Video Abstracting at a Semantical Level

Video Abstracting at a Semantical Level

Dissertation
zur Erlangung des akademischen Grades eines
Dr.-Ing.

von
Dipl.-Inf. Till von Wenzlawowicz
aus
Marburg

Vorgelegt dem
Fachbereich 3
der
Universität Bremen
im
Mai 2017

Gutachter:
Prof. Dr. rer. nat. Otthein Herzog
Prof. Dr. phil. John A. Bateman

Bremen

Datum des Promotionskolloquiums: 02.02.2018

Zusammenfassung

Eine der bekanntesten Formen von *Video Abstracts* ist der Filmtrailer. Zeitgenössische Trailer haben genreunabhängig ähnliche Strukturen, die eine automatische Generierung erlauben und den Aufbau des entsprechenden Films in Teilen wiederspiegeln. In dieser Arbeit wird ein System für Actiontrailer vorgestellt, das zusätzlich zu Actiontrailern auch andere Genres beherrscht. Dazu wurden Horror- und Komödientrailer manuell analysiert, und ihre Struktur indentifiziert und formalisiert.

Um die Modellierung der für die automatische Generierung nötigen Abstractmodelle zu ermöglichen wurde ein grafisches Programm entwickelt, das sämtliche Schritte der Generierung von Video Abstracts automatisch durchführt, gleichzeitig aber auch eine Vorschau sowie eine einfache manuelle Optimierung erlaubt.

Darauf aufbauend wurde zusätzlich zum Actiontrailermodell je ein Abstractmodell für Horror- und Komödientrailer erstellt, mit dem Trailer für Horrorfilme und Komödien generiert wurden. In einer abschließenden Evaluierung zeigte sich, dass die automatisch erstellten Trailer eine ähnliche Struktur wie die originalen Hollywoodtrailer aufweisen, auch wenn sie noch nicht gehobene künstlerischen Ansprüche befriedigen können.

Abstract

One the most common form of a video abstract is the movie trailer. Contemporary movie trailers share a common structure across genres which allows for an automatic generation and also reflects the corresponding movie's composition. In this thesis a system for the automatic generation of trailers is presented. In addition to action trailers, the system is able to deal with further genres such as horror and comedy trailers, which were first manually analyzed in order to identify their basic structures.

To simplify the modeling of trailers and the abstract generation itself a new video abstracting application was developed. This application is capable of performing all steps of the abstract generation automatically and allows for previews and manual optimizations.

Based on this system, new abstracting models for horror and comedy trailers were created and the corresponding trailers have been automatically generated using the new abstracting models. In an evaluation the automatic trailers were compared to the original trailers and showed a similar structure. However, the automatically generated trailers still do not exhibit the full perfection of the Hollywood originals as they lack intentional storylines across shots.

Danksagung

An dieser Stelle danke ich vor allem Professor Herzog, der mir die Promotion ermöglicht und mich über all die Jahre betreut, beraten und unterstüzt hat. Ebenso danke ich Professor Bateman für die Betreuung als Zweitgutachter und speziell für seine Expertise im Bereich Film. Großer Dank geht an das Graduiertenkolleg "Advances in Digital Media" von Professor Malaka für die Begleitung und Betreuung während meiner Promotionszeit und an die Klaus-Tschira-Stiftung für die Gewährung eines Stipendiums.

Ohne die Arbeit der Mitglieder des SVP-Projekts wäre meine Promotion nicht möglich gewesen. Ich danke den Teilnehmern Christoph Brachmann, Hashim Chunpir, Silke Gennies, Benjamin Haller, Philipp Kehl, Astrid Paramita Mochtarram, Daniel Möhlmann, Christian Schrumpf, Christopher Schultz, Björn Stolper und Benjamin Walther-Franks sowie den Betreuern Arne Jacobs, Thorsten Hermes und Otthein Herzog. Ebenfalls danken möchte ich den Teilnehmern des Projekts ADDiCT.

Eine große Hilfe waren mir auch meine Korrekturleser Nick, Marie, Nadir und Reginald. Danken möchte ich meinen Eltern für ihre Unterstützung, und ganz speziell Jan, Jorge, Momo und Ingrid.

Contents

List of Figures xiii

List of Tables xv

1 Introduction 1

2 Theoretical Aspects of Movies and Trailers 3
2.1 Movies . . . 3
2.1.1 Narrative Structure . . . 3
2.1.2 Narration . . . 5
2.1.3 Characters and Events . . . 6
2.1.4 Expectations and Conventions . . . 6
2.1.5 Emotions . . . 7
2.1.6 Narrative Form . . . 8
2.1.7 Shooting Film . . . 9
2.1.7.1 Mise-en-Scène . . . 9
2.1.7.2 Cinematography . . . 17
2.1.7.3 Editing . . . 25
2.1.7.4 Sound . . . 31
2.1.8 Movie Genres . . . 37
2.2 Trailers . . . 44
2.2.1 Overview . . . 44
2.2.2 Trailer Types . . . 44
2.2.3 History of Trailers . . . 47
2.2.4 Contemporary Trailer Structure . . . 51

3 Automatic Video Abstracting 53
3.1 Video Abstracting . . . 53
3.1.1 Keyframe Systems . . . 53
3.1.2 Skimming Systems . . . 56
3.1.3 Evaluation Methods . . . 63

3.2 Trailer Generation Systems 65
3.2.1 Automatic Generated Recommendation for Movie Trailers . . 66
3.2.2 Automatically Selecting Shots for Action Movie Trailers . . . 67
3.2.3 Automated Production of TV Program Trailer using Electronic Program Guide 68
3.2.4 Animation Movies Trailer Computation 69
3.2.5 Video Abstracting 70
3.2.6 Automatic Trailer Generation 72
3.2.7 Semantic Video Patterns in Action Movies 75
3.2.7.1 Syntax Analysis 75
3.2.7.2 Analyzer 79
3.2.7.3 Generator 91

4 A Multi-Genre Approach to Movie Trailers 99
4.1 Analysis of Hollywood Trailers 99
4.1.1 Genre Selection 99
4.1.2 Manual Annotation 100
4.1.2.1 Horror 100
4.1.2.2 Comedy 105
4.1.3 Structure of Horror and Comedy Trailers 109
4.2 Abstracting Application 112
4.2.1 Shortcomings of the SVP System 112
4.2.2 Requirements for a new Application 113

5 Implementation 115
5.1 Preparational Work 115
5.2 Application Design 115
5.2.1 Video Abstracting Workflow 116
5.2.2 User Interface 116
5.2.2.1 Categorizer 116
5.2.2.2 Video Abstracting Model Editor 119
5.2.2.3 Video Abstract Viewer 122
5.3 Video Abstracting Algorithms 123
5.3.1 Categorization 123
5.3.2 Abstract Building 125
5.4 Application Structure 128
5.4.1 Data Models 128
5.4.1.1 Categories 128

5.4.1.2 Abstraction Model 129
5.4.1.3 Movie Data 130
5.4.1.4 Additional Data 132
5.4.2 Application Modules 132
5.5 Implementation of Video Abstracting Models 135
5.5.1 Action Trailer Model 135
5.5.2 Semantic Footage Categories 139
5.5.3 Creating Models 141
5.6 Additional Functions 145
5.6.1 Speech Part Detection 145
5.6.2 Analyzer Module Controller 145
5.6.3 Trailer Annotation Player 146
5.6.4 SVP Data Converter 146

6 Evaluation **147**
6.1 Video Abstracting Application 147
6.2 Automatically Generated Trailers 148
6.2.1 Evaluation Methods for Video Abstracts 148
6.2.2 Scenario 148
6.2.2.1 Criteria 149
6.2.2.2 Trailer Selection 150
6.2.3 Semantics 150
6.2.3.1 Horror Trailers 151
6.2.3.2 Comedy Trailers 156
6.2.4 Technical Quality 163
6.2.5 Summary 164

7 Conclusion and Future Work **167**

Bibliography **169**

List of Figures

2.1 The Structure of a Movie according to the Paradigm by Field 4

3.1 XML description showing the results of the analyzing step 79
3.2 Categorization Example 92
3.3 Hierarchical Trailer Model of the SVP System 93
3.4 Different animation styles used in the SVP system 95

4.1 The shot-based manual annotation of the movie Bruce Almighty 101
4.2 Semantical Sequences in the *Dreamcatcher* Trailer 102
4.3 Semantical Sequences in the *Texas Chainsaw Massacre* Trailer 104
4.4 Semantical Units in the *Bruce Almighty* Trailer 106
4.5 Semantical Units in the *Hitch* Trailer 108

5.1 Categorizer Widget 117
5.2 Category Parameter Dialogue 118
5.3 Model Viewer Widget 119
5.4 Node Edit Dialogue 120
5.5 Clip/Transition Dialogue 121
5.6 Abstract Viewer Widget 122
5.7 Feature Tracks used in the Categorizing Algorithm 124
5.8 The Abstract Model 127
5.9 Category and Category Parameter Data Type 128
5.10 Abstracting Model Node Data Type 129
5.11 Clip/Transition Pair Data Type 130
5.12 UML-Diagram of the Structure MovieData 131
5.13 Overview of the Application Structure and Main Data Flow 133
5.14 Abstracting Model Excerpt 144

6.1 Comparison of Semantical Units in the Intro Phase of Horror Trailers . 152
6.2 Comparison of Semantical Units in the Story Phase of Horror Trailers . 153
6.3 Comparison of Semantical Units in the Action Phase of Horror Trailers 155
6.4 Comparison of Semantical Units in the Outro Phase of Horror Trailers . 156

6.5 Comparison of Semantical Units in the Intro Phase of Comedy Trailers . 157
6.6 Comparison of Semantical Units in the Story Phase of Comedy Trailers 159
6.7 Comparison of Semantical Units in the Action Phase of Comedy Trailers 160
6.8 Comparison of Semantical Units in the Outro Phase of Comedy Trailers 162

List of Tables

3.1 Footage categories defined by the SVP team 77
3.2 Absolute Movement Magnitude Ranges 82

4.1 Generic Structure of a Horror Trailer 110
4.2 Generic Structure of a Comedy Trailer 111

5.1 Clip Arrangement in the Abstracting Model of Action Trailers 136
5.2 Action Trailer Model, Action Phase Variant A 137
5.3 Action Trailer Model, Action Phase Variant B 138
5.4 Action Trailer Model, Action Phase Variant C 139
5.5 Clip categories used for abstract generation 140
5.6 Animation categories used for abstract generation 141
5.7 Clip Arrangement in the Abstracting Model of Horror Trailers 142
5.8 Clip Arrangement in the Abstracting Model of Comedy Trailers 143

6.1 Trailer and Corresponding Movie Selection for the Evaluation 150
6.2 Comparison of Formal Parameters of Trailers 164

Chapter 1

Introduction

In a scientific publication an abstract is normally provided to summarize the content of the work and to give the reader a brief overview of what to expect. Furthermore, it allows for positioning the publication in the scientific domain.

