) by statements in *group B*'s clips. This second presentation mode generates more dynamic documentaries since the length of the clips is shorter.

- **Captions**, on or off. If the captions are on, additional information is displayed in the caption area (see fig. 6.5), such as the question asked, the interviewee or her statements.

The documentaries are generated using the vox populi montage technique (section 2.2.6), which allows to show a range of arguments from different people (multiple POV) or to add to a video interview focused on a particular person, or on a few persons, the different opinions of the men-on-the-street (single POV). Vox Populi uses multiple POV, with the exception of the binary communicator point of view (*group A*), which has a main character and therefore uses a single POV.

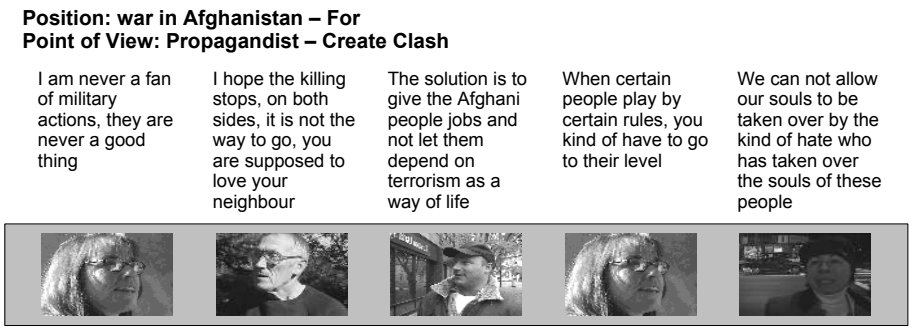

Figure 6.3: Documentary generated with Position "war in Afghanistan - For", Interviewee "Lawyer in Harvard", Point of View "Propagandist - Create Clash", Intercut on, group B "Race" = "White" (third row of select boxes).

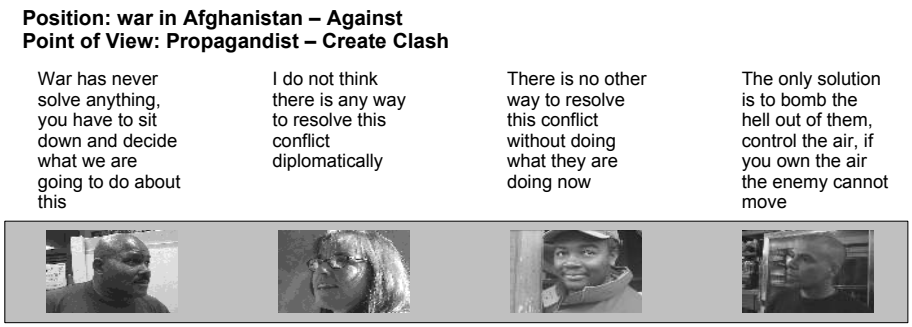

Figure 6.4: Documentary generated with Position "war in Afghanistan - Against", Interviewee "Black shop owner Stanford", Point of View "Propagandist - Create Clash", Intercut on.

Vox Populi's interface aims at showing the capabilities of the generation model we defined in the previous chapter. Even though the viewer can start exploring the repository specifying only a few options (for example, only the question asked), she needs to specify all the choices we discussed above to exploit the characteristics of our generation model. Freeing the viewer from the fact that selection of material and bias are made by somebody else requires the viewer to take charge of these activities for herself.

6.3.1.3 Examples

We give two simplified examples of generated documentaries here, the simplification being that some clips are not shown in the figures for the sake of clarity. In fig. 6.3 we show a documentary generated specifying for the **Position** "war in Afghanistan - For" and as **Interviewee** "Lawyer in Harvard" (the woman shown on the left). The **Point of View** is "Propagandist - Create Clash" and the counterarguing group (*group B*) is

Figure 6.5: A generated documentary being viewed with a SMIL player (Real Player)

selected to have **Race** "White" in the third row of select boxes. The **intercut** option is on. In fig. 6.4 we show a documentary generated specifying for the position "war in Afghanistan - Against" and as interviewee "Black shop owner Stanford" (the man on the left). The **Point of View** is "Propagandist - Create Clash", the counterarguing *group B* is not restricted to belong to any category. The **intercut** option is on. In both documentaries the presentation is biased so that the position of the chosen interviewee is weaker than that of the counterarguing party, which is composed of more interviewees, leading to a higher pathos/ethos value. Both examples are generated using the "gaze matching" editing rule (section 5.2.3.1).

6.3.2 The SMIL output

Once the documentary has been generated, the viewer can watch it with any SMIL player (we use RealPlayer version 10, see fig. 6.5). SMIL supports several types of media, of which we use text, image and video. In fig. 6.5 a video clip is played in parallel with a text caption area and a browsing area (explained below).

Video clips can be specified in SMIL using the attribute *src* for the URL of the file and *clipBegin*, *clipEnd* for the begin and end time within the file, for example:

```
src="rtsp://media.cwi.nl/IWA1.2_100.rm"
clipBegin="00:00:38.672" clipEnd="00:01:06.867"
```

Files containing the footage can thus be dynamically edited into a sequence without being physically modified. Applications using SMIL do not need to implement video editing functionality, and media files can be stored in different locations, as long as they can be accessed using URL. Vox Populi can thus use different repositories to generate documentaries, providing they are annotated according to the schema defined in chapter 4.

SMIL also supports hyperlinks. To allow the viewer to browse the content of the repository more easily, we include three hyperlinks in the generated documentary.

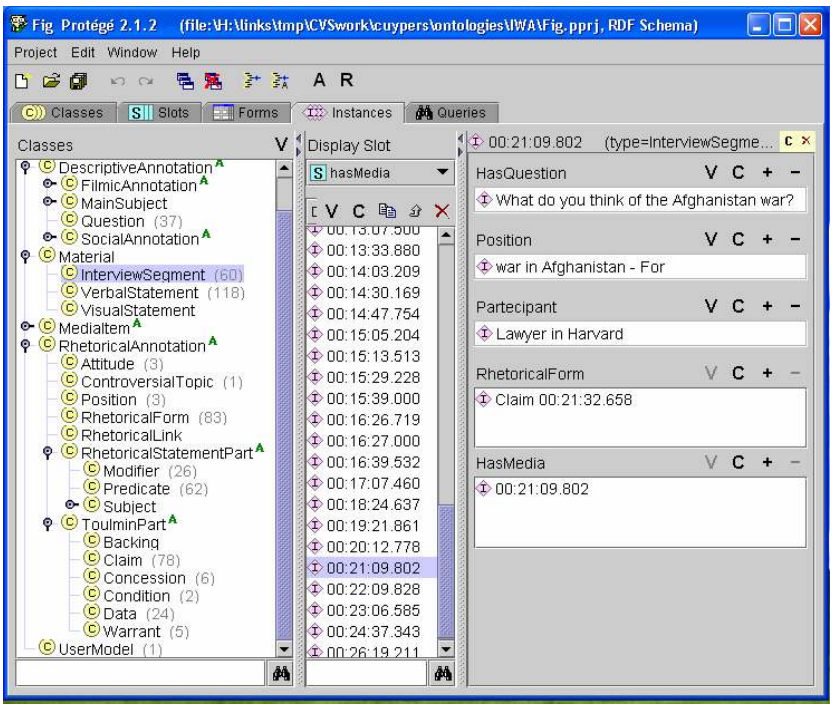

Figure 6.6: Protégé, the editor used to encode the annotations

These are represented by three text areas (“Bin Comm”, “Clash” and “Support”) which can be seen at the bottom of fig. 6.5, under the caption area, and are color-coded, from left to right, yellow, red and green. By selecting any of them, the viewer requests Vox Populi to generate another documentary where the interviewee currently presented in the video at the moment the link is clicked will become the initial character. If the viewer clicks on the yellow “Bin Comm” link, the documentary will have the point of view “Binary Communicator”. If the chosen link is the red “Clash” one, the documentary will have the point of view “Propagandist - Create clash”, while if it is the green “Support” link Vox Populi will generate a documentary with the Point of view “Propagandist - Create support”.

6.4 Vox Populi for the documentarist

In this section we describe the tools for a documentarist to annotate her material (section 6.4.1), which we used to create the IWA repository. We then discuss how a documentarist can influence the generation of a documentary by specifying how Vox Populi should apply rhetorical and cinematic editing rules (section 6.4.2).

6.4.1 The annotation tools

Protégé was used to create the annotation schema as well as the instances. In fig. 6.6 Protégé’s interface is shown, with our annotation schema on the left and one particular instance of an interview segment annotated on the right. Parts of the schema hierarchy

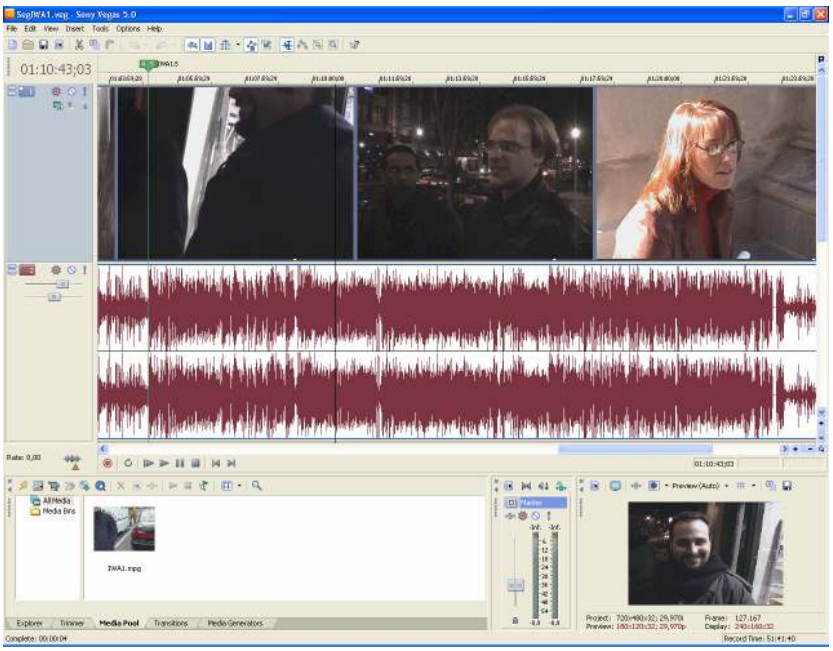

Figure 6.7: Vegas Video, the video editing software used to annotate footage

are visible in more detail, for example the statement structure (class *RhetoricalStatementPart* with subclasses subject, modifier and predicate) and the Toulmin structure (class *ToulminPart* with subclasses backing, claim, concession, condition, data and warrant).

The annotations are encoded in RDF and stored in Sesame, an open source RDF database with support for RDF Schema inferencing and querying. Sesame can use different storage systems (e.g. relational databases, in-memory, file systems), and offers an access API, which supports both local and remote access (through HTTP or RMI[5]). For better performance, we use the in-memory option for local access, but Vox Populi can also access a remote repository, which is necessary in case the annotations are distributed. Furthermore, Sesame provides access to the data using a query language called SeRQL [2]. A possible alternative to RDF for media annotations is MPEG-7 [35], which we did not choose since RDF offers better support for annotation tools.

The clips were annotated using Sony Vegas[6] to view the content (see fig. 6.7). We manually time-stamped the footage to annotate the start and end point of the interviews segments. Sony Vegas includes functionalities such as noise reduction and color correction in case the original quality of the video material is inadequate. The only feature we used was reverse rendering, which is needed to have interviewees looking in both directions to apply the gaze-matching technique (section 5.2.3.1). To reduce the documentary download time for the SMIL player, we segmented the files into shorter parts of about 20 minutes each and we rendered each part using the RealMedia[7] format for

[5] `http://java.sun.com/products/jdk/rmi/`

[6] `http://www.sonymediasoftware.com/products/vegasfamily.asp`

[7] The use of RealMedia limits the possible players to only RealPlayer, but rendering to MPEG is also possible.

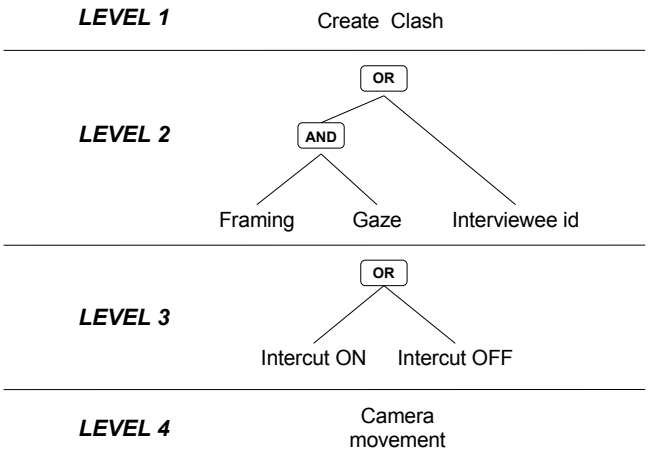

Figure 6.8: An example of a tree of rules

20, 100 and 225 Kbps for low, medium and high bandwidth respectively.

The repository we use, "Interview with America", contains approximately 8 hours of video interviews. Of this, one and a half hours of video has been annotated, containing 15 people being interviewed and 37 questions asked. In appendix D we report the interviewees description, the positions annotated and the questions asked. The annotations cover 185 video segments annotated by 118 statements. The thesaurus we use contains 155 terms and 199 relations between the terms (counting symmetric relations twice). It is not easy to estimate how much time it took to annotate the material, since the annotation schema was refined based on insights gained during the development, and the annotations had to be modified several times. A rough estimate would be 40 hours for one person. This is a considerable amount of work, but comparable to the annotation effort (e.g. transcripts) normally used in traditional documentary making (section 2.3.5). A means of reducing the annotation effort would be to use automatic video segmentation techniques, in particular audio-based segmentation [76].

6.4.2 Authoring rules architecture

When using our generation model, a documentarist can influence the documentary by specifying the priority in which the generation process applies rhetorical and cinematic rules, such as those described in sections 5.2.2 and 5.2.3. Vox Populi implements a particular architecture in order to support this mechanism. The rules Vox Populi provides operate on two data structures: the **Story** and the **Fabula**. These two structures are inspired by the homonym concepts in narrative ([4] p. 5): a fabula is a series of logically and chronologically related events that are caused or experienced by the actors in the story, while the story is the exposition of the fabula, in which facts are selected and presented in an artistic manner (e.g. using flashbacks, flash-forward, etc.). In Vox Populi the fabula coincides with the *Semantic Graph* we defined in section 5.2.1, i.e. it contains all the descriptions of the clips in the repository and the typed links among

them[8]. The story contains the ordered sequence of clips which have been selected for the documentary to be generated. Each rule can either eliminate clips from the fabula (e.g. if a rule specifies that only females must be shown in the documentary, all clips containing males are eliminated), or place clips from the fabula at a particular place in the story, e.g. place the main character at the beginning. A rule can place more candidate clips at a certain position in the story, leaving subsequent rules to select which of these clips should be removed. In this way each rule operates on the selection and story arrangement determined by the previous rule. After all rules are applied, the story becomes the generated documentary which is shown to the viewer. To determine the order in which they are applied, rules are organized in a hierarchical structure (see fig. 6.8, where we show a tree which is binary for simplification but can also be non-binary). The documentarist specifies such a structure creating a script or using one of the scripts we provide. Each rule is placed at a certain level, which determines the order of execution and the importance of the rule: higher rules have more importance. Thus, if rule B is applied after rule A, rule A makes editing decisions which rule B can only refine, but not change. A rule engine starts applying rules to the selected clips, and checks after each rule whether the result is satisfactory or not. If it is, a valid video sequence satisfying the rules has been found, otherwise the rule engine goes on with applying rules. To judge whether the solution is good enough, the rule engine calculates a weighted sum of the rules that have been executed, giving more weight to the higher rules than to lower ones. The calculation depends on the result of the rule and on the level of the rule. If this sum is greater than a threshold set by the documentarist in the script, Vox Populi considers the solution to be good enough.

We use the following formula:

$$\frac{\sum_{k=1}^{N_r} q_k \, 2^{(N_l - L_{r^k})}}{\sum_{k=1}^{N_r} 2^{(N_l - L_{r^k})}} \leq threshold$$

where N_r is the number of rules, q_k is the numerical result of rule number k (either 0 = failure OR rule not executed yet or 1 = success), N_l is the total number of levels, L_{r^k} is the level of rule k and $threshold$ is expressed as percentage. This formula calculates the weighted percentage of completed rules, giving more weight to rules on higher levels.

The Boolean operators AND and OR provide means to group rules together. Rules can be grouped on a single level using AND, in which case their weight is the same. Every OR causes the rule tree to branch. If the rule engine executes a leaf rule without reaching the threshold, it will backtrack and try the other path at the point of the OR-branching. For example, in fig. 6.8 there are 4 levels and 7 rules. The order of execution is *Create Clash - Framing - Gaze - Intercut ON - Camera movement* (unless the threshold is achieved before executing *Camera movement*, in which case the engine stops). In the case the threshold is not achieved, the rule engine tries other paths, if there are any. In our example, assuming the engine has to try all paths because the threshold is not achieved, it would try first *Create Clash - Framing - Gaze - Intercut OFF - Camera movement*, then *Create Clash - Interviewee id - Intercut ON - Camera movement* and finally *Create Clash - Interviewee id - Intercut OFF - Camera movement*. As soon as the threshold is reached, Vox Populi presents the result. If the threshold is not reached, the rule engine gives a warning but still presents the result of the path with the highest score.

[8]Strictly speaking, the *Semantic Graph* can not be equated to the fabula, since, depending on the repository, it can contain zero, one or more sets of causally related events, and therefore zero, one or more fabulas.

Different values for the threshold represent different trade-offs between computational time required to generate the documentary and quality of the generated documentary. If the documentary must be served quickly, a low threshold will cause Vox Populi to stop sooner without applying all the rules set by the documentarist, resulting in a less accurate documentary. In this case, rules in higher positions have more chances of being satisfied, so that a documentarist can give priority to the visual aspect or to the rhetorical aspect of the documentary. For example, a group of rules could be composed by the continuity rules (*Framing*, *Gaze*, *Camera movement* and *interviewee identity*) and another one by the rhetorical rules (*gaze matching* and *Intercut ON/OFF*). By placing the first group at a higher level, the documentarist requires Vox Populi to generate a documentary with preference for cinematic properties, while putting the second group at a higher level would produce a more rhetoric-driven documentary. For example, in fig. 6.8 the documentarist has given priority to the cinematic properties of the result, by placing cinematic rules at a higher level. The documentarist can also try both possibilities by combining them with OR and leave the system to determine which rule path achieves the higher score.

The rule support mechanism described in this section represents a means for the documentarist to retain some control on the outcome of the generation process. At the moment the rules structure as described in this section is inserted using scripts. The documentarist would need a better interface which, were it available, could also be used by the viewer. In this case the viewer could also have a role in this part of the process.

6.5 Projects using Vox Populi

As well as for IWA, Vox Populi has been used in two other projects, namely VJ Cultuur and Passepartout. Although the goal and the domain of both projects are different from IWA's, we have used this experience to extend and improve Vox Populi's implementation.

6.5.1 Visual Jockey

The project **VJ Cultuur**[9] aims at describing the work of VJs with respect to other art disciplines and the reciprocal influences between existing visual arts and VJ (see figure 6.9). VJ Cultuur is the result of a one-year long research about the culture of VJs in the Netherlands and consists of video interviews with 12 Dutch VJs. From the start of this project it was decided to present the video material in an interactive way using Vox Populi. Therefore, VJs were asked the same questions with the intent of presenting the material using the *vox populi* technique (section 2.2.6). Due to the short duration of the project, the annotations could only cover the questions asked and the interviewees (and not the statements made during the interviews), therefore the rhetorical editing which makes Vox Populi different from other tools could not be used. Nevertheless, VJ Cultuur was annotated using the same tools as for IWA and Vox Populi was used to generate categorical type of documentaries, based on the viewer selection of question asked and interviewee (fig. 6.10). The annotation schema (specified in chapter 4) was expanded to cater for the needs of this project, for example by including a field for the answer transcript or a field for ordering the questions.

[9] `http://www.vjcultuur.nl/`, in Dutch.

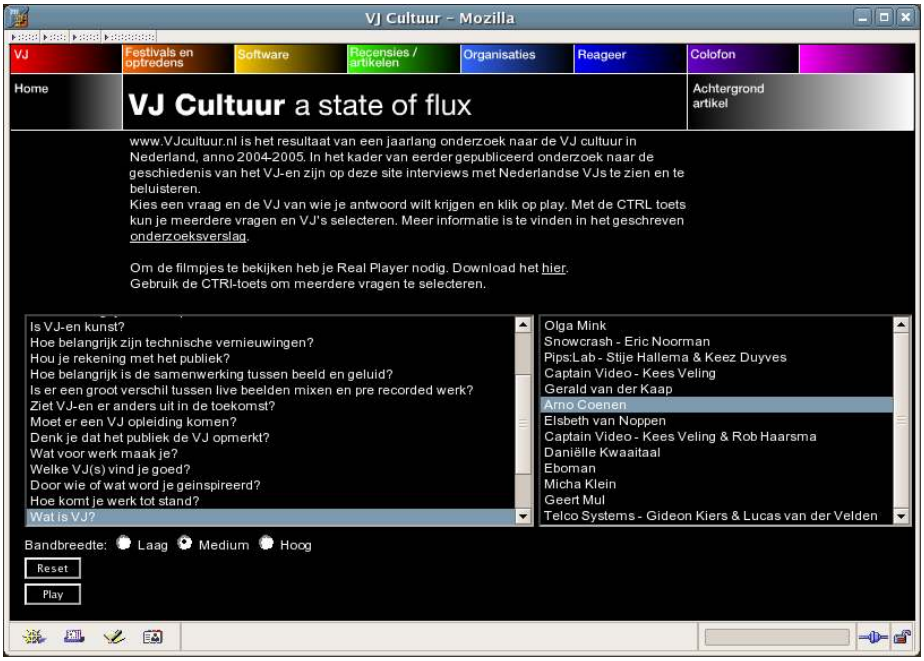

Figure 6.9: VJ Cultuur using Vox Populi

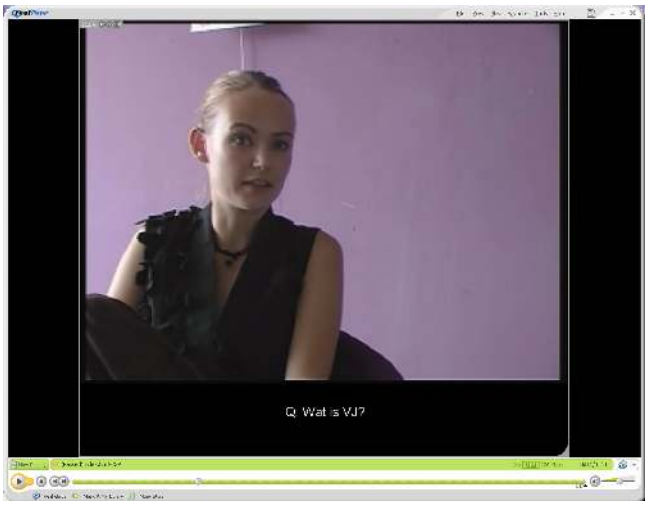

Figure 6.10: Example of a generated sequence, with question “What is VJ?” (“Wat is VJ?” in Dutch)

The annotator found Vox Populi "easy to handle and to learn, easy to trace back your steps. It probably has many more functions than I needed, so for me it could have been more simple". Even though in VJ Cultuur's case a less sophisticated tool would have performed as well, viewers were reported to be very enthusiastic about the ease of use of the system.

6.5.2 Passepartout - move.me

The move.me project is part of the European ITEA **Passepartout** project[10], which investigates the technical challenges and new forms of human-computer interactions in broadband home environments. The move.me prototype is an interactive cafe/lounge environment in which several users can manipulate the audio-visual content presented on various screens through interaction with sensor-enhanced objects, in this case cushions. The prototype achieves a tight coupling of the human body to an ambient "intelligent" system, such that the system as a whole becomes both an adaptive provider of content and a transparent "window" for social interaction among participants.

In the first phase of the project Vox Populi was used as an example of how the sensor data of the touchpad within a cushion can be mapped to higher level commands that alter the presentation of an interactive program. Depending on the current excitement state of the user, the move.me prototype decided if an adaptation of the content to be presented was required. For example, in cases where a user was overexcited, the move.me adaptation engine required that less controversial content was presented. In addition it would propose to generate a documentary that supports the beliefs of the viewer (also taken from the move.me user model). The match between these beliefs and the query structures provided by the Vox Populi interface, as described in fig. 6.2, were performed by the move.me engine, which would generate a query and transmit it to the Vox Populi engine. The material used at this stage of the move.me prototype was taken from the IWA repository.

The move.me environment showed the applicability of our approach for the automatic generation of documentaries within an interactive environment, as the behavior of the adaptation engine, completely based on biometric feedback from the user, cannot be predicted. The same is true for the perception of the user which results of new biometric feedback. The generation capabilities of Vox Populi could be used to react adequately to changes in the environment.

6.6 Preliminary technical evaluation

We have not conducted formal user studies yet since Vox Populi's main purpose is to demonstrate the feasibility of generating rhetorical video sequences. In addition, the interface would need to be improved to make it appealing to real viewers. For example, more captions should be provided to clue viewers about the context of each clip shown. Moreover, the quantity of annotated material only allows to generate short sequences. Nevertheless, in the following we provide a brief **technical evaluation** of the graph creation process described in section 5.2.1.

Cleary and Bareiss [19] evaluate their automatic linking methods comparing the links created automatically with the links created manually by professional annotators. Performances of each method are calculated using precision and recall against

[10] `http://www.citi.tudor.lu/passepartout`

	Subject	Modifier	Predicate	Vox Populi	Manual
1	*bombing*	*no mod*	*effective*	SUPPORTS	SUPPORTS
2	*bombing*	*not*	*effective*	CONTRADICTS	CONTRADICTS
3	*bombing*	*no mod*	*useless*	CONTRADICTS	CONTRADICTS
4	*bombing*	*no mod*	*waste*	CONTRADICTS	CONTRADICTS
5	*diplomacy*	*not*	*effective*	SUPPORTS	SUPPORTS
6	*economic-aid*	*no mod*	*effective*	CONTRADICTS	CONTRADICTS
7	*ground forces*	*not*	*effective*	CONTRADICTS	CONTRADICTS
8	*military-actions*	*no mod*	*effective*	SUPPORTS	SUPPORTS
9	*military-actions*	*never*	*effective*	CONTRADICTS	CONTRADICTS
10	*violence*	*must not*	*(be) used*	**not found**	CONTRADICTS
11	*violence*	*must*	*(be) used*	**not found**	SUPPORTS
12	*violence*	*not*	*necessary*	**not found**	CONTRADICTS
13	*war*	*not*	*effective*	CONTRADICTS	CONTRADICTS
14	*war*	*not*	*effective*	CONTRADICTS	CONTRADICTS
16	*war*	*not only*	*effective*	CONTRADICTS	CONTRADICTS
17	*war*	*only*	*effective*	SUPPORTS	SUPPORTS

Table 6.1: Links generated from the statement *war effective* by Vox Populi and by a manual annotator

the *golden standard* of human annotators. Such an evaluation is not possible in our case, since we do not have manually created links for our annotations. Nevertheless, to get at least an indication of performance, we tried the same experiment by manually creating links for a particular statement and comparing them with the ones generated automatically. To do so, we chose a statement which had the most automatic links in the repository, i.e. *war effective*, and selected in the repository the statements which we judged should be linked to it, specifying also the type of link. In selecting the statement with the most links we did not want to make performances look better than they are, but we were interested in seeing if all the generated links were correct, as well as how many Vox Populi was able to find. The results are shown in table 6.1. As can be seen, the links that Vox Populi can find are correctly assessed as SUPPORTS or CONTRADICTS. Vox Populi cannot find 3 of the manually created links. For the statements in row 10, 11 and 12, even though *violence* is related to *war* in the thesaurus with relation *Similar*, Vox Populi cannot establish that, for example, *war effective* and *violence must not (be) used* contradicts each other. This is because the contradiction is more in the meaning than between two terms, since *(be) used* and *effective* cannot be related by the relations we use in the thesaurus. Vox Populi cannot create links when the logic is more complex than the mechanism we defined in section 5.2.1. This limitation is due to the MEDIA-DRIVEN [HLR 6] requirement, which makes modeling domain knowledge not suited for our approach (as discussed in section 3.2.1.3). If a calculation of precision and recall based on a single example would make sense, Vox Populi would have a precision of 100% (too good to be true) and a recall of 82% (a fair result).