In a similar way, a video abstract summarizes the content of a longer video. In the domain of video abstracting, two main approaches are popular: key frames and video skims. While the key frame generation results in one or more still images, video skimming produces a shorter version of the original video. The aim of a video abstract is to preserve as much information as possible from the original. [Truong and Venkatesh, 2007]

A special case of video abstracting are movie trailers. The possibly most common reason for creating and distributing a trailer is to use it as the main advertisement for the upcoming movie in movie theaters, on television and online. Throughout the history of cinema, general structures of arranging a trailer were developed. The contemporary Hollywood trailer mimics the film by summarizing the first 2/3 of its story. It uses drama and a sound track especially composed for the movie trailer and it features elements like artistic icon and text animations of actors, producers and directors. [Hediger, 2001]

This general structure allows for an attempt at computing a trailer automatically. Breaking down the trailer into smaller units allows this structure to be formalized and algorithmically used. Utilizing several audio and video processing modules, a movie can be segmented into units which can be used as source to automatically compile a trailer. By generating textual animations and by adding a dynamically generated soundtrack, such a trailer can be completed automatically. A first approach has been shown to be successful in the master students project *SVP* [Brachmann et al., 2006] and *ADDiCT* [Asaad et al., 2008] for action movies and action trailers.

Movies of the genre action share a variety of typical elements, such as gun shots, explosions, spectacular stunt scenes, car chases and general fast-paced action shots.

These elements typically come along with distinct visual and auditory features which allow for a quite reliable automatic detection of these sequences. In Hollywood trailers these sequences are rearranged and combined with shots of locations and characters, text animations and company logos. The combination of all these different types of footage forms the basic units a trailer can be composed of.

For action movies and trailers, this set of footage categories is comparatively large and allows the choice of a variety of distinct sequence types. Furthermore, some footage categories serve similar narrative purposes, e.g., it does not make a big difference for the dramaturgy if an explosion is replaced by a gun shot. An interesting question is how the existing abstracting approach performs in other genres, especially those that do not feature so much visually and auditory spectacular sequences but rather focus on speech and acting. These sequences require the audience to interpret the meaning, either by listening to and understanding the words or by recognizing gestures and actions.

From an image and audio processing perspective, the differences between distinct footage categories are extensive in the action genre and rather subtle in the comedy genre. This demonstrates a need for further methods of segmentation.

The research question of this thesis can thus be described as follows: *To which extent is it possible to automatically generate Hollywood-like movie trailers beyond the action movie genre?*

In the next chapter a theoretical overview of the domain of film and trailers is presented. The first part is organized similarly to the production cycle of a movie. In the beginning, narration and the structure of movies and their central elements as well as the role of the audience are introduced. Subsequently the shooting and post-production of films is elaborated, and finally the analysis and classification of films by means of the genres outlined.In the second part of chapter 2 the different types of movie trailers are introduced and an overview of the historic development is given. The chapter closes with a description of the probably most common trailer type, the contemporary theatrical trailer.

Chapter 3 focuses on the domain of video abstracting and gives an overview of related approaches. A focus is put on trailer generation systems, particularly the SVP project is described in detail, as this thesis draws from it.

The extension beyond this is described in chapters 4 and 5. While chapter 4 deals with the manual analysis of additional trailer genres and the design of the video abstracting application, chapter 5 describes the implementation of this application and new trailer generation rules. Chapter 6 discusses an evaluation and the outcome of this thesis, followed by a conclusion and final thoughts in the last chapter.

Chapter 2

Theoretical Aspects of Movies and Trailers

An introduction to the understanding of movies has to start with an overview of their basic structures. Afterwards the elements which are used to tell the story are introduced in section 2.1.1. An overview of the term *genre* is given in section 2.1.8. In the following sections, the historical development is presented and the two main trailer types identified by Hediger [2001] are described. The chapter concludes with a short description of the "theatrical trailer" and an overview of the requirements and possibilities for automatic trailer generation.

In this excursus an overview is presented, since giving an exhaustive view on the features of film is not the major concern of this thesis. Thus the respective sections largely follow selected books.

2.1 Movies

The structure of this section is set up to be similar to the process of movie production. First the development of screenplays is described, followed by background information about the planning, shooting and editing of movies. Finally, principles for differentiating and organizing different kinds of movies by means of genres are explained.

2.1.1 Narrative Structure

According to Hediger [2001, p. 33] almost every contemporary movie has a narrative structure based on the paradigm by Field [1984].

Following the traditional scheme of Aristotle's *Poetics*, the story of one subject, the hero, is told [Hediger, 2001, p. 33], [Field, 1984, p. 14]. Grant [2007] explains that the narrative structure is often similar in movies, while the hero is usually a

unique one. Furthermore he identifies two narrative arcs within the plot of a film. A primary dramatic arc in which the hero has to overcome obstacles to achieve a goal and a secondary romantic arc which provides the narrative closure. [Grant, 2007, p. 15f]

The structure described by Field [1984] is illustrated in figure 2.1 and divided into three acts:

- ACT I: Beginning — setup of the location, characters, relationships
- ACT II: Middle — confrontation, introduction of the conflict/thread
- ACT III: End — resolution of the conflict

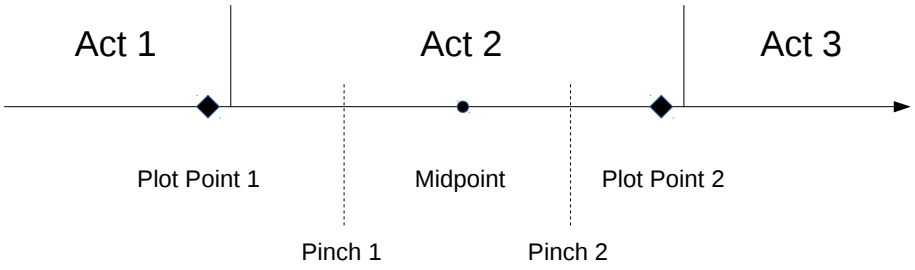

Figure 2.1 The Paradigm by Field, based on Field [1984] and Soto

According to guidelines by Field "one written page of screenplay equals one minute of screen time" [Field, 1984, p. 27], while a Hollywood movie is approximately 120 pages long. These divide into roughly 30 pages for act one, followed by the main part of 60 pages for act two and conclude with another 30 pages for act three. [Field, 1984, p. 27f]

In the first act, called *exposition*, the main characters of the story are introduced along with the setting and location (example: a group of cowboys in a small town in the Wild West). The relations between the characters are explained as well. The purpose of the first act is to "establish the dramatic premise" and to "build and expand the information of the story". [Field, 1984, p. 28]

At the end of the first act a first *plot point* occurs. A plot point can be an event or incident which changes the status quo and leads over to the second act. The second act, the *confrontation*, deals with the consequences resulting from the events of the first plot point. The hero or heroine is challenged by a threat and has to overcome all obstacles in his way to success (or sometimes failure). Near the end of the second act, another plot point leads over to the third act. In the third act, referred to as

resolution, the story is resolved and questions that have emerged during the first two acts are answered. [Field, 1984, p. 30ff]

The second act can be further divided into two parts. These are connected by the *midpoint* which is located around page 60 of a screenplay and can be an event which links the first and second half of act two together. In this context Field coins the term *pinch* which describes important narrative points in the first and second half of act two. They should be placed around pages 45 and 75. [Field, 1984, p. 131ff, p. 155ff]

Following Field's paradigm, it can be assumed that key events happen at the plot points, the midpoint and the pinches. As one page of a written screenplay corresponds roughly to one minute of film, it can be inferred that approximately at the minutes 30, 45, 60, 75 and 90 interesting story points and turns in the narration can be found. [Field, 1984, p. 27]

2.1.2 Narration

While early movies concentrated on scenic places or important events, Bordwell and Thompson say that first staged scenes were shot around 1903. Although *Cinderella* by Georges Méliès, an early narrative movie, had been realized in 1899, it wasn't until 1904 that a narrative form became a widespread technique in film making. These narratively constructed films consist of chains of events which are causally related to each other and take place in time and space. Narrative form is mostly used in fictional films but also in documentaries. [Bordwell and Thompson, 2010, p. 78, p. 456f, p. 493]

Narrative films go along with certain expectations by the audience: There will be characters and some interaction between them, incidents and a conflict or problem which will "reach a final state" [Bordwell and Thompson, 2010, p. 78] in the end. A narrative can be considered "to be a chain of events linked by cause and effect, and occurring in time and space" [Bordwell and Thompson, 2010, p. 79]. The events lead from one situation to another through a "series of changes" [Bordwell and Thompson, 2010, p. 79].

Important terms of narration are *story* and *plot* [Bordwell and Thompson, 2010, p. 80],[Chatman, 1980, p. 43f]. While a story covers all the events, either directly shown or inferred through existing cultural and expectations, the plot describes only directly perceivable visible and audible content. Bordwell and Thompson state "the plot will arrange story chronology so as to present the cause–effect chain most strikingly" [Bordwell and Thompson, 2010, p. 103].

In the classical Hollywood cinema most action is caused by individual characters

who Bordwell and Thompson describe as "causal agents". According to them, a narrative is focused on "personal psychological causes: decisions, choices and traits of character" [Bordwell and Thompson, 2010, p. 79]. Usually, such a narrative is started by the *desire* of a character which results in a *goal* to be achieved at the end of the narrative. A common practice in film making is to end a movie with "a strong degree of closure" [Bordwell and Thompson, 2010, p. 103]. Some contemporary films continue after the end credits by showing a final scene — the so called "credit cookie"— to undo closure and provide a chance for sequels. [Bordwell and Thompson, 2010, p. 79, p. 99, p. 102, p. 103]

2.1.3 Characters and Events

Typically, a Hollywood film features seven to eight significant *characters* who function as protagonists that formulate "clear-cut, long range goals" causing conflict among them [Bordwell and Thompson, 2010, p. 406].

Most of these characters have certain characteristics as pointed out by Bordwell and Thompson [2010, p. 397]:

> "classical Hollywood cinema often constructs a narrative around characters with definite traits who want to achieve specific goals. The clash of these characters' contrasting traits and conflicting goals propels the story forward in a step-by-step process of cause and effect".

Most characters' traits are tailored to achieve the desired effects in the narrative and are often portrayed in an exaggerated way compared to real life. In a narrative the characters deal with *events* which they cause and to which they react. [Bordwell and Thompson, 2010, p. 82f]

Not all events have to be shown explicitly. The audience is accustomed to naturally infer circumstances and make assumptions about implicit events. [Bordwell and Thompson, 2010, p. 80]

2.1.4 Expectations and Conventions

While watching movies, the spectators "urge for form" [Bordwell and Thompson, 2010, p. 58] as "the system of relationships within the work has not yet been completed" [Bordwell and Thompson, 2010, p. 58]. According to Bordwell and Thompson viewers have a demand for developing and completing patterns presented in narratives. People are accustomed to immediately form *expectations* of the things

to come: Will she "meet another character or arrive at her destination"? [Bordwell and Thompson, 2010, p. 58f]

The creators of movies are seeking to engage the audience in the narrative by making them curious about what might happen next. As expectations arise, the narrative will need to surprise the audience to keep them engaged by realigning their expectations continually. Further examples for expectations given by Bordwell and Thompson are a mystery which will offer a solution while reading it, "usually at the end", and "that the main character introduced in the first half of a film will be present in the second half" as well. [Bordwell and Thompson, 2010, p. 59f]

Such expectations are supported by patterns often used in narration (see chapter 2.1.6) such as the "journey pattern" in *The Wizard of Oz* (1939, by Victor Fleming), *Stagecoach* (1939, by John Ford) and *North by Northwest* (1959, by Alfred Hitchcock). *North by Northwest* also features the "search pattern" and romance plots. [Bordwell and Thompson, 2010, p. 61, p. 400].

2.1.5 Emotions

With regard to emotional response of the audience *suspense*, *surprise* and *curiosity* are named by Bordwell and Thompson [2010, p. 59f, p. 62]. They describe *suspense* as involving "a delay in fulfilling an established expectation" and as "leaving something suspended", like the following "elements in a pattern" and the "urge for completion" [Bordwell and Thompson, 2010, p. 59]. *Surprise* is defined by Bordwell and Thompson as "a result of an expectation that is revealed to be incorrect" [Bordwell and Thompson, 2010, p. 59]. *Curiosity* is the "ability of the spectator to wonder about prior events" [Bordwell and Thompson, 2010, p. 59f], like why a character is going somewhere.

Examples of emotions in movies given by Bordwell and Thompson are anxiety or sympathy spurred by suspense, satisfaction and relief vivified by gratified expectations. Cheated expectations and curiosity may puzzle the viewer or cause keener interest. [Bordwell and Thompson, 2010, p. 62]

A classical example of suspense presented by Bordwell and Thompson is a climax scene from *North by Northwest*: "Eve is dangling over the edge while Roger is clutching one of her hands and Leonard grinds his foot into Roger's other hand. It is a classic, not to say clichéd, situation of suspense" [Bordwell and Thompson, 2010, p. 404].

2.1.6 Narrative Form

The narrative form of movies consists of three parts: *opening*, *development* and *closing*. Most films begin either *in medias res* or with an *exposition*, according to Bordwell and Thompson. The first named form of a beginning starts directly with an action and only explains little background while an exposition introduces the characters, relations and locations first. Such an exposition explains background information and is a narrative convention. Often a variety of feasible causes and corresponding effects for what is shown are presented in the first quarter of a movie in order to raise expectations. This is often called the *setup*. [Bordwell and Thompson, 2010, p. 60, p. 90].

What normally follows is the development of the plot — changing the characters' situation by cause and effect in certain ways — which can be categorized in common as *plot patterns*, as mentioned before. *Change in knowledge* is "the most common general pattern" according to Bordwell and Thompson. In such a plot something new is learned by a character, with the key knowledge being revealed at the final turning point. Furthermore an often used plot pattern is *goal-oriented* in which the protagonist has to achieve a goal or reach a desired situation by taking certain steps. [Bordwell and Thompson, 2010, p. 91]

Other typical plot patterns are *searches* and *investigations*, which Bordwell and Thompson describe as "instances of the goal plot" [Bordwell and Thompson, 2010, p. 91]. Examples for movies incorporating the search pattern are *Raiders of the Lost Ark* (1981, by Steven Spielberg), in which the protagonist is looking for the Ark of the Covenant, *Le Million* (1931, by René Clair), in which a missing lottery ticket is being searched for, and *North by Northwest*, in which Roger Thornhill is trying to find George Kaplan. The investigation pattern is very common in detective films, in which the characters look for information instead of an object. [Bordwell and Thompson, 2010, p. 91]

Other plot patterns are time or space oriented. Some films make use of flashbacks to show the events leading to the current situation, for example in *The Usual Suspects* (1995, by Bryan Singer). Other plots deal with a *deadline*, a certain moment in time that needs to be met — like synchronizing a time machine and lighting to get back to the present in *Back to the Future* (1985, by Robert Zemeckis). Space oriented plot patterns focus the action on a single location, like a home (*Long Day's Journey into Night*, 1962, by Sidney Lumet) or a train (*The Tall Target*, 1951, by Anthony Mann). It is also quite common to combine different plot patterns. Many films following the journey pattern use the deadline pattern as well (for example *The Wizard of Oz* and *North by Northwest*), investigations making use of flashbacks (like in *The Usual*

Suspects).[Bordwell and Thompson, 2010, p. 91]

Each plot pattern raises specific expectations of the audience which will become engaged during the course of the film. Some genres make use of "cheating of an expectation" in order to create surprise and "delay an expected outcome" to raise suspense. Cheating of expectations is extensively used in comedy. As Bordwell and Thompson point out, *genres* "depend heavily on expectations". [Bordwell and Thompson, 2010, p. 59f, p. 91f].

Similarly to the opening, the end of a movie is not simply a stop. In most films "the ending resolves, or closes off, the chains of cause and effect" [Bordwell and Thompson, 2010, p. 92]. While the development usually concentrates at a peak with the highest suspense or tension, the *climax*, the ending brings the causal issues developed during the plot to a resolution. The climax is typically the point when the plot seems to have only a few possible closures which let the audience expect only one specific outcome. Such a closure of the narrative goes along with emotional satisfaction of the viewer. [Bordwell and Thompson, 2010, p. 92]

Some narratives do not provide closure but feature an open end instead. Such a plot does not resolve the outcome and consequences of the events in the narrative and asks the viewers to imagine the story's conclusion "or to reflect on other ways in which our expectations have been fulfilled". [Bordwell and Thompson, 2010, p. 92]

2.1.7 Shooting Film

The previous part focused on the theoretical and narrative parts of film making. In the following part technical and practical aspects are described. Bordwell and Thompson named *mise-en-scène*, *cinematography*, *editing* and *sound* as the essential parts of a movie's style [Bordwell and Thompson, 2010, p. 4].

2.1.7.1 Mise-en-Scène

Mise-en-scène — french for "putting into the scene" [Bordwell and Thompson, 2010, p. 118] — covers the arrangement of setting, actors and costumes as well as lighting and staging. It can be understood as staging for the camera, similar to theater. [Bordwell and Thompson, 2010, p. 121, p. 492]

Setting Although often serving as "a container for human events", sometimes the *setting* can become part of the narrative. Usually either existing locations — commonly used in early movies — or especially constructed sites and studios are used to film the action. Colors can be attributes to a setting as well and even support the

development of the narrative as in the change from "cold, steely colors" [Bordwell and Thompson, 2010, p. 123] to brighter, vivid colors in *Play Time*(1967) by Jacques Tati. It is not always required to have a full sized setting. Often miniatures and/or painted or photographed backgrounds were used to enhance the setting and to realize fantasy worlds. Nowadays, computer generated images are often used instead. [Bordwell and Thompson, 2010, p. 121, p. 123]

Sometimes objects from the setting serve a purpose in the storytelling. Such objects are called *property*, or *prop*, and are often used in comedies to serve humorous functions. [Bordwell and Thompson, 2010, p. 123]

Costumes may have certain functions in movies as well and can support narrative patterns and thematic developments by working in conjunction with the setting. Occasionally even to such an extent that director Erich von Stroheim (known for *Sunset Boulevard* (1950), *Blind Husbands* (1919)) was rumored to have his actors wear special underwear, e.g. in *Foolish Wives* (1922) which "would instill the proper mood" [Bordwell and Thompson, 2010, p. 125], although it was not shown in the final film. [Bordwell and Thompson, 2010, p. 128]

Sometimes costumes in films are quite stylized and colorful. For example in *Freak Orlando* (1981, by Ulrike Ottinger) highly intense primary colors are used. In some films costumes "play important motivic and causal roles", as Bordwell and Thompson [2010, p. 125, p. 128] point out. They give the example of dark glasses which the character Guido is wearing in *8½* (1963, by Federico Fellini) "to shield himself from the world" [Bordwell and Thompson, 2010, p. 125], or of different hats used to symbolize the change in a character, e.g., from a housewife to a reporter in *His Girl Friday* (1940, by Howard Hawks). [Bordwell and Thompson, 2010, p. 128]

Costumes often cooperate with the setting. While the setting provides a neutral background, the costumes help the audience "to pick out the characters" [Bordwell and Thompson, 2010, p. 126] — emphasizing the importance of color design [Bordwell and Thompson, 2010, p. 126].