6.7 Conclusions for future video generation approaches

In this section we report some conclusions derived from our experience in using Vox Populi. For the IWA project we played the role of the documentarist, first shooting the material and then annotating it. From our experience, the annotation effort can be divided into two types: deciding the best way to cut the material into clips and deciding how to describe each clip. The former turned out to be very time-consuming, since clips must be cut so that when composed dynamically in a sequence, the sound is understandable and sentences do not start or stop abruptly. Clips must be very often "cleaned" from interviewer's voice, interviewee's pauses or noisy expressions such as laughter. The documentarist is forced to search for the point where to cut and to listen to the clip to verify the cut is correct. The same experience was reported also by the annotator of the VJ project. Even though the annotations in this case were much simpler, the need to cut the clips made annotating the material very time-consuming. As we mentioned in section 6.4.1, the annotation process could be greatly facilitated by automatic audio-based segmentation methods, which could provide candidate points where to make a cut. Any approach which requires video segmentation should consider investigating which tools are available to make this task at least semi-automatic.

A better SMIL player would have also been beneficial to our approach. Even though SMIL allows to dynamically edit video sequences, it only offers some of the features of a professional video editing software. This limitation is made worse by the fact that SMIL is not well-supported by players, of which RealPlayer was the only sufficiently working option. As long as SMIL is not well-supported, video generation approaches that focus particularly on editing effects should consider a different output format. For example, Final Cut Pro[11], a professional video editing tool, can export and import XML editing directives. Instead of generating video in SMIL, XML could be generated and imported in Final Cut Pro to execute the rendering. The downside of this solution is that it is difficult to automate, since a different software must be invoked to render the video before the viewer can see the result of her request.

Protégé, the annotation editor and Sesame, the annotation storage, served our purposes very well, with the additional advantage of being open-source and free.

6.8 Summary

In this chapter we presented Vox Populi, an implementation of the generation model we defined in chapters 4 and 5. We described the system architecture (section 6.2), as well as how viewers can request a documentary and documentarists can author a documentary repository. For the former, we described the interface and the generated output (section 6.3), while for the latter we discussed the annotation tools and Vox Populi's rule mechanism for editing (section 6.4). We then presented some experiences from the use of our system in two other projects, VJ Cultuur and Passepartout (section 6.5), and we provided a brief technical evaluation (section 6.6). Finally, we reported some conclusions for automatic video generation approaches (section 6.7).

[11] http://www.apple.com/finalcutstudio/finalcutpro/

Chapter 7

Conclusions

In this chapter we present an overview of the thesis and we discuss how we answered the research questions we formulated. We then present future work to expand the borders of what we have achieved.

7.1 Introduction

This chapter starts with a short summary of the thesis research contributions (section 7.2). The summary prepares the reader for the discussion that follows. In section 7.3 we discuss the applicability of our approach in different scenarios. This applicability is hindered by the effort required to annotate the material. We therefore discuss possible solutions to this problem. We then examine two issues related to automatic generation approaches, namely the influence of the annotations on the viewer choices and the limitations of operating under an open-world assumption.

In section 7.4 we discuss possible alternative solutions for the presentation forms we adopted in this thesis as an answer to *Research Question Documentary Form [1]*. We examine how the rhetorical form could be used both at the macro and at the micro level, and how some narrative properties can be used to compose a documentary without violating the MEDIA-DRIVEN [HLR 6] requirement.

In section 7.5 we examine issues related to the modeling and the creation of arguments. Improving the process of argument creation might require a revision of our annotation schema and generation process, which we defined as answers to *Research Question Annotation Schema [2]* and *Research Question Generation Process [3]*, respectively. The model for logos, pathos and ethos is discussed, taking into account the differences between verbal and non-verbal information and the role that non-verbal information in visuals could play. We then present how the model of Toulmin provides more information than we currently use, and how an alternative discourse theory, RST, can provide an alternative way of encoding arguments. We then show how capturing more semantics can provide a finer granularity for the argumentation relations in the *Semantic Graph*.

In section 7.6 we discuss the feedback mechanism we defined as an answer to *Research Question Author Support [4]*, and suggest how its principles could be applied to other generation systems.

Having discussed the last research question, in section 7.7 we present future directions we see for our work.

7.2 Contributions of the thesis

In the thesis we present a model to automatically generate documentaries, and an implementation of it. We focus on matter-of-opinion documentaries based on interviews. Our model has the following characteristics, which are lacking in previous automatic generation approaches:

- it allows the viewer to select the subject and the point of view of the documentary;
- it allows the documentarist to add material to the repository without having to specify how this material should be presented (data-driven approach);
- it generates documentaries according to presentation forms used by documentarists.

To build the model, we first studied the domain of video documentaries. As a result of this analysis, we formulated *high-level requirements* that specify what aspects of documentary making need to be captured in an automatic video generation model. We then reviewed related work examining which existing approaches can satisfy, at least partially, the high-level requirements we set, and specified with *low-level requirements* how feasible solutions from literature should be included in our model. The low-level requirements divide the model into two components: an *annotation schema* capturing information and a *generation process* manipulation this information to create documentaries. We defined both components so that they satisfy the high-level and low-level requirements and we implemented the model in a demonstrator called *Vox Populi*. Vox Populi was also used in two other projects.

7.3 Automatic video generation: common issues

In the thesis we present an alternative way of authoring documentaries, which a documentarist might decide to use instead of the traditional documentary making process. The use of our approach is not necessarily an alternative to traditional documentary making: our system could also be used to suggest interesting editing possibilities, based on different points of view. The documentarist might adopt, or expand on, generated sequences to create a final static version, or she could use the system as a way of browsing the content of the footage, instead of using transcripts and logs.

Our approach could also be used in other cases. For example, there are many content providers and publishers that offer online access to their multimedia archives. The **BBC Motion Gallery**[1] contains thousands of shots from the vast and diverse archives of the BBC and CBS News. The archive **het Geheugen van Nederland**[2], ("the memory of the Netherlands") consists of a large digital collection of photos, texts, films and audio fragments from several Dutch cultural institutes. These archives usually offer keyword based access. The viewer can query for concepts in the content and for properties of the footage, e.g. when it was shot and where. Keyword search allows the retrieval of a list of relevant (parts of) documents, but it is unable to build a single new document using the parts interesting to the viewer. The viewer has to go through

[1] `http://www.bbcmotiongallery.com`
[2] `http://www.geheugenvannederland.nl/`

this list viewing one or more items, selecting the interesting parts and mentally placing them in a coherent document. Our approach would allow the presentation of the content in a documentary-like form.

7.3.1 The annotation effort

An obstacle to the adoption of our approach is the effort needed to annotate the material. We designed the annotation schema to be as simple as possible, by focusing on the process of creating arguments and encoding as little of the content as needed for a rhetorical approach. Still, the annotations we require are considerably more complex than keywords and have to be done manually since they cannot (yet easily) be obtained automatically from video and audio processing methods. This is because we need symbolic information such as positions, while the finest granularity that state-of-the-art automatic approaches in multimodal video indexing are able to detect consists of, for example, explosions in action movies, goals in soccer games, or a visualization of stock quotes in a financial news broadcast ([67]). Although this information is sufficient for some applications (e.g. automatic generation of trailers for action movies [27]), it can not provide the level of semantic we need in our approach. This problem is known as the semantic gap [65].

In our case, as well as in the case of all the systems we described in chapter 3, annotations are therefore manual. Annotating archives such as the BBC motion gallery would be extremely time and resource consuming, and the annotation effort is determinant for the adoption of the approach (as argued in [73]). Obtaining annotations automatically would greatly facilitate the adoption of approaches that use them.

On the other hand, traditional documentary making already includes some form of annotations, such as logs and transcripts (section 2.3.5). Documentarists are therefore used to some annotation effort. The information contained in logs and transcripts could be used for the annotations we need for our approach, e.g. the question asked, the data about the interviewee, the position expressed. This information could also be entered at shooting time, by gathering annotations during media capturing, either automatically (e.g. timestamps) or manually from the documentarist. This has the advantage that information is captured at the moment it is available, and kept for future reference (such an approach is used in Nack's smart camera work [49]). In our street interview scenario, this would mean, for example, annotating the question and the interviewee during the interview. In this way, traditional documentary making could be modified to facilitate the adoption of our approach.

Another source of metadata for video could be provided by recent research efforts to facilitate non expert users with video content creation. For example, the **MINDFUL DOCUMENTARY** [5] is a system aiding filmmakers during the shooting phase. The system consists of a device containing a camera and a processing unit with commonsense reasoning capabilities. The filmmaker shoots a scene and informs the system of what she has shot by entering a description in the device. Using commonsense reasoning, the device provides a list of suggestions for next shots that the filmmaker can follow, save for later or disregard. Adams's **MEDIA CREATION ENVIRONMENT** [1] is a system targeted at helping non-expert users to make home videos. This environment is supposed to aid the user in all tasks involved with producing a film, i.e. authoring, screen playing, shooting, and in particular low-level editing, using well-formed media transformations in terms of high-level film constructs (e.g. tempo). These systems can provide metadata while shooting by simply asking the user to record whether she has taken the suggested shot or not. If she has, the suggestion then can be used as metadata

associated with the suggested shot.

Another possibility to reduce the annotation effort is provided by collaborative annotation initiatives. For example, **filmEd** [61] combines videoconferencing with collaborative indexing, browsing, annotation and discussion of video content between multiple groups at remote locations. Another example is the **Echo Chamber** project[3] (section 1.1), an open source, investigative documentary where users can transcribe and rate footage to form a sort of collaborative filmmaking.

7.3.2 The documentarist's influence

In our approach the viewer can choose the subject and the point of view of the documentary. Considering that the documentary generation process relies on the annotations of the material, this choice is influenced by the person who provides these annotations, i.e. the annotator. Typically, but not necessarily, the annotator is the same person as the documentarist. The annotator has the power of removing footage by not annotating it, making it effectively non-existent for the generation process. Furthermore, the annotator segments and interprets the interviewees' answers in arguments and statements, building the thesaurus at the same time. All these operations are influenced by the point of view of the annotator, analogously to how the point of view of a documentarist influences traditional film making. The viewer therefore does not completely determine the subject and point of view of the generated documentary. We can conclude that in traditional film making the final result is determined only by the documentarist, while in our approach the final result is determined by both the annotator/documentarist and the viewer. The influence of a single annotator is further minimized in the case of collaborative annotations. Ideally, when multiple sources of annotations from different annotators are present, the viewer is able to choose not only the subject and the point of view of the documentary, but also which annotations/annotator to use for the generation.

7.3.3 Automatic video generation under the open-world assumption

Our approach allows the documentarist to add annotated material to the repository and have the generation process automatically use it to generate documentaries. This means that the content of the repository cannot be foreseen beforehand, or, in other words, that the model needs to operate under an open-world assumption rather than a closed-world one. Since our annotations do not capture all the information contained in video clips, some generated combinations of clips may make no sense or be either unqualified or offensive, which clearly produces the opposite effect to the intended one. On the contrary, in systems operating under a closed-world assumption, such as Terminal Time [44] (section 3.2.1.3), the material is created specifically for the automatic generation task. This allows complete control over the content and the possible sequences that can be generated. Analogously to what Nack [48] concludes, some unreliability is the price one has to pay for operating in an open-world setting.

7.4 Alternative presentation forms for documentaries

We discuss here issues related to:

[3] `http://www.echochamberproject.com/`

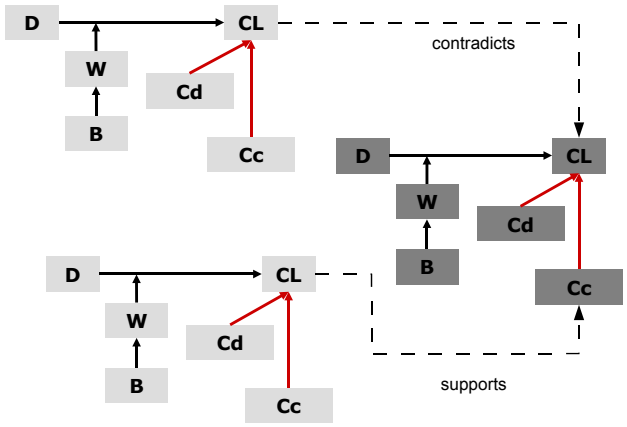

Figure 7.1: Example of narrative progression based on Toulmin

RESEARCH QUESTION 1 *(*DOCUMENTARY FORM*) What characteristics of the presentation forms used by documentaries on matter-of-opinion issues must be modeled?*

In modeling presentation forms used in documentaries, we use the rhetorical form at the micro-level and the categorical form at the macro-level (chapters 2 and 5). This does not need to be so: these two forms could change role with respect to the level and/or be used at both levels. Furthermore, even though the narrative form is used in documentaries, we could not include it in our approach because of our requirements. In the following we discuss two possible alternatives to the presentation forms we adopted that could still satisfy the requirements we set: using only the rhetorical form and using narrative-like forms. These alternatives could be based on the information we already captured in our annotations and they would not require modifying the annotation schema we defined.

7.4.1 Rhetorical form

At the macro-level, we currently use categories to organize content into structures larger than a single argument. Another possible choice is to use only the rhetorical form, since the documentaries we aim to generate are rhetorical in nature. Categories would then need to be replaced by rhetorical means to structure the content. These means should be compatible with the MEDIA-DRIVEN [HLR 6] requirement, i.e. they should not rely on particular knowledge about the domain, except what is encoded in the thesaurus as relations between terms. A possible structure based only on the rhetorical form can be obtained by chaining together arguments using the information encoded with the model of Toulmin (see fig. 7.1). When composing an argument, instead of using only single clips supporting or contradicting parts of the target argument (the darker Toulmin schema in the figure), if these clips correspond to the claims of other arguments, the generation process could include those arguments with the sequence. In the figure, the two claims on the left (belonging to the lighter Toulmin schemas) are included in the composed argument together with the arguments of which they are the

claim. In this way the generated sequence represents a chain of arguments leading to the debate of a single claim.

7.4.2 Narrative form

The use of the narrative form, i.e. the generation of a documentary based on a story, is against the MEDIA-DRIVEN [HLR 6] requirement (section 3.4). Nevertheless, if we do not require the narrative to have a story logic, and we aim at representing other narrative characteristics, alternative approaches might be possible. For example, some studies have described the role of **dramatic intensity** over the course of a narrative. Freytag [28] describes the increase and decrease of dramatic intensity in a story as a triangle, starting with low intensity in the *Exposition* (where the characters and the setting are introduced), raising through the *Inciting Incident* (which introduces the major conflict of the play) to a climax in the *Peripeteia* (reversal or change of fortune: a character produces an effect which is the opposite of his intentions), and then decaying to the *Denouement* (the unraveling, or untying, of the complexities of the plot).