The narrative progression of a film can be supported by setting and costumes. Bordwell and Thompson give the example of *Women in Love* (1969, by Ken Russel) which starts with "shallow middle-class life" [Bordwell and Thompson, 2010, p. 128] expressed by "saturated primary and complementary colors" [Bordwell and Thompson, 2010, p. 128]. Later on in the narrative less color is used "as the characters discover love on a country estate" [Bordwell and Thompson, 2010, p. 128], and in the end of the film have become almost black and white as the enthusiasm of the protagonists faded. [Bordwell and Thompson, 2010, p. 128]

Makeup is quite related to costumes so many of the aforementioned aspects apply to makeup as well. Initially makeup was required as it was difficult to capture the faces of actors well on the first film strips. Later the purpose was to improve the actor's appearance in the cinema. In some movies the use of makeup is avoided completely while others use it elaborately. For example, by adding a false beard, nose and eyebrows in *Ivan the Terrible* (Part 1, 1945, by Sergei Eisenstein). Actors of historical characters have often used makeup extensively. [Bordwell and Thompson, 2010, p. 128]

Nowadays makeup is mostly intended to be unnoticed but sometimes it's used, e.g., to highlight certain facial aspects or hide unwanted wrinkles. A common contemporary convention of makeup is for female actors to wear lipstick but male actors often utilize makeup as well. Besides aiding the look, makeup may also support certain character traits. Eyebrows, for example, are often expressively shaped, lengthened to enlarge a face or shortened to make it look more compact. "Thick, straight eyebrows, commonly applied to men, reinforce the impression of a hard, serious gaze". [Bordwell and Thompson, 2010, p. 128]

Especially in recent horror and science fiction films character traits and story events are supported by "bumps, bulges, extra organs, and layers of artificial skin" [Bordwell and Thompson, 2010, p. 130].

Lighting is not just used to illuminate a scene but can provide guidance and let the viewers focus on specific parts of the image. Bright areas might highlight key events while dark or shadowed ones may hide certain details or create suspense. The result of lighting of objects are highlights and shadows. [Bordwell and Thompson, 2010, p. 131]

Bordwell and Thompson describe highlights as "patch of relative brightness on a surface", like an illuminated face or fingertips. Light may also provide hints for the texture of objects, like sparkling looks for smooth surfaces and diffuse ones for softer or rougher surfaces. [Bordwell and Thompson, 2010, p. 131]

Shadows can be either *attached* (also known as *shaded*) or *cast shadows*. Attached shadows are caused by the shape of an object, e.g., the shadow from a nose on a character's face, while cast shadows are caused by objects blocking the light. Lighting supports the viewers by sensing "a scene's space" [Bordwell and Thompson, 2010, p. 131]. Bordwell and Thompson give the examples of bright and dark stripes creating the illusion of a prison cell. [Bordwell and Thompson, 2010, p. 131].

> "Light is everything. It expresses ideology, emotion, colour, depth, style. It can efface, narrate, describe. With the right lighting, the ugliest face,

> the most idiotic expression can radiate with beauty or intelligence" *Frederico Fellini, director* cited in [Bordwell and Thompson, 2010, p. 131].

For the purpose of filming, Bordwell and Thompson name four features of lighting: *quality*, *direction*, *source* and *color*. The *quality* of illumination might be *hard*, providing sharp and distinct shadows with crisp edges. Or it might be *soft*, without distinctive edges and "diffuse illumination", like from a clouded sky. [Bordwell and Thompson, 2010, p. 132]

Bordwell and Thompson describe five different lighting *directions*:

- frontal lighting — reduction of shadows
- sidelighting/crosslighting — shaping character's features
- backlighting — creating silhouettes or contours
- underlighting — light from below, distorted look, often used for "horror effects", may also be used realistically, for example as fireplace
- top lighting — highlighting for example the cheekbones

While some movies, like documentaries, rely on the available natural light, almost all fictional movies use artificial lighting to have more control of the captured images. Most extra light sources support the existing lights of the setting and create a consistent illumination without being noticed. [Bordwell and Thompson, 2010, p. 132]

A common approach among filmmakers is to use a *key light* and a *fill light* for each subject. The key light is "providing the dominant illumination and casting the strongest shadows" [Bordwell and Thompson, 2010, p. 133]. The fill light is dimmer and eliminates or softens the key light's shadows. [Bordwell and Thompson, 2010, p. 133f]

In classical Hollywood cinema *three-point lighting* was a common illumination method. This method relies on three light sources for each major person. The two previously described key and fill lights are supplemented by a *backlight* which is located above and behind the character. In this setup the *key light* illuminates the figure frontally and the *fill light* is located near the camera. Such artificial lighting may require the light sources to be rearranged when the camera's position changes. This helps to compose the shot, although constantly changing light sources are unlikely in reality. [Bordwell and Thompson, 2010, p. 134f]

Such a lighting design is referred to as *high-key lighting*. This "overall approach to illumination" results in diffuse low contrast light with fairly transparent shadows. It is used by filmmakers in comedies, adventure films and almost every drama. [Bordwell and Thompson, 2010, p. 135]

In comparison *low-key illumination* rarely uses a fill light and thus tends to result in more contrast and "sharper, darker shadows" [Bordwell and Thompson, 2010, p. 136]. Because of such properties, low-key lighting is often used in somber or mysterious scenes, occurring in horror films and film noir. [Bordwell and Thompson, 2010, p. 136]

Most of the time filmmakers try to use pure white lighting. However sometimes color filters are used, for example a slight tint of orange in order to simulate the light of a candle. Other movies feature colored light to emphasize a character's mood, for example blue light for uncertainty and terror in *Ivan the Terrible*, Part 2. [Bordwell and Thompson, 2010, p. 136]

Staging denotes the behavior of figures through movement and performance. A *figure* in a movie can express feelings and thoughts and might be a person, an animal, a robot or just a shape [Bordwell and Thompson, 2010, p. 138].

One common technique bringing such figures to life is the animation of puppets via *stop-motion* or *stop-action*. Each frame of the movie is shot individually and the puppets are moved in-between. Other techniques for non-human figures are animated cartoons and computer-generated images. These recently developed digital techniques are often combined with motion-capturing to get realistic figure movement. [Bordwell and Thompson, 2010, p. 139]

In spite of this, most figures are still portrayed by human actors and thus needed to be performed. Bordwell and Thompson differentiate performances into visual and audible elements. Among the visual elements are appearance, gestures and facial expressions. The audible ones contain voice and effects. In some movies only visual or audible performances might be perceivable. For example, in silent movies or through an invisible narrator. [Bordwell and Thompson, 2010, p. 139]

An actor is not just reciting his or her lines from the storyboard in a believable manner but is at all times part of the visual concept. Therefore he or she has to play his or her character even when he or she is not speaking. [Bordwell and Thompson, 2010, p. 141]

The most important tools for conveying a character are facial expressions, in particular during the era of silent movies. These movies became quite popular all over the world since simple facial expressions, such as happiness, fear and anger, are comprehensible in different cultures [Bordwell and Thompson, 2010, p. 141]. Contemporary movies use close-up shots intensively which require precise control over facial expressions by the actors. Most important for expressive facial features for acting are the mouth, the eyes and the eyebrows. Together they communicate

the reactions of a character to dramatic events to the audience. [Bordwell and Thompson, 2010, p. 141]

The eyes of the actors are of special importance. They provide information by line of sight, the utilization of the eyelids and the eyebrows' shape. Bordwell and Thompson note that while in everyday life the eyes wander around about 50% of the time during a conversation, actors have to look constantly at each other and try not to blink. On one hand looking away would suggest distraction or avoidance. On the other hand blinking, as a reaction to something, may be interpreted as a sign of anxiety or surprise. [Bordwell and Thompson, 2010, p. 141]

Besides the face of an actor the body is often visible as well and thus important for the performance. Attitude and personality of a character can be communicated by their posture — how they walk, stand or sit. Especially in silent movies the hands and gestures were also important tools for acting. According to Bordwell and Thompson the "hands are to the body what eyes are to the face: They focus our attention and evoke the character's thoughts and feelings." [Bordwell and Thompson, 2010, p. 141f]

Acting goes along with the question of the performance's realism. This depends on the audience's comprehension of what a realistic performance is. This has changed over time. Additionally not every film aims to be realistic but features a specific style of acting. Contemporary examples for exaggerated, non-realistic acting can be found among others in mass-productions from Hollywood, India and Hong Kong. Bordwell and Thompson name martial-arts films with Jet Li or Jackie Chan and comedy films with Jim Carrey as examples, in which non-realistic performances are expected by the audience. [Bordwell and Thompson, 2010, p. 139f].

Typically performances are to some degree both *individualized* and *stylized*. In the case of a protagonist the audience expects a unique character which is neither over- nor underplayed. An example for a rather individualized character is *Don Vito Corleone* from *The Godfather* (1977, by Francis Ford Coppola) who has "a complex psychology, a distinctive appearance and voice, and a string of facial expressions and gestures that make him significantly different from the standard image of a gang boss" [Bordwell and Thompson, 2010, p. 140]. The stylization of the performance is described as middle-ranged, neither flat nor too exaggerated. [Bordwell and Thompson, 2010, p. 140]

Besides individualized characters, anonymous *types* became common in classical Hollywood cinema — typically stereotypes like "the Irish cop on the beat, the black servant, the Jewish pawnbroker, the wisecracking waitress or showgirl" [Bordwell and Thompson, 2010, p. 140]. By *typecasting*, actors were chosen to match a desired type. Actors who played characters belonging to a typical social class or movement

were chosen through similar system called *typage*. [Bordwell and Thompson, 2010, p. 140]

A performance can be combined out of several different shots, and each one can be taken several times. Thus the best played takes of a shot can be selected, providing the filmmaker a lot of choices. The performance can be additionally enhanced through editing. When a close-up shot of widened eyes is followed by an image of a gun, the expression of fear is enhanced. [Bordwell and Thompson, 2010, p. 145]

Another difference to traditional acting in theaters can be noted. While the distance between the audience and the actors in the movie theater is limited by the size of the theater itself, the proximity can be very small due to the camera that can make objects appear bigger and vice versa. This results in a need for different ways of acting depending on the camera's distance. Small figures need large gestures while even tiny movements may suffice in a close-up shot. [Bordwell and Thompson, 2010, p. 145]

Space & Time Focusing the attention of the audience on certain crucial parts of the image is a basic task in filmmaking. Among other purposes, e.g. causing emotions, suspense and curiosity, this is realized through mise-en-scène by utilizing costumes, lighting, setting and figures. These elements allow the filmmaker to create a three-dimensional space[1] for staging the plot. [Bordwell and Thompson, 2010, p. 147]

Bordwell and Thompson compare a film shot to a painting, as being a "flat array of colors and shapes" [Bordwell and Thompson, 2010, p. 148]. In most shots, filmmakers try to achieve balance and to evenly place points of interest, although it is also quite common to balance only the left and right parts of the shot. Filmmakers assume that the viewers focus on the top half of the image, since this is where faces of characters would be expected. [Bordwell and Thompson, 2010, p. 148f]

> "The audience is only going to look at the most overriding thing in the frame. You must take charge of and direct their attention. It's also the principle of magic: what is the single most important thing? Make it easy for them to see it, and you're doing your job."
>
> *David Mamet, director* cited in [Bordwell and Thompson, 2010, p. 148]

But not all shots are balanced, sometimes frames are unbalanced on purpose to accomplish certain effects. An example described by Bordwell and Thompson is an

[1] Most films use a three-dimensional space, however cartoons for example are often limited to two dimensions [Bordwell and Thompson, 2010, p. 147].

unbalanced image with a doorway to suggest that new characters may enter a scene. [Bordwell and Thompson, 2010, p. 149]

Another way to guide attention is by using contrast. If the film is black and white, bright and light areas, often costumes or faces, attract more attention than darker regions. In colored films, warm colors like red, orange or yellow stand out, and cool colors like green and purple appear not so prominent. Some movies use dominating color tones which allow for strong contrast by using a completely different color. [Bordwell and Thompson, 2010, p. 150]

Similar to contrast, motion can guide the viewers' attention. A single moving element in a film may catch attention when the rest is not moving. Similarly an object moving with a different speed than the rest of the image will stand out. [Bordwell and Thompson, 2010, p. 151]

Although a movie is normally shown on a two-dimensional screen the audience immediately perceives a three-dimensional space through so called *depth cues*. All elements of mise-en-scène, i.e., lighting, setting, costumes and staging, may help to provide such depth cues. A space as described here has a *volume* — it is solid and takes up room — as well as multiple *planes* — layers in which figures and objects are arranged in the scene. Volume in a film is indicated by aspects like movement, shape and shading. Bordwell and Thompson term the different planes in a scene, according to their distance to the camera, *foreground*, *middle ground* or *background*. [Bordwell and Thompson, 2010, p. 151f]

Apart from a blank screen, which only has a single plane, a frame has at least two planes. One is the background and the other one(s) hold objects in the foreground — operating by virtue of the so called *overlap* depth cue. Because these objects block the view onto the background plane, they are perceived as being closer to the camera. [Bordwell and Thompson, 2010, p. 152]

Another depth cue for overlapping is different color palettes. As colors become cool and pale with growing distance, these are often chosen for background planes, for example for settings. "Warm or saturated colors tend to come forward" [Bordwell and Thompson, 2010, p. 152], so they are often used for foreground bits such as costumes. However, small differences in color contrasts may sometimes help perceiving a space as three-dimensional as well. Another very important depth cue, implying planes and volume at the same time, is *movement*. [Bordwell and Thompson, 2010, p. 152]

The human visual system is used to assign sharp outlines, clear textures and pure colors with closer objects. More distant objects appear blurred and grayed due to atmospheric haze. Such a depth cue is called *aerial perspective* and is often combined with specific lighting and blurring of background planes through lens focus. Along

with depth comes *size diminution*. This describes the effect that objects which are distant to the camera appear smaller than closer objects. Size diminution thus helps in perceiving a deep space with greater separation of the planes. Through utilizing a *linear perspective*, a huge depth can be suggested by convergence of parallel lines in a point far away. This point may be either *central* or *off-center*. [Bordwell and Thompson, 2010, p. 152f]

Regardless of the actual techniques used, different kinds of shot compositions can be compared: *shallow-space* and *deep-space*. As the names suggest the planes are either quite close or very distant from each other. These extremes are quite rare, most shot compositions are in between them. [Bordwell and Thompson, 2010, p. 154]

Besides space, *time* is an important factor in mise-en-scène as well. Bordwell and Thompson point out that mise-en-scène is "not only *what* we see but *when* we see it and for how long" [Bordwell and Thompson, 2010, p. 156]. As humans perceive even slight motions quite easily, the audiences' attention can be focused even by very small or slow movements. Furthermore some filmmakers take advantage of the human sense for rhythm. Even slight changes in routine tasks, like the preparation of a meal carried out in the same manner every day, can be noticed by the audience and communicate emotional pressure. [Bordwell and Thompson, 2010, p. 156]

As a new shot starts, the audience will quickly get a first impression of the whole spatial arrangement which builds up expectations. After this initial assessment, the viewers will intuitively look for further cues, especially for the existence of movement. Such cues are time-bound, as Bordwell and Thompson point out, as different kinds of movement heavily depend on timing. The scanning of the frame follows the depth of a composition as well. Some deep-space shots utilize this by placing an event in the background which stirs up expectations about what will go on in the foreground. A lot of attention is gained when an object or figure moves from the background to the foreground. [Bordwell and Thompson, 2010, p. 157]

The placement of actors also guides attention, as normally the audience supposes that a character's face will convey more story information than his or her back. Bordwell and Thompson call this the "power of *frontality*" [Bordwell and Thompson, 2010, p. 158]. Strengthening this principle, viewers will focus on characters turned towards them and neglect those not facing them. [Bordwell and Thompson, 2010, p. 158]

2.1.7.2 Cinematography

The term *cinematography* describes *how* things are filmed. According to Bordwell and Thompson three different fields are covered by cinematography: the process of

photography, the framing of the shot and its duration. [Bordwell and Thompson, 2010, p. 167]

Photography Photography covers several aspects: *tonalities*, *speed of motion* and *perspective* [Bordwell and Thompson, 2010, p. 167].

- **Tonalities** A film may have strong or weak colors, it can be grayish or entirely black and white. Depending on the chemical configuration a film stock may have stronger or weaker *contrast*. A more light sensitive, faster film will in general lead to a "low-contrast" image while a slow film with less light sensitivity yields to higher contrasts. In color movies, certain specific film stock like "Technicolor became famous for its sharply distinct, heavily saturated hues". [Bordwell and Thompson, 2010, p. 167f]

  Apart from using different film stocks the tonalities can also be influenced in laboratory processes and later by computer grading. *Tinting* and *toning* are techniques used to either boost brighter or darker parts of the image in certain colors. Conventional colors used here are, for example, blue for night scenes, red for scenes showing fires and amber for indoor scenes. Some old films were even *hand colored*, meaning that picture by picture was colored manually. [Bordwell and Thompson, 2010, p. 169f]

  Tonalities in film also depend on the *exposure* during the shooting of the film. That means controlling the amount of light which is getting onto the film through the optical system of the camera. Over- and underexposure may be used to sharpen contrasts between indoor and outdoor shots and for other effects. Sometimes *filters* are used to limit specific colors being captured. This allows the filmmaker to shape the range of tonalities even more. An example presented by Bordwell and Thompson is *day for night*, a technique which allows night scenes to be shot during the day. [Bordwell and Thompson, 2010, p. 170f]

- **Speed of Motion** The pace of the events in front of the camera can not only be controlled by actors but also by filming techniques. By changing the number of frames recorded per second in relation to those played back later in the cinema, the motion can be slowed down or sped up. This yields *slow-motion* or *fast-motion* effects. Slow-motion is often used to show enormous power, to emphasize or to indicate a dream or fantasy. Some comedies utilize fast-motion for humorous purposes. [Bordwell and Thompson, 2010, p. 172]

  Another common manipulation of time is the *time-lapse* which shortens longer processes dramatically. A sunset may be shown in a few seconds or "a flower

sprout, bud, and bloom in a minute" [Bordwell and Thompson, 2010, p. 173] by taking for example one frame per minute. *High-speed* cameras allow for slowing down extremely short events like shattering glass by shooting several hundreds or thousands of frames per second. [Bordwell and Thompson, 2010, p. 173]

- **Perspective** Some *perspective* relations were described before (see chapter 2.1.7.1, such as diminution which describes the shrinking size of an object if it is moving farther away from the viewer. A camera works in a similar fashion as the human eye except for allowing the use of different *lenses* and therefore varying perspectives. [Bordwell and Thompson, 2010, p. 173f]

  Bordwell and Thompson differentiate three general kinds of *lenses* according to their *focal length*[2]: the *short-focal-length* or wide-angle lens, the *middle-focal-length* or medium lens and the *long-focal-length* or telephoto lens. [Bordwell and Thompson, 2010, p. 174]

  A typical wide-angle lens has a focal-length of usually less than 35mm and often causes distortion of straight lines around the frame border. Another property of such lenses is the exaggeration of depth which makes the shot composition look deeper and motions backward or forward seem faster. [Bordwell and Thompson, 2010, p. 174]

  The focal length of a medium lens is normally around 50mm. Such a lens avoids noticeable distortions by providing straight and perpendicular lines leading to a correct naturalistic perspective. The perceived depth of a scene composition should be in between the wide-angle and the telephoto lens. [Bordwell and Thompson, 2010, p. 174]

  With its length of 100mm or more a telephoto lens is much longer than the previously described lenses. These lenses have a flattening effect on a shot by reducing the "cues for depth and volume" [Bordwell and Thompson, 2010, p. 174] which also makes movement towards the camera seem slower. Often such lenses are used for sports coverage where the photographer may not be close to the objects to be captured. [Bordwell and Thompson, 2010, p. 17ff]

  Apart from lenses with a fixed focal length, a *zoom lens* can be used to adjust the focal length during shooting. Until the late 1950s it was common to zoom before starting the actual capture. Sometimes the zoom is used to simulate

[2] The focal length is the distance between the middle of the lens and the actual film strip [Bordwell and Thompson, 2010, p. 174].

changes of the camera position without actually moving it back or forth. This may transform scale and depth. [Bordwell and Thompson, 2010, p. 176f]

Besides its focal length an important property of a lens is its *depth of field*. This describes the range of distance between the lens and objects which allows for capturing images in sharp *focus*. Normally the depth of field from a wide-angle lens is greater than the one of a telephoto lens. A lens with a smaller depth of field allows the filmmaker to have only a certain plane captured sharply — a technique called *selective focus*. The opposite is desired when using *deep focus*. By applying this technique, all objects are in focus, no matter their distance. Another technique is called *racking* or *pulling focus*. Here the focal point is changed during the duration of the shot. For example, starting with a focused plane in the foreground and ending with a different one in the background in order to shift the attention of the audience. [Bordwell and Thompson, 2010, p. 177f]

Apart from *special effects* used in mise-en-scène, like models or computer-generated images, cinematography allows for other special effects as well. Such special effects may be the combination of two or more separately shot planes of action through *superimposition*. This is done by exposing the film multiple times. Other more advanced techniques allow for the combination of multiple film strips in *process* or *composite shots* through *projection process work* and *matte process work*. [Bordwell and Thompson, 2010, p. 179f]

The basic principle of projection process work is the combination of a setting projection and actors performing in front of this screen. This is done to avoid bringing the whole film team to location [Bordwell and Thompson, 2010, p. 180].