Therefore, a narrative evolution in the generated documentary can be obtained by using different intensities at different moments of the video. A possible implementation could be to relate intensity to the rhythm of the presentation. Since rhythm is related to the length of the clips, the generated video could be made in such a way that clip lengths would follows an intensity curve (Freytag's triangle for example). Alternatively, intensity could be mapped to the degree of contrast between arguments in the video. In this case, the documentary might start with people supporting each other, then transition to a clashing phase and then resolve again with arguments supporting each other.

Narrative evolution can also be achieved by using well-known metaphors, such as the day or the journey (or both, the one-day journey), which provide a progression in time or space. For example, the documentary could start with clips shot in the morning and/or at a particular location, and progress in time till the evening and/or another location.

7.5 Modeling and creating arguments

We discuss here issues related to:

RESEARCH QUESTION 2 *(*ANNOTATION SCHEMA*) What information should be captured in an annotation schema for an automatic video generation approach where:*

- *the viewer can specify the subject and the point of view,*
- *the documentarist can collect material to be used for documentaries, without having to specify how this material should be presented to the viewer,*
- *the material is presented according to presentation forms used by documentarists?*

and

RESEARCH QUESTION 3 *(*GENERATION PROCESS*) How must a generation process be defined for an automatic video generation approach where:*

- *the viewer can specify the subject and the point of view,*
- *the documentarist can collect material to be used for documentaries, without having to specify how this material should be presented to the viewer,*
- *the material is presented according to presentation forms used by documentarists?*

Information modeling in video and video generation are closely related, since the former is designed to support the latter, and the latter determines what information needs to be modeled by the former. Therefore, we address in this section issues related to research questions 2 and 3 together.

Arguments are central to our approach. We discuss here alternatives to the argument model and argument creation process we present in this thesis, beginning with logos, pathos and ethos. Because we focus on verbal information, logos is central to our approach, while pathos and ethos play a comparatively minor role. We discuss the issues of making pathos and ethos more prominent, using alternative models, and how visuals can contribute to this. We then examine features of the model of Toulmin we are not yet using, to see whether they can offer more argument creation possibilities. The model of Toulmin defines a role for the part of an argument. An alternative is to examine the relation between the different parts of an argument, as it is done in RST [43], a theory of discourse. We examine whether RST's approach can be useful to ours. Finally, we consider whether the argument creation process would be facilitated by having a *Semantic Graph* with more types of argumentation links than just SUPPORTS and CONTRADICTS, and we present some solutions to achieve this, namely the use of all the relations in our thesaurus and the use of ontologies.

7.5.1 Logos, pathos and ethos

In this thesis we make the simplifying assumption that logos, pathos and ethos can be modeled separately and do not interact with each other. A different approach might investigate whether there are interrelations that make logos, pathos and ethos dependent on each other. The existence of dependencies would require a redesign of the generation process, which uses logos, pathos and ethos as three separate techniques.

In our approach we use logos to model verbal arguments, while pathos and ethos are used to assess how convincing an interviewee appears to the viewer. Due to the lack of a theory capable of establishing how the characteristics of video influence a viewer inclination towards the positions expressed in the video (as discussed in section 3.2.2), we use a simple model for pathos and ethos. Nevertheless, our generation process could use, were they available, ethos and pathos models that provide a more accurate rating, as long as the price in terms of annotations would not be too excessive (as discussed in section 7.3.1). In this sense, more accurate models would improve the quality of the final result, but as long as they provide a numerical rating for a clip, the generation process would not need to be modified. These models can in fact be considered input data to the generation process, which is mostly logos based.

Logos is easier to model because it is based on verbal information, which has been studied extensively. Using existing theories such as the model of Toulmin, we can decompose interviewees' answers into arguments parts and statements, which can be recombined and reused in other contexts. For ethos and pathos, an argument cannot be decomposed into or composed from smaller units, i.e. there is no equivalent to statements and argument structures. If we would have similar pathos/ethos entities to

argument structures and statements, ethos and pathos could be used to build arguments in the same way as we do with logos.

7.5.2 Statements and non-verbal information

In interview documentaries, information is mostly conveyed by the interviewees' replies to questions. Consequently, in our approach we focus on verbal information more than on non-verbal information and, in particular, on speech more than other media types. Verbal information has been the subject of extensive research, which makes modeling easier than for non-verbal information. Arguments are more easily expressed using verbal information, which makes composing arguments based on verbal information easier. Nonetheless, according to Hampe, a documentary also needs visual evidence: ([33] pp. 53-54):

> "*If a film—or a video—isn't composed primarily of visual evidence, then even though you recorded it with a camera and show it on a screen, it really isn't a film. [...] Getting people to talk about the subject of the documentary is important, but mainly for research purposes. Yes, you always hope for a great sound bite that will drive home a point. But if all you have is people* ***telling*** *about the topic, you lack the visual evidence to make a documentary film.* "

Moreover, film editing is able to combine the expressive power of video and audio, for example by using particular transitions between shots, or by varying the pace of the presentation. How to model non-verbal information contained in visuals and its interaction with verbal information is for us an important future work direction.

Our approach could be extended to model non-verbal information if we observe that statements are not limited to encoding verbal information. Statements can also encode visual and non-visual information which is non-verbal in nature, but can be associated with a verbal message. We concentrate on visual information. For example, images like "do not smoke" carry a verbal message because they use a language, the sign language. Other types of visual information can make a statement as well. For example, a video sequence of a river being polluted by a factory can express the statement "*Factories pollute the environment*", although the association is not as strong as for verbal information. Such a sequence can be used to build an argument against a verbal statement such as "*Industrialization is positive for the world*".

Associating statements to non-verbal information can be more difficult than formally summarizing what has been said, since it requires an interpretation of the impressions visuals can evoke in viewers. Nevertheless, some visuals have a strong connotative meaning that could be annotated with statements, for example flags, elderly people and children can be associated with something we have to protect. We call these examples **visual statements**.

We could encode visual statements in the same way as verbal statements, with the three-part structure and with terms from the thesaurus. Once encoded, visual statements could be manipulated together with verbal ones, and linked to other statements with either CONTRADICTS or SUPPORTS links. Ethos and pathos would need to be extended beyond interview video sequences, since they are currently defined only for sequences that show humans. Having done this, visual statements could be used to compose arguments in the same way as with verbal statements (section 5.2.2).

Visual statement would also allow another possibility for editing: **counterpoint editing**. This technique consists of presenting the sound from one clip, such as an

interview, together with the images from another. Rabiger says about counterpoint editing ([54] p. 272):

> "*By creating juxtaposition that requires interpretation, film is able to counterpoint antithetical ideas and moods with great economy. At the same time it can kindle the audience's involvement with the dialectical nature of life.* "

In counterpoint editing, a video clip described by a visual statement is overimposed to a rhetorically related video clip. An example of counterpoint editing is a sequence where the video track shows a river polluted by a factory (encoded by the visual statement "*Industrialization bad*") while in the audio track a voice is explaining the advantages of industrialization (encoded by the verbal statement "*Industrialization good*"). In [9] we present such an approach using images, which are superimposed to interviews fragments.

7.5.3 The model of Toulmin

In our approach, we use the Toulmin model to describe the role of each part of a discourse in building an argument. We do not, however, use all the granularity provided by Toulmin. Our support/counterargue strategies use Toulmin basically as it would be composed by only three parts: the claim, the premises that support the claim (i.e. data, warrant and backing) and the premises that refute the claim (i.e. concession and condition) (section 5.2.2.1). Future work could explore the use of differences in counterarguing/supporting the specific Toulmin part, and how.

Another feature of Toulmin we included in our annotation schema, but we do not use yet, is that parts of the argument can be implicit (as explained in section 3.2.1.1). For example, the argument "we have been attacked, therefore war is the right answer" might be considered to have an implicit warrant:"*if you are attacked, you must react with violence*". By encoding this implicit information, the generation process could counterargue the above-mentioned argument with a statement such as "*violence is not the answer to violence*". Therefore, the model of Toulmin might offer other ways of counterarguing or supporting a claim by describing implicit assumptions.

We also considered alternatives to the model of Toulmin, by looking at Rhetorical Structure Theory [43], a theory which is also able to model the different parts in an argument. The main difference between RST and Toulmin is that Toulmin describes the role of each part of a discourse in building an argument, while RST describes the particular relation between sentences in a coherent discourse. Some of the relations RST defines can be applied to describe the relation between a claim and its premises. For example:

- Otherwise : "we fight back, otherwise they will do it again".
- Concession: "even if we fight back, they will do it again".

The two arguments mentioned in the example contradict each other, and this is because of the different relations between the statement "*we fight back*" and the statement "*they will do it again*". Since RST captures this information, particular strategies could be defined to support or counterargue an argument depending on the relation between the claim and its premises. In our case, because Toulmin has less granularity than RST, we would need to account for the contradiction by encoding the statements, for

example the first case as "*fighting is self-defense*" and the second as "*fighting is not self-defense*". Using different discourse theories is another possible approach which could be considered to improve our capability of composing arguments.

7.5.4 Argumentation links and ontologies

Providing presentation forms different from the ones we describe in the thesis could expand the applicability of our approach. The support/counterargue rhetorical forms we currently use are based on the premise that clips either *support* or *contradict* each other. Having finer granularity relations could allow the definition of more complex presentation forms. In more detail, the *Semantic Graph* we use in our approach contains two types of argumentation links, namely SUPPORTS and CONTRADICTS. These two types are approximations of the possible relations between two statements. If we compare this level of granularity with approaches such as ScholOnto [63] (discussed in section 3.2.1.5), we see that the generic positive/negative meaning expressed by supports/contradicts links is further specified in, for example, *confirms*, *is (in)consistent with*, *takes issue with*, *raises problem*, *refutes*. This added semantics leads to a *Semantic Graph* containing more detailed typed links, possibly allowing more complex presentation models than the ones we defined in chapter 5. As we discussed when formulating the low-level requirements (chapter 3), ScholOnto's approach, and any other approach involving manually annotating links between items, are not feasible in our case. The only viable alternative is to define a process that automatically creates subtler links than supports/contradicts ones. This process cannot depend on the specific concepts of a particular domain, since this would limit the scope of application of our model. Capturing domain knowledge, as in Auteur [48] (discussed in section 3.3.2.2), is not feasible for our approach because the repository can grow to include contributions which are not known a priori.

A possibility for increasing the granularity of the links in the semantic graph could be to better exploit the semantics contained in the thesaurus. At the moment we do not consider the specific semantics of the relations *Specialization* and *Generalization*, since we treat these relations as providing generic associations between terms (section 5.2.1.2). Considering a more specific role might provide more typed links for the *Semantic Graph*. Another possibility could be to use, instead of a thesaurus, an ontology. An ontology is able to capture more semantics than a thesaurus, allowing to define more complex rules to create links. On the other hand, an ontology imposes more constraints on the annotator's freedom to choose the terms she thinks best describe the clip, and forces her to check that she is using the semantics of each term according to the semantics defined in the ontology. A thesaurus can be easily expanded or modified by the annotator, by adding or removing terms and adding or removing relations. Changing an ontology would not be as easy because terms acquire a semantic role depending also in their position in the hierarchy. Moreover, having rules defined on an ontology might require that the ontology does not change, analogously to Cleary and Bareiss's approach ([19] p. 39, discussed in section 3.2.1.6). The drawback of such an approach with respect to ours is that the annotation effort increases. Having more semantics would then come at the expense of making people less keen on applying this method, since the annotation effort is fundamental for adoption, as we discussed in section 7.3.1. Nevertheless, this is a possible alternative that we might experiment with in the future.

7.6 Providing feedback to the author

We discuss here issues related to:

RESEARCH QUESTION 4 *(*AUTHOR SUPPORT*) How must a generation process be defined so that it can give to the documentarist an indication of the quality of the documentaries it can generate?*

The method we presented to give feedback to the documentarist (section 5.3) is focused on how well the single statements are connected to each other in the *Semantic Graph*, since the process of argument creation relies on the SUPPORTS/CONTRADICTS links in the graph. Other possible feedback methods could be to calculate every possible user request, and check whether the generation process is able to build a documentary. These documentaries could be evaluated based on richness of content (i.e. whether they contain sufficient clips) and on their cinematic properties (i.e. how well the editing rules we established in section 5.2.3 are satisfied).

The feedback method we provide is based on our particular graph creation process and therefore it cannot be applied to other systems. Nevertheless, the idea to use performance measurements not only to *test* the modeling (i.e. the annotations in our case), but also to give exact feedback to *improve* the modeling, is relevant to other approaches as well. Particularly, we think that similar feedback methods could be applied to initiatives based on collaborative annotation efforts. When the role of the annotator is shared among different people, inconsistent annotations are more likely to be produced, and the use of annotation support tools could be valuable.

7.7 Future directions

The discussion in the previous section pointed out some future directions for our research. We examine here the ones we think are the most interesting and promising.

First of all, the role of **non-verbal visuals** in communicating a message or strengthening a verbal message needs to be further researched, and more theories relating video to viewer reactions need to be developed. Documentarists stress the importance of providing visual evidence and a visual story which complement what is said in the documentary. In section 7.5.2 we discussed including non-verbal visual sequences in our generation mechanism, by associating a verbal statement to non-verbal information. Other approaches are needed to establish how and when to use visuals, and also to understand the effects of combining non-verbal visuals with auditory verbal messages, e.g. using counterpoint editing.

A better understanding of visuals might lead to a better modeling of pathos and ethos. We especially see potential improvement in modeling **pathos** more accurately, since pathos has a strong appeal for viewers. For example, a model for pathos might include the emotions of the interviewees' facial expressions, for example using Ekman and Friesen [26] Facial Action Coding System (FACS), a standard for facial expression annotations.

A fundamental issue concerns the annotation effort. In section 7.3.1 we mentioned some approaches potentially able to reduce this effort. In section 6.7 we discussed the advantages of using automatic video segmentation approaches. Further research is needed to establish which approach can help produce the complex metadata required by our approach.