The *rear projection* technique was used to add background images to shots. This was done by projecting the image from behind a translucent screen. A major drawback of this technique is that it lacks credible depth cues. [Bordwell and Thompson, 2010, p. 179].

Front projection was introduced in the late 1960s and helped to avoid the issues of rear projection by providing sharper focus. The camera films the actors in front of a high-reflective screen on which the background image is projected. A two-way mirror allows the projected background image to originate from the exact spot from where the camera captures the whole set-up. Nowadays a blue or green screen replaces the projected image. Later this respective single-colored screen is replaced with the desired background by digital techniques.

[Bordwell and Thompson, 2010, p. 181]

Another common way to compose an image out of different sources is to combine several shots by *matte work*. A part of the setting is captured or even painted on the matte while areas on the film strip are left blank. During the laboratory work another shot in which the actors were captured is added to compose the final sequence. [Bordwell and Thompson, 2010, p. 181]

By using *traveling matte* actors can not only be overlaid onto but also integrated into the background footage. Mattes paintings were used in most mainstream movies until the late 1990s and then supplemented and replaced by digital imagery. [Bordwell and Thompson, 2010, p. 181f]

Some of the special effects are not only part of cinematography but also require preparation during shooting. This makes them part of mise-en-scène as well. Especially the success of digital effects blurred the border between mise-en-scène and cinematography. [Bordwell and Thompson, 2010, p. 182f]

Framing Through framing, the image that the audience will see is defined by setting the *vantage point*. Here the camera is located with respect to the action. Bordwell and Thompson say that framing helps the filmmaker "to transform everyday reality into cinematic events" [Bordwell and Thompson, 2010, p. 186]. This is done by the frame's *shape and size*, how the frame dictates *on- and offscreen space*, the camera's *angle, level, height and distance* and the frame's *mobility*. [Bordwell and Thompson, 2010, p. 186]

- **Dimensions and Shape** As the actual size of the film's projected image depends on the cinema's screen, the *aspect ratio* is the most important dimensional property of films. It describes the relation of the image's width to its height. There are many standard aspect ratios. The most important classical ones are 1.33:1 for early movies and later on the *Academy ratio* of 1.37:1. Besides these formats, a collection of *widescreen* aspect ratios has been established. Among these 1.85:1 is the most common in North America and 1.66:1 the most common in Europe. In the 1950s the *CinemaScope* system was introduced. It uses *anamorphic* optics, both in production and screening, to store a horizontally squeezed image on the film reel. Films shot in CinemaScope had an aspect ratio of 2.35:1 until the 1970s, when it was changed to 2.40:1 [Bordwell and Thompson, 2010, p. 189]. The widescreen format is especially suitable for horizontal shot compositions. That is why it was initially used in

genres like Westerns, travelogues, musicals and historical epics which feature sweeping settings. [Bordwell and Thompson, 2010, p. 187ff].

Although most movies have a rectangular format *masks* may be used in some shots to achieve a different form. One example given by Bordwell and Thompson is the use of an opening and closing *iris* which leaves a round image as a transition between shots. A different kind of mask results in vertical slices, representing an actor's view through a fence. Among other common effects is the *multi-frame* or *split-screen*. Here multiple shots are combined in a single frame. This is often used for telephone conversations to show both participants at the same time or to build up suspense as several events can be seen simultaneously to "gain a godlike omniscience" [Bordwell and Thompson, 2010, p. 191].

- **Onscreen and Offscreen** By capturing a shot with a camera, two areas are created the *onscreen* and the *offscreen* space of the setting. While the onscreen space covers everything currently visible in the frame Bordwell and Thompson cite film aesthetician Noël Burch describing the offscreen space as consisting of six different zones: the area following each of the four frame borders, the zone behind the action and the background behind the camera or the audience respectively. [Bordwell and Thompson, 2010, p. 191]

  Actors may interact with figures or objects in the offscreen space by pointing at or looking in an offscreen zone. Sound originating in the offscreen space may also give clues about what is happening there. Offscreen space is also used for surprising effects. For example a hand reaching abruptly into the frame and indicating the presence of another person. [Bordwell and Thompson, 2010, p. 191f]

- **Angle, Level, Height and Distance** Besides defining the offsceen space the frame also positions the perspective from which the action is shown. Filmmakers distinguish three common *angles* in which the camera may capture the action: the most common *straight-on angle*, the *high angle* with a down-looking camera and the *low angle*, where the camera is pointing upwards to the action. [Bordwell and Thompson, 2010, p. 193f]

  Most shots are oriented horizontally, although *canted* frames are sometimes used to symbolize that the "world is out of kilter" [Bordwell and Thompson, 2010, p. 194f]

  The *height* from which a frame is captured relates partly to the angle. The reason for this is that a high angle requires the vantage point to be above the

material to be filmed. Some movies may also use a specific height for purposes of visual style. [Bordwell and Thompson, 2010, p. 194]

The *camera distance* of a shot is normally measured by using the human body as a standard. Bordwell and Thompson describe the following common shot distances: [Bordwell and Thompson, 2010, p. 195]

- extreme long shot — landscapes, bird's-eye view
- long shot — background dominates
- medium long shot — humans framed from the knees up
- medium shot — humans framed from the waist up
- medium close-up — humans framed from the chest up
- close-up — showing only the face, hands, feet or similar sized objects
- extreme close-up — portion of a face or object

Although there is no exact definition of these shot distances, these terms are clear enough for use throughout the whole movie industry. [Bordwell and Thompson, 2010, p. 196]

- **Functions** All of these qualities of framing have no absolute meaning, their function depends on the individual movie and the context of their usage. In some situations a character seems powerful when he or she is filmed from a low angle and "dwarfed and defeated" [Bordwell and Thompson, 2010, p. 196] when filmed from above. However this is no general rule and depend on the context of the narrative. [Bordwell and Thompson, 2010, p. 196]

  Nevertheless, the distance of the camera often has the function of establishing and reestablishing the setting and the location of characters. A *point-of-view(POV)* shot simulates the view through the eyes of a character and implies that the world of the movie is seen in the same way as the character does. Certain details that would not be noticed otherwise can be brought to the viewer's attention by using close ups. Framing can also yield comic effects and pictorial jokes. [Bordwell and Thompson, 2010, p. 196ff]

- **Mobile Frame** Compared to other pictorial kinds of art, film has one specific property that others do not: the ability to *move* the frame in relation to the material *being* filmed. The types of mobile framing can be grouped into *pan*, *tilt*, *tracking* or *dolly shots* and *crane shots*. [Bordwell and Thompson, 2010, p. 199]

  In a *pan* shot the camera is rotated around its vertical axis without changing its position. This gives an impression of a head looking left or right to the

audience. A *tilt* is a rotation of the camera around its horizontal axis. Both rotational movements unroll the scenes space either horizontally or vertically. [Bordwell and Thompson, 2010, p. 199]

In comparison a *tracking* or *dolly* shot describes a shot, in which the camera is moved on the ground. This may be forth or back, sideways, diagonally or even circular. A *crane* shot allows for the camera to be moved off the ground and in different directions. It may be achieved by using either a mechanical arm or filming from a helicopter or a plane. [Bordwell and Thompson, 2010, p. 199]

Another way of filming is by using body-mounted cameras which allow the simulation of a character's perspective. Such a camera may be either a *handheld* one, resulting in a shaky, documentary-like point-of-view image, or a stabilized *Steadicam*. This camera allows smooth movements and balanced shots. Finally, a different way to realize mobile framing is the utilization of a zoom lens. This method can provide mobility without physically moving the camera. [Bordwell and Thompson, 2010, p. 200, p. 202]

A common motivation for changing the camera's perspective is movement between the figures. This requires *reframing* in order to reestablish balance in the composition of the shot. Often such reframings are seldom noticed by the audience since they are motivated by the characters' movement. *Following shots* — as the name suggests — follow objects or figures: "a pan may keep a racing car centered, a tracking shot may follow a character from room to room, or a crane shot may pursue a rising balloon". [Bordwell and Thompson, 2010, p. 203]

But the camera may not only move in reaction to character movement. It is also quite common for the camera to back up and leave the characters when something important is to be disclosed. Otherwise unnoticed clues may be focused on by a moving camera or a setting may be introduced before characters enter. Camera movement independent from figures can also link different characters together. [Bordwell and Thompson, 2010, p. 203f, p. 210]

Besides such purposes a moving frame also consumes screen time and may either build up arcs of expectation and fulfillment by panning quickly or surprise the viewers by showing unexpected things. By moving slowly, a mobile frame may delay the fulfillment of the audience's expectation and thus support building up suspense. [Bordwell and Thompson, 2010, p. 205]