Another interesting research direction is about how to organize the content in a way that is engaging for the viewer, without requiring extensive annotations of the video material. In the related work chapter we discuss how Terminal Time and Auteur can generate interesting narratives, the former in a closed-world setting, while the latter using extensive annotations. Our open-world automatic generation approach needs more **rhetorical and narrative forms** that require a reasonable annotation effort, are domain-independent and interesting for the viewer. This research direction requires investigation of narrative as well as rhetoric, the latter being more of an undiscovered area since narrative in video has been already studied.

At a lower level than narrative, the process of **argument creation** could also be improved. The model of Toulmin is capable of encoding more information than we currently use, such as implicit argument parts. Using this information could enable the building of more and/or more complex arguments. Alternatively, we could use a different discourse theory, such as RST [43], to capture the exact relation between statements, which Toulmin cannot encode. Finally, argumentation relations between clips could be further specialized by encoding more semantic in the annotations, for example by using an ontology instead of a thesaurus.

As a closing remark, we hope that this thesis will stimulate further research in automatic video generation. Many important related works are already a decade old, while the interest of the man-on-the street for producing and using video has steadily increased over recent years. The interest to document reality with video is growing at the same pace as the technology enabling video production. As we have experienced during the course of our work, automatic video generation is closely connected with development in society. Different projects, outside the research world, would have used automatic video presentation methods were they available, and were interested in using our approach. All this interest definitely deserves a place in the multimedia community's research agenda. As a side effect of this, we will gain a deeper understanding of interpreting media, and the ways media (and viewers) can be manipulated, of which the work presented in the thesis is an example.

Appendix A

Typographical legend

In this thesis we use typographical conventions to facilitate understanding what concept a term relates to.

Definitions

Research questions are defined by:

RESEARCH QUESTION 1 *(*ONE*) This is a research question*

and referred to as *Research Question Zero [1].*

High-level requirements are defined by:

HIGH-LEVEL REQUIREMENT 1 *(*ONE*) This is an high-level requirement*

and referred to as ONE [HLR 1].

Low-level requirements are defined by:

LOW-LEVEL ANNOTATION REQUIREMENT 1 *(*TWO*) This is an annotation low-level requirement*

for annotations and by:

LOW-LEVEL PROCESS REQUIREMENT 2 *(*THREE*) This is a process low-level requirement*

for the process. Low-level requirements are referred to as TWO [LLR 1].

Concepts

Keywords are indicated like **automatic video editing**. Systems or authors in related work are indicated as **SCHOLONTO**. Projects or initiatives are indicated as **Interview with America**.
Citation are indicated like this:

"*this is the text of a citation* "

We indicate an **interview** text as follows: "This is the text of an interview", while a particular statement from the interview is "*I cannot think of a better solution than war*". When such a **statement** is formally annotated, it becomes s:war m:most p:effective. The abbreviations s, m and p stand for subject, modifier and predicate, the three parts of a statement.

The terms used to formally annotate a statements are contained in a **thesaurus**, and are indicated like this: *war*, *most* and *effective*. Terms in the thesaurus are related by **relations**: *Similar* (sometimes abbreviated with *Sim.*), *Opposite* (sometimes abbreviated with *Opp.*), *Generalization* (sometimes abbreviated with *Gen.*), *Specialization* (sometimes abbreviated with *Spec.*) and *Id*.

Statements are annotated with their role in the **argument**, which can be one of the following: claim, data, warrant, backing, concession and condition.

Finally, we indicate **opinions** as "*War in Afghanistan - For*".

Special terms

These are the special terms we use in this thesis:

no mod indicates the absence of a modifier in a statement.
Semantic Graph indicates the graph representing the content and the relations in a repository.
SUPPORTS link joins two clips whose statements support each other.
CONTRADICTS link joins two clips whose statements contradict each other.

Appendix B

Research Questions and Requirements

In this appendix we report for ease of reference all the research questions and the requirements we have defined in the thesis.

Research Questions

RESEARCH QUESTION 1 *(*DOCUMENTARY FORM*) What characteristics of the presentation forms used by documentaries on matter-of-opinion issues must be modeled?*

RESEARCH QUESTION 2 *(*ANNOTATION SCHEMA*) What information should be captured in an annotation schema for an automatic video generation approach where:*

- *the viewer can specify the subject and the point of view,*
- *the documentarist can collect material to be used for documentaries, without having to specify how this material should be presented to the viewer,*
- *the material is presented according to presentation forms used by documentarists?*

RESEARCH QUESTION 3 *(*GENERATION PROCESS*) How must a generation process be defined for an automatic video generation approach where:*

- *the viewer can specify the subject and the point of view,*
- *the documentarist can collect material to be used for documentaries, without having to specify how this material should be presented to the viewer,*
- *the material is presented according to presentation forms used by documentarists?*

RESEARCH QUESTION 4 *(*AUTHOR SUPPORT*) How must a generation process be defined so that it can give to the documentarist an indication of the quality of the documentaries it can generate?*

High-level requirements

HIGH-LEVEL REQUIREMENT 1 *(*PRESENTATION FORM*) A generation model for matter-of-opinion interview documentaries must be able to organize information using the rhetorical form. The following aspects of the rhetorical form must be modeled:*

- *Presenting the material as a propagandist or a binary communicator.*
- *Presenting supporting/opposing positions in relation to each other.*
- *Composing arguments according to logos, pathos and ethos.*

Together with the rhetorical form at the micro-level, the model may use the narrative or the categorical form at the macro-level.

HIGH-LEVEL REQUIREMENT 2 *(*SUBJECT-POINT OF VIEW*) The model must allow the viewer to specify the subject and the point of view of the automatically generated documentary. Possible points of view are the propagandist's point of view and the binary communicator's point of view. In the case of the propagandist's point of view, the viewer must be able to choose the documentary position with respect to the chosen subject, either* against, for *or* neutral.

HIGH-LEVEL REQUIREMENT 3 *(*CONTEXT*) The model must be able to capture the context of statements so that an interviewee is not misquoted by taking her words out of context.*

HIGH-LEVEL REQUIREMENT 4 *(*MONTAGE TECHNIQUE*) It must be possible to present interviews with the vox populi montage technique.*

HIGH-LEVEL REQUIREMENT 5 *(*CONTINUITY RULES*) The model must assemble material at the rough cut using spatial, temporal, graphical and role continuity rules.*

HIGH-LEVEL REQUIREMENT 6 *(*MEDIA-DRIVEN*) The documentarist must be able to add material at any time to the repository used to generate documentaries without being required to specify explicitly how to present it to the viewer.*

Low-level requirements

LOW-LEVEL ANNOTATION REQUIREMENT 1 *(*LOGOS ARGUMENT*) The annotation schema must capture the form of arguments, which is independent from the subject matter, using the model of Toulmin.*

LOW-LEVEL ANNOTATION REQUIREMENT 2 *(*PATHOS ARGUMENT*) The pathos value of an interviewee can be determined on the basis of:*

- *whether the interviewee is appealing to the viewer.*
- *the framing distance of the video clip and the gaze direction of the interviewee*

LOW-LEVEL ANNOTATION REQUIREMENT 3 *(*ETHOS ARGUMENT*) The ethos value of an interviewee can be determined on the basis of the social categories the interviewee belongs to and a user profile determining how important for the user these categories are.*

LOW-LEVEL ANNOTATION REQUIREMENT 4 *(*ANNOTATION STRUCTURE*) Annotations must use a structure composed of* ***P-annotations*** *and a* ***controlled vocabulary****. In particular, media items must be described with a fixed sentence-like structure of the form <****subject****> <****modifier****> <****action****> <****object****>.*

LOW-LEVEL ANNOTATION REQUIREMENT 5 *(*ANNOTATION CONTENT*) Annotation must be clip-based and must model content on the denotative, connotative and cinematic level.*

LOW-LEVEL PROCESS REQUIREMENT 6 *(*GRAPH GENERATION*) The generation process must create, using inferencing from the annotations, a semantic graph where the nodes correspond to the media items and the edges are argumentation relations having one of two possible polarities, positive (i.e. supports) or negative (i.e. contradicts).*

LOW-LEVEL PROCESS REQUIREMENT 7 *(*ARGUMENT COMPOSITION*) The generation process can compose arguments together using actions such as* ***rebuttals****,* ***undercutters****,* ***sequential weakening*** *and* ***parallel weakening****.*

LOW-LEVEL PROCESS REQUIREMENT 8 *(*QUALITY FEEDBACK*) The generation process must provide feedback to the documentarist about the quality of the documentaries it can generate.*

LOW-LEVEL PROCESS REQUIREMENT 9 *(*CATEGORICAL FORM*) The generated documentary must use the categorical form to assemble content on the macro-level.*

Appendix C

Documentaries

Lowered costs allow more documentaries

> "*The diversification of television consumption, through cable, satellite, the Internet, and video facilities—with production equipment becoming smaller, better and cheaper—makes video presentations possible that were once prohibitively difficult and expensive. Distribution is evolving toward the diversity of book publication so a growing need for diverse screen authorship already exists. Lowered production costs and an increased outlet should mean increased freedom for the individual voice—the kind of freedom presently available in the print media ([54], page 10).* "

Categorical documentaries

Bordwell ([12] pp 133-134) describes *Olympia*, a categorical documentary:

> "*One classic documentary organized categorically is Leni Riefenstahl's Olympia, Part 2, made in 1936 as a record of the Berlin Olympics. Its basic category is the Olympic Games as an event, which Riefenstahl has to condense and arrange into two feature-length films. Within these films, the games are broken down into subcategories—sailing events, sprinting events, and so on.[. . .] Because categorical form tends to develop in fairly simple ways, it risks boring the spectator. If the progression from segment to segment depends too much on repetition ("And here's another example. . . "), our expectation will be easily satisfied. The challenge to the filmmaker using categorical form is to introduce variations and thereby to make us adjust our expectations.*
> *In categorical form, patterns of development will usually be simple. The film might move from small to large, local to national, personal to public, and so on.[. . .] Riefenstahl organizes Olympia according to a large-scale ABA pattern. The early part of the film concentrates on the games as such, rather than on the competition among athletes and countries. Later she shifts to setting up more dramatic tension by focusing on some of the individual athletes and whether they will be successful in their events. Finally, in the diving sequence at the end, there is again no differentiation among*

> *participants, and the sheer beauty of the event dominates. Thus Riefenstahl achieves her thematic goal of stressing the international cooperation inherent in the Olympics.* "

Ways documentarists use to make a categorical film more interesting are for ex. to choose a category which is exciting or broad or unusual. Another way is the use of film techniques.

> "*The diving sequence at the end of Olympia is famous for its dazzling succession of images of divers filmed from all angles.* "

Furthermore, categorical films can mix in other form of documentaries (Bordwell [12] p 134):

> "*Finally, the categorical film can maintain interest by mixing in other kinds of form. While overall the film is organized around its category, it can inject small-scale narratives. At one point Olympia singles out one athlete, Glen Morris, and follows him through the stages of this event, because he was an unknown athlete who unexpectedly won the decathlon.* "

Rhetorical documentaries

Talking about the way a film tries to convince the viewers, Bordwell says ([12], p 140):

> "*Films can use all sorts of arguments to persuade us to make such choices. Often, however, these arguments are not presented to us* ***as*** *arguments. The film will frequently present arguments as if they were simply observations or factual conclusions. Nor will the film tend to point out other opinions.* "

There are a number of rhetorical figures that support arguments (such as enthymeme or syllogism, see [16]) Bordwell ([12] p 141) says:

> "*Further, filmmakers can back up an argument by exploiting familiar, easily accepted, argumentative patters. Students of rhetoric call such pattern enthymemes*[1]*, arguments that rely on widespread opinion and usually conceal some crucial premises.* "

An example of this is to present a problem and then a solution, and show that the solution works well, implying that that is the best solution, even though other solutions (not shown in the film) might work better.

Documentary types

Documentary types according to Bordwell ([12] p 130):

- Interview documentaries (also called talking heads documentaries) records testimonies about events or social movements.

[1] An enthymeme is a syllogism in which one of the premises is implicit. An example of an enthymeme is: Some politicians are corrupt. Therefore, Senator Jones could be corrupt.

- Compilation film produced by assembling images from archival sources, for ex. newsreel footage and instructional films.
- Direct-cinema records an ongoing event as it happens, with no interference by the documentarist (see also Cinéma-vérité)[2].
- Cinéma-vérité records an ongoing event as it happens, with minimal interference by the documentarist (see also Direct-cinema).
- Nature documentary exploring the world of nature.
- Portrait documentary documenting scenes from the life of a compelling person.
- Synthetic documentary including more than one of the above mentioned types.

Logging

Hampe ([33] pp 283-284) in logging footage on computers:

> "*Quite often I have the footage log typed into a database program. This gives me the ability to sort the shots by scene and take number, by location, or by topic. Then, in addition to a sequential log of shots as they occur on the film or tape, I can create custom logs:*
>
> - *showing the scenes in the order in which they will be edited, and identifying the best takes*
> - *showing all the footage shot at a specific location, regardless of what tape it is on*
> - *showing all the shots dealing with a specific topic or person*
>
> "

On the importance of visual

Hampe ([33] pp 283-284) says

> "*If you consider your footage to be visual evidence and build your documentary as a visual argument, then you will rarely make the mistake of putting in a shot that doesn't belong simply because it's a* ***visual*** *that seems to* ***illustrate*** *something being said on the sound track.* "

Furthermore, Hampe ([33] pp 53-54) again says:

> "*If a film—or a video—isn't composed primarily of visual evidence, then even though you recorded it with a camera and show it on a screen, it really isn't a film. [. . .] Getting people to talk about the subject of the documentary is important, but mainly for research purposes. Yes, you always hope for a great sound bite that will drive home a point. But if all you have is people* ***telling*** *about the topic, you lack the visual evidence to make a documentary film.* "

[2] Actually Bordwell does not distinguish between Direct-cinema and Cinéma-vérité, but these are two distinct genres (see also Rabiger [54] p 25): for the first it is not admissible that the documentarist influences what is being filmed in any way, while in the second the documentarist can also appear on screen and cause or stimulate what is being filmed.

Appendix D

Technical details

Annotations contained in the IWA repository

The repository we use, "Interview with America", contains approximately 8 hours of video interviews. Of this, one and a half hours of video has been annotated, containing 15 people being interviewed and 37 questions asked. The annotations cover 185 video segments annotated by 120 statements. The thesaurus we use contains 155 terms and 199 relations between those terms (counting symmetric relations twice).

Interviewee
Woman with bike in Boston
Lawyer in Harvard
The other parking guard at Stamford
Corporate party worker
German in Boston
Internet Cafe barman
Girl in Starbucks
Old man working in the garden
Ethiopian in Boston
New York Empire State Building worker
Newspaper reader in Starbucks
Cameroon Parking Guard at Stamford
Black shop owner Stamford
Argentinean student in Boston
Second Argentinean student

Table D.1: Description of the interviewees contained in the IWA repository

Position
war in Afghanistan - Against
war in Afghanistan - Neutral
war in Afghanistan - For

Table D.2: Positions annotated in the IWA repository

Question
What changed after 09-11-2001?
Have you ever been threatened?
How will it be after the war?
What do you think about the media?
What do you think about the Anthrax?
Is this an attack against America?
What do you think of the Afghanistan war?
What do you think of the casualties among civilians?
Are you concerned about your own safety?
Is the economical situation a ground for terrorism?
Is there a continuous threat of terrorism?
What are the consequences of the war?
How do people feel about the war?
Is war the solution?
Do you think the war will last till the winter?
What will happen if the war extends?
Are people discussing the matters?
What is the position of the media?
What do you think about the situation?
Does the government hide the truth?
Is media presenting only the bad face of the public opinion?
What is the difference between terrorism and violence?
Will the military action give results?
What happens in your home country?
When is this going to end?
Do Americans treat you differently now?
Could 9-11 have been prevented?
What is the solution?
Has Europe a different point of view?
Why is the American opinion different from the European one?
What are the roots of the problem?
Why did they do what they did?
What is the ground for terrorism?
Will war make things worse?
What is the biggest problem?
Is the anthrax related to terrorism?
What do Americans think?

Table D.3: Questions annotated in the IWA repository

Bibliography and Filmography

[1] B. Adams, S. Venkatesh, and R. Jain. IMCE: Integrated media creation environment. In *2004 International Conference on Multimedia and Expo., Taipei, Taiwan.*, June 2004.

[2] Aidministrator Nederland B.V. *SeRQL user manual*, April 4, 2003.

[3] Aristotle. *Rhetoric*. Modern Library, NY, 1954. translated by W. Rhys Roberts.

[4] M. Bal. *Introduction to the theory of narrative*. University of Toronto Press, second edition, 1997.

[5] B. Barry. *Mindful Documentary*. PhD thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, 2005.

[6] H. Beck and H. S. Pinto. Overview of Approach, Methodologies, Standards, and Tools for Ontologies. The Agricultural Ontology Service (UN FAO).

[7] T. J. Bench-Capon. Specification and Implementation of Toulmin Dialogue Game. In *Proceedings of JURIX 98*, pages 5–20, 1998.

[8] S. C. Bernard. *Documentary Storytelling for Video and Filmmakers*. Focal Press, 2004.

[9] S. Bocconi. VOX POPULI: Automatic Generation of Biased Video Sequences. In *First ACM Workshop on Story Representation, Mechanism and Context*, pages 9–16, October 2004.

[10] S. Bocconi, F. Nack, and L. Hardman. Supporting the Generation of Argument Structure within Video Sequences. In *Proceedings of the sixteenth ACM Conference on Hypertext and Hypermedia 2005*, pages 75–84, September 2005.

[11] S. Bocconi, F. Nack, and L. Hardman. Using Rhetorical Annotations for Generating Video Documentaries. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME) 2005*, July 2005.

[12] D. Bordwell and K. Thompson. *Film Art: An Introduction*. McGraw-Hill, 7 edition, 2003.

[13] E. Branigan. *Narrative Comprehension and Film*. Routledge, New York, 1992.

[14] J. Broekstra, A. Kampman, and F. van Harmelen. Sesame: A Generic Architecture for Storing and Querying RDF and RDF Schema. In I. Horrocks and J. Hendler, editors, *The Semantic Web - ISWC 2002*, number 2342 in Lecture Notes in Computer Science, pages 54–68, Berlin Heidelberg, 2002. Springer.

[15] K. Brooks. *Metalinear Cinematic Narrative: Theory, Process, and Tool*. PhD thesis, MIT, 1999.

[16] G. Burton. Silva Rhetoricae. http://humanities.byu.edu/rhetoric/.

[17] C. I. Chesñevar, A. G. Maguitman, and R. P. Loui. Logical models of argument. *ACM Computing Surveys (CSUR)*, 32(4):337–383, 2000.

[18] J. Clark. XSL Transformations (XSLT) Version 1.0. W3C Recommendation, 16 November 1999.

[19] C. Cleary and R. Bareiss. Practical methods for automatically generating typed links. In *Proceedings of the the seventh ACM conference on Hypertext*, pages 31–41, Bethesda, Maryland, United States, March 1996.

[20] M. Crampes. Auto-Adaptive illustration through conceptual evocation. In *Proceedings of the 2st ACM international conference on Digital libraries*, pages 247–254, Philadelphia, Pennsylvania, United States, July 23-26, 1997. ACM.

[21] M. Crampes and S. Ranwez. Ontology-supported and ontology-driven conceptual navigation on the World Wide Web. In *Proceedings of the 11th ACM Conference on Hypertext and Hypermedia*, pages 191–199, San Antonio, Texas, USA, May 30 – June 3, 2000. ACM.

[22] M. Crampes, J. P. Veuillez, and S. Ranwez. Adaptive narrative abstraction. In *The Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia*, pages 97–105, Pittsburgh, PA, June 20-24, 1998. ACM, ACM Press. Edited by Kaj Grønbæck, Elli Mylonas and Frank M. Shipman III.

[23] M. Davis. *Media Streams: Representing Video for Retrieval and Repurposing*. PhD thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, 1995.

[24] M. Davis. Editing Out Video Editing. *IEEE MultiMedia*, 10(2):54–64, 2003.

[25] Dublin Core Community. Dublin Core Element Set, Version 1.1, 2003.

[26] P. Ekman and W. Friesen. *Unmasking the face. A guide to recognizing emotions fron facial clues*. Prentice-Hall Trade, 1975.

[27] FB 3 Digital Media, University of Bremen. Semantic Video Patterns, 2006.

[28] G. Freytag. *Technique of the Drama*. S.C. Griggs and Company, Chicago, sixth edition, 1895. Translated by Elias J. MacEwan.

[29] J. Geurts, S. Bocconi, J. van Ossenbruggen, and L. Hardman. Towards Ontology-driven Discourse: From Semantic Graphs to Multimedia Presentations. In D. Fensel, K. Sycara, and J. Mylopoulos, editors, *Second International Semantic Web Conference (ISWC2003)*, pages 597–612, Sanibel Island, Florida, USA, October 20-23, 2003. Springer-Verlag.

[30] J. Greimas. *Structural Semantics: An Attempt at a Method*. Lincoln: University of Nebraska Press, 1983.

[31] W. Grosso, H. Eriksson, R. Fergerson, J. Gennari, S. Tu, and M. Musen. Knowledge Modeling at the Millennium (The Design and Evolution of Protege-2000). Technical Report SMI Report Number: SMI-1999-0801, Stanford Medical Informatics (SMI), 1999.

[32] M. Halliday. Digital Cinema: An Environment for Multi-Threaded Stories. Master's thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, 1993.

[33] B. Hampe. *Making documentary films and reality videos: a practical guide to planning, filming, and editing documentaries of real events*. Henry Holt and Company, 1997.

[34] L. Hardman. *Modelling and Authoring Hypermedia Documents*. PhD thesis, University of Amsterdam, 1998. ISBN: 90-74795-93-5, also available at http://www.cwi.nl/∼lynda/thesis/.

[35] ISO/IEC. Overview of the MPEG-7 Standard (version 8). ISO/IEC JTC1/SC29/WG11/N4980, Klagenfurt, July 2002.

[36] C. Lagoze and H. V. de Sompel. The Open Archives Initiative: Building a low-barrier interoperability framework. *JCDL2001*, 2001.

[37] D. B. Lenat. CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11), 1995.

[38] C. A. Lindley. The FRAMES Processing Model for the Synthesis of Dynamic Virtual Video Sequences. *Database and Expert Systems Applications (DEXA'98), IEEE Computer Society*, 0:119–122, 1998.

[39] C. A. Lindley. Generic Film Forms for Dynamic Virtual Video Synthesis. *Multimedia Computing and Systems (icmcs), IEEE Computer Society*, 2:97–101, 1999.

[40] C. A. Lindley. A video annotation methodology for interactive video sequence generation. In *Digital content creation*, pages 163–183, New York, NY, USA, 2001. Springer-Verlag New York, Inc.

[41] S. Little, J. Geurts, and J. Hunter. Dynamic Generation of Intelligent Multimedia Presentations through Semantic Inferencing. In *6th European Conference on Research and Advanced Technology for Digital Libraries*, pages 158–189, Pontifical Gregorian University, Rome, Italy, September 2002. Springer.

[42] P. Lorentz. The River, 1937. IMDB entry: http://www.imdb.com/title/tt0029490/.

[43] W. C. Mann, C. M. I. M. Matthiesen, and S. A. Thompson. Rhetorical Structure Theory and Text Analysis. Technical Report ISI/RR-89-242, Information Sciences Institute, University of Southern California, November 1989.

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

[45] S. Melnik and S. Decker. Wordnet RDF Representation. http://www.semanticweb.org/library/, 2001.

[46] M. Moore. Roger & Me, 1989. IMDB entry: http://www.imdb.com/title/tt0098213/.

[47] M. Murtaugh. *The Automatist Storytelling System*. PhD thesis, Massachusetts Institute of Technology, 1996.

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

[49] F. Nack and W. Putz. Designing Annotation Before It's Needed. In *Proceedings of the 9th ACM International Conference on Multimedia*, pages 251–260, Ottawa, Ontario, Canada, September 30 - October 5, 2001.

[50] S. Newman and C. Marshall. Pushing Toulmin too far: learning from an argument representation scheme. Xerox PARC Technical Report SSL-92-45, 1992., 1992.

[51] A. Ortony. On making believable emotional agents believable. In R. Trappl, P. Petta, and S. Payr, editors, *Emotions in humans and artifacts*, pages 189–211. MIT Press, 2003.

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

[53] V. I. Pudovkin. *Film Technique and Film Acting*. Bonanza Books, 1949. translated by Ivor Montagu.

[54] M. Rabiger. *Directing the Documentary*. Focal Press, 1998.

[55] L. Riefenstahl. Olympia, Part Two: Festival of Beauty, 1936. IMDB entry: http://www.imdb.com/title/tt0030523/.

[56] C. Rocchi and M. Zancanaro. Generation of Video Documentaries from Discourse Structures. In *In Proceedings of Ninth European workshop on Natural Language Generation*, April 2003.

[57] L. Rutledge, M. Alberink, R. Brussee, S. Pokraev, W. van Dieten, and M. Veenstra. Finding the Story — Broader Applicability of Semantics and Discourse for Hypermedia Generation. In *Proceedings of the 14th ACM Conference on Hypertext and Hypermedia*, pages 67–76, Nottingham, UK, August 23-27, 2003. ACM, ACM Press.

[58] W. Sack. Coding News And Popular Culture. In *The International Joint Conference on Artificial Intelligence (IJCA93). Workshop on Models of Teaching and Models of Learning*, Chambery, Savoie, France, 1993.

[59] W. Sack and M. Davis. IDIC: Assembling Video Sequences from Story Plans and Content Annotations. In *IEEE International Conference on Multimedia Computing and Systems (ICMCS)*, Boston, Massachusetts, May 14-19, 1994.

[60] N. Sawhney, D. Balcom, and I. Smith. HyperCafe: Narrative and Aesthetic Properties of Hypervideo. In *Proc. of the Seventh ACM Conference on Hypertext*, pages 1–10, 1996.

[61] R. Schroeter, J. Hunter, and D. Kosovic. FilmEd - Collaborative Video Indexing, Annotation and Discussion Tools Over Broadband Networks. In *Proceedings of International Conference on Multi-Media Modeling*, pages 346–353, Brisbane Australia, 2004.

[62] K. Schwarz, T. Kouwenhoven, V. Dignum, and J. van Ossenbruggen. Supporting the decision process for the choice of a domain modeling scheme. In *Formal Ontologies Meet Industry*, Verona, Italy, June 9-10, 2005.

[63] S. B. Shum, E. Motta, and J. Domingue. ScholOnto: an Ontology-Based Digital Library Server for Research Documents and Discourse. *International Journal on Digital Libraries*, 3(3), August/September 2000.

[64] S. B. Shum, V. Uren, G. Li, J. Domingue, and E. Motta. Visualizing Internetworked Argumentation. In P. A. Kirschner, S. J. B. Shum, and C. S. Carr, editors, *Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making*, pages 185–204. Springer-Verlag: London, 2003.

[65] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval: the end of the early years. *IEEE trans.*, pages 1349–1380, 2000.

[66] T. G. A. Smith and G. Davenport. The Stratification System. A Design Environment for Random Access Video. In *ACM workshop on Networking and Operating System Support for Digital Audio and Video.*, San Diego, California, 1992.

[67] C. Snoek and M. Worring. Multimodal Video Indexing: A Review of the State-of-the-art. *Multimedia Tools and Applications*, 25(1):5–35, 2005.

[68] J. F. Sowa. *Conceptual Structures: Information Processing in Mind and Machine*. Addison-Wesley Publishing Company, 1984.

[69] R. Stam, R. Burgoyne, and S. Flitterman-Lewis. *New Vocabularies in Film Semiotics: Structuralism, Post-Structuralism, and Beyond.* Routledge, 1992.

[70] M. Theune, S. Faas, D. Heylen, and A. Nijholt. The virtual storyteller: Story creation by intelligent agents. In S. Göbel, N. Braun, U. Spierling, J. Dechau, and H. Diener, editors, *E 03: Technologies for Interactive Digital Storytelling and Entertainment*. Fraunhofer IRB Verlag, 2003.

[71] S. Toulmin, R. Rieke, and A. Janik. *Introduction to Reasoning*. MacMillan Publishing Company, 2 edition, 1984.

[72] M. van Assem, M. R. Menken, G. Schreiber, J. Wielemaker, and B. Wielinga. A Method for Converting Thesauri to RDF/OWL. In *Proceedings of the Third International Semantic Web Conference (ISWC'04)*, number 3298 in Lecture Notes in Computer Science, pages 17–31, Hiroshima, Japan, November 2004. Springer.