The mobile frame's velocity may also be rhythmic. Especially in musical films,

the camera movement's speed may accentuate characteristics of a singing or dancing performance. [Bordwell and Thompson, 2010, p. 205]

Long Take In early cinema (ca 1895–1905) the shots of a movie had a rather long duration and often the film consisted of only one shot. By 1916 *continuity editing* (see chapter 2.1.7.3) was established and thus shots became shorter. The average shot length of an American movie between the late 1910s and the early 1920s was around 5 seconds and grew with the emergence of sound films to about 10 seconds. [Bordwell and Thompson, 2010, p. 213]

Some directors however wanted to shoot films with a small number of longer shots, the so-called *long takes*. Typically, a scene is composed of a number of shorter shots but in such films a scene is often filmed in one long take which is then called a *sequence shot*. It is typical for a long take to be captured as medium or long shot. Such lengthy shots may have a formal pattern on their own. They may have their own development, trajectory and shape. [Bordwell and Thompson, 2010, p. 214, p. 216]

2.1.7.3 Editing

Early films before 1904 contained only a single shot as just noted. In contrast to these contemporary movies are made up of between 1000 and 3000 shots and thus require considerable work on editing to be organized. [Bordwell and Thompson, 2010, p. 223]

The main purpose of *editing* is "the coordination of one shot with the next" [Bordwell and Thompson, 2010, p. 223] and to exclude footage that is not desired. This is normally done by keeping only the best take of a shot. The transition between two shots may be varied. It can be a *cut*, the second shot immediately follows the first one. It may be a *fade-out*, the first shot is slowly darkened or a *fade-in*, the second shot is brightened from a dark frame. Or it can be a *dissolve*, the two shots are blended over each other. Another possible transition is a *wipe* in which the second shot is revealed by a moving boundary on the first shot without dissolving. Fades, dissolves or wipes are not perceived as instantaneous changes as a cut is. [Bordwell and Thompson, 2010, p. 223f]

Before the usage of computers in film editing, transitions were done in the laboratory during the production process by physically manipulating the film strip. Very rarely, films do not rely on cutting at all through carefully planning and filming in a manner ready to be shown. Nowadays editing is mostly done with digital footage on computers. [Bordwell and Thompson, 2010, p. 224]

Film Editing Dimensions According to Bordwell and Thompson, there are four different areas over which filmmakers have choice and control while editing. These are *graphic*, *rhythmic*, *spatial* and *temporal* relations between a shot A and a shot B. [Bordwell and Thompson, 2010, p. 225]

- **Graphic Relations** The pictorial relations between two adjacent shots may be any of the qualities of mise-en-scène and cinematography. These are lighting, setting, the acting of figures in space and time as well as photography, framing and the mobility of the camera. [Bordwell and Thompson, 2010, p. 226]

  While linking the shots the editor can work with similarity and difference to create smooth continuity and sudden contrast. Graphic similarities between two shots — such as shapes, colors, movements or the overall composition — are described by Bordwell and Thompson as *graphic match*. It is often desired by filmmakers to achieve a rough graphic continuity when linking shots. Explicit graphic matching is rather rare in narrative cinema. "The director will usually strive to keep the center of interest roughly constant across the cut, to maintain the overall lighting level, and to avoid strong color clashes from shot to shot" [Bordwell and Thompson, 2010, p. 226].

- **Rhythmic Relations** By choosing the length of an individual shot in comparison to its adjacent shots the editor is able to control the rhythmic relations of the composition. Although other cinema techniques contribute to the rhythm of a film as well, the pattern of shots' duration is mostly responsible for the audience's perception of a movie's rhythm. [Bordwell and Thompson, 2010, p. 230]

  A shot's duration can also be used to stress an action or to deaccentuate it. For example, some white frames can be added in order to indicate a violent impact. Furthermore a shot can be extended to grant the audience time to calm down its reactions after a spectacular event in the scene. Similarly, shorter shots can accelerate the pace of a scene while lengthened shots may slow it down. A slow change from longer to shorter shots, or acceleration of the pace, may start a tense sequence. A shift in tempo allows the filmmaker to influence how much time the audience has to perceive and think about what is shown in the film. For example a sequence of very short shots leaves only little time to become aware of what is happening. [Bordwell and Thompson, 2010, p. 230]

- **Spatial Relations** Apart from graphics and rhythm, editing also serves the purpose of composing film space. Through editing it is possible to pretend

that any points share some kind of connection, regardless of their location. [Bordwell and Thompson, 2010, p. 231]

A very common pattern to form spatial relations through editing is to begin "with a shot that establishes a spatial whole and follow this with a shot of a part of this space" [Bordwell and Thompson, 2010, p. 231]. An example for such an analytical breakdown might first show a medium long shot of a group of figures, and continue with a medium shot of a single figure. Another way may be to compose a whole space by combining shots showing several parts of the setting individually, and using assisting parts of mise-en-scène to construct a connection between them. [Bordwell and Thompson, 2010, p. 231]

Often shots captured in different locations are combined to show a setting that does not exist in the actual way it is presented. The audience will get the illusion that objects, figures or events, which are shown in consecutive shots, are related although they may be captured totally independent from each other. The phenomenon is called the *Kuleshov effect*. [Bordwell and Thompson, 2010, p. 231f]

Crosscutting is a technique used quite often to show different spaces and actions simultaneously through editing. [Bordwell and Thompson, 2010, p. 233]

- **Temporal Relations** A basic purpose of editing is to bring the events of the plot into an *order*. Most films use a chronological arrangement. Some movies utilize *flashbacks* to change that order and to show prior events or memories. Similarly a *flash-forward* can be used to show events which will happen in the future and "to tease the viewer with glimpses of the eventual outcome of the story action". [Bordwell and Thompson, 2010, p. 233]

  Besides the order of shots, the duration of a movie's action can be changed by *elliptical editing*. This is achieved by leaving out non-interesting footage. It may be realized by using classical transitions — like dissolves, wipes or fades — to symbolize that time has passed. Or the editor may cut from the beginning of a redundant action directly to its end. Bordwell and Thompson explained this giving the example of a man climbing stairs: Instead of following him the whole time he may be only shown on the bottom of the staircase, followed by a shot showing how he arrives at the top. To cover the extracted time the editor may either leave some *empty frames* showing only the empty staircase, or he may insert a shot from a different event to avoid the blend via a *cutaway*. [Bordwell and Thompson, 2010, p. 233]

  Instead of shortening the plot's time, *overlapping editing* uses repetition of the

end of the first shot in the beginning of the second shot. While most films show an action only single, some movies utilize overlapping editing to repeat spectacular events several times from different angles, e.g. martial arts films by Jackie Chan. [Bordwell and Thompson, 2010, p. 234]

Continuity Editing Although the techniques described here open up a great variety of possibilities for editing films, a quite narrow set of these have come to form a dominant style of editing. This style is called *continuity editing*. Since the first edited films it was the aim of filmmakers to tell their stories through organizing the shots in a coherent and clear way to provide *narrative continuity*. [Bordwell and Thompson, 2010, p. 234, p. 236]

Such continuity shall "allow space, time, and action to flow over a series of shots" [Bordwell and Thompson, 2010, p. 236]. This can be achieved by keeping the graphic aspects approximately consistent: balanced figures, symmetrical deployment, consistent lighting and centered action. [Bordwell and Thompson, 2010, p. 236]

In addition it is quite usual to link camera distance and shot duration. A long shot is cut to last longer than a medium shot and close-ups shots are displayed even shorter. The motivation behind this is to give the viewer more time to grasp the usually more detailed long shots. [Bordwell and Thompson, 2010, p. 236]

Axis of Action The basic principle which allows for spatial continuity is the *axis of action*, also called *center line* or the *180° line*. On this line, the main action of the scene — may that be a walking person, a conversation or a car race — is happening. To ensure spatial continuity it is important that the camera always stays on the same side of the line. In other words the camera might only be placed in the half-circle defined by the axis of action, that is in the 180°area. [Bordwell and Thompson, 2010, p. 236]

A very typical shot pattern for capturing a scene involving the two characters A and B is to start with a medium shot showing both figures, followed by a shot over the shoulder of character A focusing on character B and a shot over B's shoulder focusing character A. The first shot in this pattern is an *establishing shot* which shows the whole spatial arrangement of figures and objects in the scene and establishes the 180° line as well. The next shots follow the *shot/reverse-shot* pattern, as shots are captured from alternating ends of the 180° line and normally show characters in a three-quarter view [Bordwell and Thompson, 2010, p. 236, p. 238f].

This *180° system* has several advantages:

- consistent relative positions

- consistent eyelines
- consistent screen direction

[Bordwell and Thompson, 2010, p. 236]

As long as the camera does not cross the 180° line, the positions of figures and objects stay the same in relation to each other. In a dialogue scene with a shot/reverse-shot pattern each character will occupy the same screen part, no matter at which angle the camera captures the action from within the area on one side of the 180° line. Similarly, the direction in which the characters are looking stays the same. The screen directions remain as well. A motion starting at the beginning of the scene will continue to move in the same direction on the screen if the 180° system is respected, e.g. from left to right. [Bordwell and Thompson, 2010, p. 236f]

This system ensures that the viewer can follow the action and can always be aware of the scene's spatial configuration of figures and objects in relation to the action. [Bordwell and Thompson, 2010, p. 237]

Spatial information may also be presented through the *eyeline match*. By first showing a character which is looking offscreen, followed by a shot with the object being looked at, spatial continuity is enforced and the two shots are spatially connected. Although it is not necessary, directors often combine shot/reverse-shot patterns with eyeline match to help the audience recognize the locations of characters, even if they are not visible in each shot. [Bordwell and Thompson, 2010, p. 239f]

Another prominent type of shot used in continuity editing is the *reestablishing shot*. It is used, as the name suggests, to reestablish the overview over the scene's arrangement. A common pattern in classical continuity editing is *establishing/breakdown/reestablishing*. One aim of continuity editing is to let cuts pass unnoticed by the viewers. A tactic to support hiding cuts is *match on action* when a figure's movement begins in a first shot and continues in a second one. The continued movement between the two shots will grab the viewer's attention and connect the shots across the cut. [Bordwell and Thompson, 2010, p. 240, p. 242]