[73] J. van Ossenbruggen, F. Nack, and L. Hardman. That Obscure Object of Desire: Multimedia Metadata on the Web (Part I). *IEEE Multimedia*, 11(4):38–48, October – December 2004. based on http://ftp.cwi.nl/CWIreports/INS//INS-E0308.pdf.

[74] W3C. Synchronized Multimedia Integration Language (SMIL 2.0) Specification. W3C Recommendation, August 7, 2001. Edited by Aaron Cohen.

[75] W3C. RDF Vocabulary Description Language 1.0: RDF Schema. W3C Recommendation, 10 February 2004. Edited by Dan Brickley and R.V. Guha.

[76] T. Zhang and C.-C. J. Kuo. Heuristic approach for generic audio data segmentation and annotation. In *MULTIMEDIA '99: Proceedings of the seventh ACM international conference on Multimedia (Part 1)*, pages 67–76, Orlando, Florida, United States, 1999.

Index

Summary

Abstract

The context of this research is one or more online video repositories containing several hours of documentary footage and users possibly interested only in particular topics of that material. In such a setting it is not possible to craft a single version containing all possible topics the user might like to see, unless including all the material, which is clearly not feasible. The main motivation for this research is, therefore, to enable an alternative authoring process for film makers to make all their material dynamically available to users, without having to edit a static final cut that would select out possible informative footage.

We propose a methodology to automatically organize video material in an edited video sequence with a rhetorical structure. This is enabled by defining an annotation schema for the material and a generation process with the following two requirements:

- the data repository used by the generation process could be extended by simply adding annotated material to it
- the final resulting structure of the video generation process would seem familiar to a video literate user.

The first requirement was satisfied by developing an annotation schema that explicitly identifies rhetorical elements in the video material, and a generation process that assembles longer sequences of video by manipulating the annotations in a bottom-up fashion.

The second requirement was satisfied by modeling the generation process according to documentary making and general film theory techniques, in particular making the role of rhetoric in video documentaries explicit.

A specific case study was carried out using material for video documentaries. These used an interview structure, where people are asked to make statements about subjective matters. This category is characterized by rich information encoded in the audio track and by the controversy of the different opinions expressed in the interviews.

The approach was tested by implementing a system called Vox Populi that realizes a user-driven generation of rhetoric-based video sequences. Using the annotation schema, Vox Populi can be used to generate the story space and to allow the user to select and browse such a space. The user can specify the topic but also the characters of the rhetorical dialogue and the rhetoric form of the presentation.

Presenting controversial topics can introduce some bias: Vox Populi tries to control this by modeling some rhetoric and film theory editing techniques that influence the bias and by allowing the user to select the point of view she wants the generated sequence to have.

Overview

We present a model to automatically generate documentaries and an implementation of it. We focus on matter-of-opinion documentaries based on interviews. Our model has the following characteristics, which are lacking in previous automatic generation approaches:

- it allows the viewer to select the subject and the point of view of the documentary;
- it allows the documentarist to add material to the repository without having to specify how this material should be presented (data-driven approach);
- it generates documentaries according to presentation forms used by documentarists.

This thesis answers the following research questions:

RESEARCH QUESTION 1 *(*DOCUMENTARY FORM*) What characteristics of the presentation forms used by documentaries on matter-of-opinion issues must be modeled?*

RESEARCH QUESTION 2 *(*ANNOTATION SCHEMA*) What information should be captured in an annotation schema for an automatic video generation approach where:*

- *the viewer can specify the subject and the point of view,*
- *the documentarist can collect material to be used for documentaries, without having to specify how this material should be presented to the viewer,*
- *the material is presented according to presentation forms used by documentarists?*

RESEARCH QUESTION 3 *(*GENERATION PROCESS*) How must a generation process be defined for an automatic video generation approach where:*

- *the viewer can specify the subject and the point of view,*
- *the documentarist can collect material to be used for documentaries, without having to specify how this material should be presented to the viewer,*
- *the material is presented according to presentation forms used by documentarists?*

RESEARCH QUESTION 4 *(*AUTHOR SUPPORT*) How must a generation process be defined so that it can give to the documentarist an indication of the quality of the documentaries it can generate?*

Chapter 2

To determine what needs to be modeled, we analyze the domain of documentaries and the process of documentary making. This analysis leads to the definition of HIGH-LEVEL REQUIREMENTS, which specify the presentation forms a documentary generation model can use, and how to edit video material into a correct (according to

traditional film making) sequence. These high-level requirements provide an answer to *Research Question Documentary Form [1].*

In more detail, these high-level requirements restate the first two bullet points, while the third one is further specified using an analysis of the domain. The requirements point out the presentation forms that can be used in documentaries, namely the *narrative form* (where the presentation of information is organized into stories), the *categorical form* (where the presentation of information is organized into categories) and the *rhetorical form* (where the presentation of information is organized according to points of view, positions and arguments). We consider two levels in a story: the level of the scene, called micro-level, and the overall structure, called the macro-level. The narrative and categorical forms can be used at the macro-level, while the rhetorical form must be used at the micro level. The rhetorical form is particularly relevant for our domain, namely matter-of-opinion documentaries. This form is composed of *points of view* (propagandist and binary communicator), which communicate *positions* (e.g. "*war in Afghanistan - For*"), which in turn are expressed by *arguments*. Arguments are based on *logos*, *pathos* and *ethos* techniques. The high-level requirements also specify that the model should implement a montage technique often used in documentaries to present interviews. This technique, called vox populi, consists of showing in a rapid sequence how interviewees answer related questions. To avoid misquoting an interviewee, the generation model is required to encode context information for the statements made during interviews. For the editing part, the analysis of the documentary-making process requires the generation model to include *continuity editing* rules as used in traditional film making.

Chapter 3

Having defined what aspects of the domain need to be modeled, we examine how related work has solved similar problems, and determine which existing technical solutions are feasible given the high-level requirements we set. This analysis leads to the definition of LOW-LEVEL REQUIREMENTS. These requirements are divided into two groups. The first group specifies what *data structure* can represent video material for the purpose of documentary generation. The second group determines the characteristics of a *process*that is capable of generating documentaries according to the high-level requirements.

In more detail, the first group of requirements concerning the annotations specify that the video material should be segmented into discrete units called *clips*. The description of the clips should capture *connotative* as well as *denotative* aspects of the video material, using *property-based* annotations and a *controlled vocabulary*. Arguments contained in interviews and based on logos should be encoded by an argument model, the *model of Toulmin*. Arguments based on pathos and ethos should be evaluated using a cognitive model, the *OCC model*. In addition to the OCC model, film theory provides another method to evaluate pathos, based on the *cinematic properties* of the clip, namely gaze direction and framing distance. The second group of requirements specify that the generation process should dynamically create, using the annotations, a data structure (the *Semantic Graph*) that provides information about the argumentation relations (SUPPORTS and CONTRADICTS) among media items in the repository. Furthermore, based on argumentation theory the requirements define a means of composing arguments from single statements, such as *rebuttals* and *undercutters*, and specify that the *categorical form* should be used as the presentation form at the macro-level.

Chapter 4

Guided by the high-level requirements and the first group of low-level requirements, we examine the content of video to determine the characteristics of the information we need to model. Based on this analysis, we specify an annotation schema capable of encoding the *rhetorical form* and the *categorical form*, and the *cinematic properties* of video to support automatic editing. The definition of this annotation schema provides an answer to *Research Question Annotation Schema [2].*

In more detail, two components of the rhetorical form are modeled, namely arguments and positions. Arguments based on *logos* are encoded by modeling verbal information contained in the auditory and visual channel. The arguments are modeled using three-part sentence-like descriptions of what an interviewee says, called *statements*, a *thesaurus* for the controlled vocabulary of terms used in the statements and the *model of Toulmin* for the role each statement plays in an argument. Arguments based on *pathos* are modeled using non-verbal information contained in the visual channel, by modeling the clip cinematic properties *framing distance* and *gaze direction*. *Ethos* is modeled based on the OCC model, by using verbal and non-verbal information to determine *social categories* an interviewee belongs to, such as gender, race, education level, and a user profile that values how important these categories are for the viewer. *Positions* are modeled as a subject and the interviewee's attitude with respect to that subject, e.g. "*war in Afghanistan - For*". Further we define the categories to support the categorical form, namely *categories related to interviews*, such as question asked, *location categories* describing where the clip was shot, such as the geographical location, and *temporal categories* describing when the clip was shot, such as the time of the day. Finally, the cinematic properties of video are modeled to support the *continuity rules*, such as gaze direction for the *gaze continuity* rule and framing distance for the *framing continuity* rule (both properties are also required to calculate pathos).

Chapter 5

Having encoded the information needed to generate documentaries of the form specified by the high-level requirements, we describe a process capable of generating these documentaries. This generation process first creates the *Semantic Graph*, a data structure that establishes the argumentation relations among media items, then manipulates this structure to form arguments using video clips. The selected video clips are presented according to the rhetorical form and the categorical form, in a video sequence that also satisfies the continuity rules and the montage specified in the high-level requirements. The definition of this generation process provides an answer to *Research Question Generation Process [3].* We then describe methods to provide the documentarist with a means of verifying the correctness of the annotations. The specification of these methods answers *Research Question Author Support [4].*

In more detail, the generation process dynamically creates the *Semantic Graph* in two steps: the first one generates possible candidate targets for linking using the statements and the relations in the thesaurus. The second one verifies which of the possible targets is associated to media items present in the repository. The result of these two steps is a graph where the edges are the argumentation relations CONTRADICTS and SUPPORTS and the nodes correspond to media items. This structure is used to assemble arguments that show supporting or conflicting positions, using actions such as rebuttals or undercutters. Pathos and ethos are used to assess which side in a conflicting argument appears more convincing to the viewer. This allows the generation of video

sequences that express a particular point of view, i.e. the propagandist where one side is more convincing than the other, or the binary communicator where both sides appear equally convincing. Selected video clips are then edited using rhetoric-driven editing such as *shot-reverse shot* and continuity rules such as *framing continuity*. The process then uses the categorical form to assemble more arguments together and form longer video sequences. The resulting generation process is driven by the viewer requests, as specified in the SUBJECT-POINT OF VIEW [HLR 2] requirement.

The *feedback method* aims at pinpointing where the annotations do not fully support the *Semantic Graph* creation. These methods are based on the definition of indexes that measure the performance of the two steps used to create the graph. The documentarist can also use these indexes for two other purposes: to suggest possible annotations to be specified in the thesaurus, and to fine-tune the process of graph creation.

Chapter 6

Having defined the model, we provide an implementation of it with a demonstrator called *Vox Populi*. Vox Populi's architecture consists of a web user interface, through which the viewer specifies the subject and the point of view of the documentary, core functionality running inside a web server and a storage back-end for the annotations. The documentarist can create the repository using an annotations editor and video editing tools. Vox Populi has also been used in two other projects, the *Visual Jockey* and *Passepartout - move.me* projects. We further report the result of a technical evaluation and our own experiences in using the system as documentarists.

Chapter 7

In this chapter we present an overview of the thesis. Our research contributions are the high-level requirements in chapter 2, the low-level requirements in chapter 3, the automatic video generation model composed of the annotation schema in chapter 4 and of the generation process in chapter 5, and the implementation of the model in chapter 6. We then discuss general issues for an automatic video generation approach (the annotation effort required, the influence of the documentarist/annotator in the process, and the consequences deriving from the open-world assumption we make), as well as issues related to each research questions. We conclude examining future directions for our work, the most promising of which is the modeling of non-verbal information and its influence on pathos.

Samenvatting

Het onderzoek beschreven in dit proefschrift betreft het voor het publiek toegankelijk maken van videomateriaal dat gebruikt wordt voor het produceren van videodocumentaires. In een traditioneel productie proces selecteert een documentairemaker de fragmenten die zij geschikt acht en monteert deze vervolgens tot een documentaire. Hoewel het oorspronkelijk videomateriaal objectief genoemd mag worden ziet het publiek een documentaire die gekleurd is door keuzes van de documentairemaker. Desalniettemin is het resterende videomateriaal mogelijk interessant voor het publiek. In dit proefschrift beschrijven we een methode voor een alternatief productie proces voor documentaires. Dit proces stelt de filmmaker in staat het beschikbare videomateriaal, afhankelijk van de specifieke interesse van het publiek, te presenteren.

Onze methode selecteert, afhankelijk van de wensen van het publiek, videofragmenten en monteert deze automatisch tot een logische eenheid. Om een juiste selectie te kunnen maken moet een automatisch productieproces gedefinieerd worden en moeten de videofragmenten voorzien zijn van een expliciete beschrijving. Zowel de beschrijving, alsmede het productieproces dat de beschrijving gebruikt, voldoen aan de volgende twee requirements:

- Het bestaande videoarchief moet dynamisch uitbreidbaar zijn. Dit betekent dat nieuw materiaal toegevoegd moet kunnen worden zonder dat reeds bestaand materiaal moet worden aangepast.
- De automatisch gegenereerde video moet begrijpelijk zijn en aansluiten bij het verwachtingspatroon van het publiek.

Ons productiemodel formaliseert de retorische elementen in een videofragment en algemene filmtheoretische kennis, alsmede specifieke kennis voor het produceren van documentaires. Dit model bestaat uit een generatieproces dat relevante videofragmenten selecteert op basis van hun retorische en filmtheoretische beschrijving.

We hebben onze methode praktisch getest door het implementeren van het Vox Populi systeem. Vox Populi genereert automatisch een sequentie van videofragmenten waarbij de fragmenten geselecteerd worden op basis van door het publiek ingevoerde gegevens, zoals het onderwerp, de karakter types en de retorische structuur. Tenslotte, probeert Vox Populi de inherente subjectiviteit in een documentaire inzichtelijk te maken door het publiek te laten kiezen van uit welk perspectief de video genereerd wordt.

Curriculum Vitae

Stefano Bocconi was born on the 12th of October, 1967 in Florence, Italy. He attended classical high-school (Italian "Liceo Classico") from 1982 till 1987 where he graduated with the final mark of 60/60. From 1987 till 1994 he studied Electronic Engineering at the Universitá degli Studi in Florence, with specialization in Control Theory. He graduated with a thesis titled "Retraining of Neural Networks" and his final mark was with 110/110 cum laude. From 1994 till 2001 he worked in industry as a programmer and software engineer, both in Italy and in the Netherlands. At the beginning of 2002 he went back to academia as a PhD student at the Centrum voor Wiskunde en Informatica, Amsterdam, where he worked till the end of 2005. In 2006 he obtained a research position at the University of Turin, Italy.

His main research interest is multimedia semantics, as well as discourse, narrative and, in general, all the disciplines necessary to make user-driven presentations or stories out of a collection of media items.

SIKS Dissertation Series

====
1998
====

1998-1 Johan van den Akker (CWI)
DEGAS - An Active, Temporal Database of Autonomous Objects

1998-2 Floris Wiesman (UM)
Information Retrieval by Graphically Browsing Meta-Information

1998-3 Ans Steuten (TUD)
A Contribution to the Linguistic Analysis of Business Conversations within the Language/Action Perspective

1998-4 Dennis Breuker (UM)
Memory versus Search in Games

1998-5 E.W.Oskamp (RUL)
Computerondersteuning bij Straftoemeting

====
1999
====

1999-1 Mark Sloof (VU)
Physiology of Quality Change Modelling; Automated modelling of Quality Change of Agricultural Products

1999-2 Rob Potharst (EUR)
Classification using decision trees and neural nets

1999-3 Don Beal (UM)
The Nature of Minimax Search

1999-4 Jacques Penders (UM)
The practical Art of Moving Physical Objects

1999-5 Aldo de Moor (KUB)
Empowering Communities: A Method for the Legitimate User-Driven Specification of Network Information Systems

1999-6 Niek J.E. Wijngaards (VU)
Re-design of compositional systems

1999-7 David Spelt (UT)
Verification support for object database design

1999-8 Jacques H.J. Lenting (UM)
Informed Gambling: Conception and Analysis of a Multi-Agent Mechanism for Discrete Reallocation.

====
2000
====

2000-1 Frank Niessink (VU)
Perspectives on Improving Software Maintenance

2000-2 Koen Holtman (TUE)
Prototyping of CMS Storage Management

2000-3 Carolien M.T. Metselaar (UVA)
Sociaal-organisatorische gevolgen van kennistechnologie;
een procesbenadering en actorperspectief.

2000-4 Geert de Haan (VU)
ETAG, A Formal Model of Competence Knowledge for User Interface
Design

2000-5 Ruud van der Pol (UM)
Knowledge-based Query Formulation in Information Retrieval.

2000-6 Rogier van Eijk (UU)
Programming Languages for Agent Communication

2000-7 Niels Peek (UU)
Decision-theoretic Planning of Clinical Patient Management

2000-8 Veerle Coup (EUR)
Sensitivity Analyis of Decision-Theoretic Networks

2000-9 Florian Waas (CWI)
Principles of Probabilistic Query Optimization

2000-10 Niels Nes (CWI)
Image Database Management System Design Considerations, Algorithms
and Architecture

2000-11 Jonas Karlsson (CWI)
Scalable Distributed Data Structures for Database Management

====
2001
====

2001-1 Silja Renooij (UU)
Qualitative Approaches to Quantifying Probabilistic Networks

2001-2 Koen Hindriks (UU)
Agent Programming Languages: Programming with Mental Models

2001-3 Maarten van Someren (UvA)
Learning as problem solving

2001-4 Evgueni Smirnov (UM)
Conjunctive and Disjunctive Version Spaces with Instance-Based
Boundary Sets

2001-5 Jacco van Ossenbruggen (VU)
Processing Structured Hypermedia: A Matter of Style

2001-6 Martijn van Welie (VU)
Task-based User Interface Design

2001-7 Bastiaan Schonhage (VU)
Diva: Architectural Perspectives on Information Visualization

2001-8 Pascal van Eck (VU)
A Compositional Semantic Structure for Multi-Agent Systems Dynamics.

2001-9 Pieter Jan 't Hoen (RUL)
Towards Distributed Development of Large Object-Oriented Models,
Views of Packages as Classes

2001-10 Maarten Sierhuis (UvA)

Modeling and Simulating Work Practice
BRAHMS: a multiagent modeling and simulation language
for work practice analysis and design

2001-11 Tom M. van Engers (VUA)
Knowledge Management:
The Role of Mental Models in Business Systems Design

====
2002
====

2002-01 Nico Lassing (VU)
Architecture-Level Modifiability Analysis

2002-02 Roelof van Zwol (UT)
Modelling and searching web-based document collections

2002-03 Henk Ernst Blok (UT)
Database Optimization Aspects for Information Retrieval

2002-04 Juan Roberto Castelo Valdueza (UU)
The Discrete Acyclic Digraph Markov Model in Data Mining

2002-05 Radu Serban (VU)
The Private Cyberspace Modeling Electronic Environments
inhabited by Privacy-concerned Agents

2002-06 Laurens Mommers (UL)
Applied legal epistemology;
Building a knowledge-based ontology of the legal domain

2002-07 Peter Boncz (CWI)
Monet: A Next-Generation DBMS Kernel For Query-Intensive
Applications

2002-08 Jaap Gordijn (VU)
Value Based Requirements Engineering: Exploring Innovative
E-Commerce Ideas

2002-09 Willem-Jan van den Heuvel(KUB)
Integrating Modern Business Applications with Objectified Legacy
Systems

2002-10 Brian Sheppard (UM)
Towards Perfect Play of Scrabble

2002-11 Wouter C.A. Wijngaards (VU)
Agent Based Modelling of Dynamics: Biological and Organisational
Applications

2002-12 Albrecht Schmidt (Uva)
Processing XML in Database Systems

2002-13 Hongjing Wu (TUE)
A Reference Architecture for Adaptive Hypermedia Applications

2002-14 Wieke de Vries (UU)
Agent Interaction: Abstract Approaches to Modelling, Programming and
Verifying Multi-Agent Systems

2002-15 Rik Eshuis (UT)
Semantics and Verification of UML Activity Diagrams for Workflow Modelling

2002-16 Pieter van Langen (VU)
The Anatomy of Design: Foundations, Models and Applications

2002-17 Stefan Manegold (UVA)
Understanding, Modeling, and Improving Main-Memory Database Performance

====

2003
====

2003-01 Heiner Stuckenschmidt (VU)
Ontology-Based Information Sharing in Weakly Structured Environments

2003-02 Jan Broersen (VU)
Modal Action Logics for Reasoning About Reactive Systems

2003-03 Martijn Schuemie (TUD)
Human-Computer Interaction and Presence in Virtual Reality Exposure Therapy

2003-04 Milan Petkovic (UT)
Content-Based Video Retrieval Supported by Database Technology

2003-05 Jos Lehmann (UVA)
Causation in Artificial Intelligence and Law - A modelling approach

2003-06 Boris van Schooten (UT)
Development and specification of virtual environments

2003-07 Machiel Jansen (UvA)
Formal Explorations of Knowledge Intensive Tasks

2003-08 Yongping Ran (UM)
Repair Based Scheduling

2003-09 Rens Kortmann (UM)
The resolution of visually guided behaviour

2003-10 Andreas Lincke (UvT)
Electronic Business Negotiation: Some experimental studies on the interaction between medium, innovation context and culture

2003-11 Simon Keizer (UT)
Reasoning under Uncertainty in Natural Language Dialogue using Bayesian Networks

2003-12 Roeland Ordelman (UT)
Dutch speech recognition in multimedia information retrieval

2003-13 Jeroen Donkers (UM)
Nosce Hostem - Searching with Opponent Models

2003-14 Stijn Hoppenbrouwers (KUN)
Freezing Language: Conceptualisation Processes across ICT-Supported Organisations

2003-15 Mathijs de Weerdt (TUD)
Plan Merging in Multi-Agent Systems

2003-16 Menzo Windhouwer (CWI)
Feature Grammar Systems - Incremental Maintenance of Indexes to Digital Media Warehouses

2003-17 David Jansen (UT)
Extensions of Statecharts with Probability, Time, and Stochastic Timing

2003-18 Levente Kocsis (UM)
Learning Search Decisions

====
2004
====

2004-01 Virginia Dignum (UU)
A Model for Organizational Interaction: Based on Agents, Founded in Logic

2004-02 Lai Xu (UvT)
Monitoring Multi-party Contracts for E-business

2004-03 Perry Groot (VU)
A Theoretical and Empirical Analysis of Approximation in Symbolic Problem Solving

2004-04 Chris van Aart (UVA)
Organizational Principles for Multi-Agent Architectures

2004-05 Viara Popova (EUR)
Knowledge discovery and monotonicity

2004-06 Bart-Jan Hommes (TUD)
The Evaluation of Business Process Modeling Techniques

2004-07 Elise Boltjes (UM)
Voorbeeldig onderwijs; voorbeeldgestuurd onderwijs, een opstap naar abstract denken, vooral voor meisjes

2004-08 Joop Verbeek(UM)
Politie en de Nieuwe Internationale Informatiemarkt, Grensregionale politi
”ele gegevensuitwisseling en digitale expertise

2004-09 Martin Caminada (VU)
For the Sake of the Argument; explorations into argument-based reasoning

2004-10 Suzanne Kabel (UVA)
Knowledge-rich indexing of learning-objects

2004-11 Michel Klein (VU)
Change Management for Distributed Ontologies

2004-12 The Duy Bui (UT)
Creating emotions and facial expressions for embodied agents

2004-13 Wojciech Jamroga (UT)
Using Multiple Models of Reality: On Agents who Know how to Play

2004-14 Paul Harrenstein (UU)
Logic in Conflict. Logical Explorations in Strategic Equilibrium

2004-15 Arno Knobbe (UU)
Multi-Relational Data Mining

2004-16 Federico Divina (VU)
Hybrid Genetic Relational Search for Inductive Learning

2004-17 Mark Winands (UM)
Informed Search in Complex Games

2004-18 Vania Bessa Machado (UvA)
Supporting the Construction of Qualitative Knowledge Models

2004-19 Thijs Westerveld (UT)
Using generative probabilistic models for multimedia retrieval

2004-20 Madelon Evers (Nyenrode)
Learning from Design: facilitating multidisciplinary design teams

====
2005
====

2005-01 Floor Verdenius (UVA)
Methodological Aspects of Designing Induction-Based Applications

2005-02 Erik van der Werf (UM))
AI techniques for the game of Go

2005-03 Franc Grootjen (RUN)
A Pragmatic Approach to the Conceptualisation of Language

2005-04 Nirvana Meratnia (UT)
Towards Database Support for Moving Object data

2005-05 Gabriel Infante-Lopez (UVA)
Two-Level Probabilistic Grammars for Natural Language Parsing

2005-06 Pieter Spronck (UM)
Adaptive Game AI

2005-07 Flavius Frasincar (TUE)
Hypermedia Presentation Generation for Semantic Web Information Systems

2005-08 Richard Vdovjak (TUE)
A Model-driven Approach for Building Distributed Ontology-based Web Applications

2005-09 Jeen Broekstra (VU)
Storage, Querying and Inferencing for Semantic Web Languages

2005-10 Anders Bouwer (UVA)
Explaining Behaviour: Using Qualitative Simulation in Interactive Learning Environments

2005-11 Elth Ogston (VU)
Agent Based Matchmaking and Clustering - A Decentralized Approach to Search

2005-12 Csaba Boer (EUR)
Distributed Simulation in Industry

2005-13 Fred Hamburg (UL)
Een Computermodel voor het Ondersteunen van Euthanasiebeslissingen

2005-14 Borys Omelayenko (VU)
Web-Service configuration on the Semantic Web; Exploring how semantics meets pragmatics

2005-15 Tibor Bosse (VU)
Analysis of the Dynamics of Cognitive Processes

2005-16 Joris Graaumans (UU)
Usability of XML Query Languages

2005-17 Boris Shishkov (TUD)
Software Specification Based on Re-usable Business Components

2005-18 Danielle Sent (UU)
Test-selection strategies for probabilistic networks

2005-19 Michel van Dartel (UM)
Situated Representation

2005-20 Cristina Coteanu (UL)
Cyber Consumer Law, State of the Art and Perspectives

2005-21 Wijnand Derks (UT)
Improving Concurrency and Recovery in Database Systems by Exploiting Application Semantics

====
2006
====

2006-01 Samuil Angelov (TUE)
Foundations of B2B Electronic Contracting

2006-02 Cristina Chisalita (VU)
Contextual issues in the design and use of information technology in organizations

2006-03 Noor Christoph (UVA)
The role of metacognitive skills in learning to solve problems

2006-04 Marta Sabou (VU)
Building Web Service Ontologies

2006-05 Cees Pierik (UU)
Validation Techniques for Object-Oriented Proof Outlines

2006-06 Ziv Baida (VU)
Software-aided Service Bundling - Intelligent Methods & Tools
for Graphical Service Modeling

2006-07 Marko Smiljanic (UT)
XML schema matching – balancing efficiency and effectiveness
by means of clustering

2006-08 Eelco Herder (UT)
Forward, Back and Home Again - Analyzing User Behavior on the Web

2006-09 Mohamed Wahdan (UM)
Automatic Formulation of the Auditor's Opinion

2006-10 Ronny Siebes (VU)
Semantic Routing in Peer-to-Peer Systems

2006-11 Joeri van Ruth (UT)
Flattening Queries over Nested Data Types

2006-12 Bert Bongers (VU)
Interactivation - Towards an e-cology of people, our technological
environment, and the arts

2006-13 Henk-Jan Lebbink (UU)
Dialogue and Decision Games for Information Exchanging Agents

2006-14 Johan Hoorn (VU)
Software Requirements: Update, Upgrade, Redesign - towards a Theory of
Requirements Change

2006-15 Rainer Malik (UU)
CONAN: Text Mining in the Biomedical Domain

2006-16 Carsten Riggelsen (UU)
Approximation Methods for Efficient Learning of Bayesian Networks

2006-17 Stacey Nagata (UU)
User Assistance for Multitasking with Interruptions on a Mobile Device

2006-18 Valentin Zhizhkun (UVA)
Graph transformation for Natural Language Processing

2006-19 Birna van Riemsdijk (UU)
Cognitive Agent Programming: A Semantic Approach

2006-20 Marina Velikova (UvT)
Monotone models for prediction in data mining

2006-21 Bas van Gils (RUN)
Aptness on the Web

2006-22 Paul de Vrieze (RUN)
Fundaments of Adaptive Personalisation

2006-23 Ion Juvina (UU)
Development of Cognitive Model for Navigating on the Web

2006-24 Laura Hollink (VU)
Semantic Annotation for Retrieval of Visual Resources

2006-25 Madalina Drugan (UU)
Conditional log-likelihood MDL and Evolutionary MCMC

2006-26 Vojkan Mihajlovic (UT)
Score Region Algebra: A Flexible Framework for Structured Information Retrieval