As with most film techniques the continuity system may not always be strictly adhered to. Sometimes a *cheat cut* is required in which the settings arrangement may be changed from one shot to the next, e.g. the narration requires so. Some movies might even cross the axis of action on purpose if the layout of the setting is clear and the crossing will not confuse the audience too much. [Bordwell and Thompson, 2010, p. 240, p. 243]

The 180° system also helps when the filmmaker wants to show two narrative lines of action in parallel by crosscutting. Bordwell and Thompson give the example of

a car crash in which one party is always driving from left to right and the other is driving right to left before the crash happens. These consistent orientations help to understand immediately which part of the scene is currently onscreen, even if two locations are shown alternately. Through crosscutting the audience can gain "unrestricted knowledge of causal, temporal, or spatial information by alternating shots from one line of action in one place with shots of other events in other places" [Bordwell and Thompson, 2010, p. 248]. [Bordwell and Thompson, 2010, p. 246ff]

Order, Frequency and Duration Bordwell and Thompson describe the temporal dimension of continuity editing by employing the terms of Genette [1983]: *order*, *frequency* and *duration*. The continuity system usually presents the events of the story in an 1-2-3 order, meaning that first a cause is shown and then the resulting effects. The most prominent exception are flashbacks which are mostly used to show things that happened in the past. Typically events are only shown once in continuity editing and are not repeated although certain events might be shown again in flashbacks to be recalled. [Bordwell and Thompson, 2010, p. 249]

Usually the screen duration is not longer than the duration of the story. Either the amount of time passing on screen and in the story are equal, or story time is omitted by means of a *temporal ellipsis* (see also *temporal editing* in chapter 2.1.7.3). In this case daily events like dressing, washing and having breakfast are shortened. [Bordwell and Thompson, 2010, p. 249, p. 254]

Besides such ellipsis which should go unnoticed some narratives call for more clearly visible hints that time has passed. Before the 1960s it was common to use dissolves, fades, or wipes for this purpose, following the rule that a dissolve symbolizes a short and a fade a longer time lapse. [Bordwell and Thompson, 2010, p. 254]

Later films often use cuts to symbolize time passing. Bordwell and Thompson present the example from Stanley Kubrick's *2001* in which a shot of a spinning bone in the air is followed by a shot showing a space station spinning as well to skip millions of years. Nowadays lighting, locale or figure positions changes are more common in contemporary movies. [Bordwell and Thompson, 2010, p. 254]

Another device for temporal ellipsis is the *montage sequence* which shows "a large-scale process or a lengthy period — a city waking up in the morning, a war, a child growing up, the rise of a singing star." In the 1930s some clichés for montage were established, like fluttering calendar pages or the special issues of newspapers to hint at the time passing. [Bordwell and Thompson, 2010, p. 254]

2.1.7.4 Sound

A movie's sound track is built independently from the visual track and can be manipulated separately as well. Often sound is perceived as a support for a movie's visuals. Even the terminologies "watching" a film and "viewers" treat sound as a less important part of a film, although most people these days would reject this neglect of sound. In the era of silent films there was live music in form of an orchestra, an organ or a piano. [Bordwell and Thompson, 2010, p. 269f]

Films with recorded sound were established in 1926. Since the 1970s, with the introduction of Dolby multi-channel sound, the quality rose and by the early 1990s almost all movies featured a digital sound track. Sound designer Michael Kirchberger noted: "Tracks are fuller and more of a selling point". Nowadays sound in cinemas has gained importance, as sometimes sound effects or dialogues are audible even before the image is shown and may carry important story information. [Bordwell and Thompson, 2010, p. 269].

Human perception is used to connect visual and audible information. Sound and image presented together will be perceived as one event. Because of this, the sound added to a movie may change the perception of the images. It may clarify, contradict or make them ambiguous. Especially horror movies draw on sound conventions. Bordwell and Thompson [2010, p. 270] remark that the noise of a creaking door lets the viewer expect to see the person who entered. However in a horror film a person already being in that room will be shown with fearful eyes, instead of another person entering the room. Sound coming from invisible or not yet seen sources is quite common in horror and mystery films. [Bordwell and Thompson, 2010, p. 270]

Film Sound Basics Central terms of sound perception are its *loudness*, its *pitch* and its *timbre*. [Bordwell and Thompson, 2010, p. 273]

Loudness, also known as a sound's volume, is the amplitude of a sound's vibration. The loudness of film sounds is steadily adjusted, for example, when background noises are lowered to make a conversation between two characters audible. The volume of the sound also gives hints on distance. Sounds farther away are normally perceived as being more quiet than closer ones. This effect is also known as *sound perspective* [Bordwell and Thompson, 2010, p. 273, p. 278]

A sound's *pitch* is described by its frequency. Few sounds in daily life consist of only one frequency, most are a mixture. Nevertheless a sound's pitch can often be perceived as high or low and furthermore help the audience to separate music, voices and noises. Furthermore the pitch of a sound may give hints on objects itself, as hollow ones normally sound dull while denser and harder objects have higher-pitched

tones. [Bordwell and Thompson, 2010, p. 273]

The *timbre* of a sound describes a sound's harmonic aspects and is also called its color or tone quality. Timbre helps in recognizing familiar sound. Bordwell and Thompson say that timbre depicts a sound's texture and "feel". [Bordwell and Thompson, 2010, p. 273]

All these three aspects of sound collude to form the sonic texture which also influences how the audience perceives a filmic experience. [Bordwell and Thompson, 2010, p. 274]

Editing Sound Three kinds of sound are differentiated in film production: speech, music and noise (or sound effects). Similarly to the visual part of the film, the editor selects the sound which will fit the desired purpose. Such sound does not have to be recorded together with the image. Although manipulations of the soundtrack are not as noticeable as visual ones, they are nevertheless demanding as well. In animated cartoons it is a common technique to arrange music, dialogue and sound effects in advance and to add the images later on to simplify synchronization of audio and video. [Bordwell and Thompson, 2010, p. 274f]

The sounds used in a movie do not have to be created especially for the current one. Instead many editors have either their own corpus or use and reuse existing media from sound libraries. A very famous example for a reused sound is the "Wilhelm scream" which was used in more than a hundred films. [Bordwell and Thompson, 2010, p. 275]

Like editing techniques, the sound in a film is arranged to guide and focus attention by making the crucial parts more prominent. Often dialogue bears important information, so it is usually processed to be clearly understandable. Sound effects are usually not as important but contribute to a realistic setting. Thus they often go unnoticed. In action parts however sound effects are very important "while music can dominate dance scenes, transitional sequences, or emotion-laden moments without dialogue". [Bordwell and Thompson, 2010, p. 275]

As sounds in film are selected to serve specific functions, they are usually not very realistic and provide a clearer and simpler version than sounds in real-life. Some are even quite unrealistic to guide the attention and manipulate the perception as the narrative requires. In a crowded scene usually only the important sounds, e.g. dialogue, can be clearly heard while other sounds are much more quiet. [Bordwell and Thompson, 2010, p. 275]

In contemporary films, sound is often *dry recorded* in a special non-reflective room and then electronically manipulated to create exactly the desired effect like the sound

of a telephone conversation. [Bordwell and Thompson, 2010, p. 276]

Editing film sound is not only about choosing a specific sound, it is also about arranging it together by *mixing*. Bordwell and Thompson remarked that one should not think of handling single sonic events but as a constant stream of sound which is embedded in a specific pattern of time and layers. [Bordwell and Thompson, 2010, p. 276]

While editing a dialogue scene which features a shot/reverse shot pattern it is a common practice to use *dialogue overlap* — a spoken line continues across a cut — to take attention from the changes in the image. Sound may also artificially be mixed to emphasize certain elements. For example, only a discussion in the background of a shot might be audible, even if other characters are shown talking in the foreground. [Bordwell and Thompson, 2010, p. 277]

Nowadays it is not unusual to use more than a dozen separate audio tracks in post-production to create a smooth soundtrack. In the 1930s, it was common that music was *sneaking in* and *sneaking out*, meaning that music was turned up when there were parts of the film without dialogue and turned down as people start to speak. In contrast current Hollywood movies often use a more abrupt soundtrack with higher dynamic range, made possible by sound systems like Dolby. [Bordwell and Thompson, 2010, p. 277]

The musical score used in films may either be arranged from already existing material or be specially composed for the film. Either way "the rhythm, melody, harmony, and instrumentation of the music can strongly affect the viewer's emotional reactions" [Bordwell and Thompson, 2010, p. 279]. An example from the film *Local Hero* (1983, by Bill Forsyth) in which different music themes are used in different settings, rockabilly music in Texas versus emotional folkish composition when being located at the Scottish seaside. Conventionally, certain facets of the narrative are connected to specific musical themes. [Bordwell and Thompson, 2010, p. 279]

Film Sound Dimensions Besides the sonic fundamentals of film and how they may be used, Bordwell and Thompson furthermore describe the relation of sound and other filmic elements in the dimensions of *rhythm*, *fidelity*, *space* and *time* [Bordwell and Thompson, 2010, p. 280].

- **Rhythm** As sound requires a duration to exist rhythm is a major aspect of sound, similar to its role in the mise-en-scène (see chapter 2.1.7.1) and editing (see chapter 2.1.7.3). [Bordwell and Thompson, 2010, p. 281]

  "Rhythm involves, minimally, a *beat*, or pulse; a *tempo*, or pace; and a pattern of *accents*, or stronger and weaker beats" [Bordwell and Thompson, 2010,

p. 281]. Typically these aspects are most present in the music parts of a soundtrack but rhythm may be found as well in sound effects like different sounds from galloping horses or even in the sounds of different machine guns firing bullets. Even speech contains a rhythm: namely characteristics in frequency and amplitude as well as patterns in pacing. [Bordwell and Thompson, 2010, p. 281]

Because both the motion in the images and the editing will provide a rhythm as well, the filmmaker may either try to coordinate the three rhythms or create disparities. In most films these rhythms are matched. For example characters are dancing to the rhythm given by the music. In animated cartoons the figures are drawn after the sound track completed to synchronize their movement perfectly with the music. A technique known as *Mickey-Mousing*. [Bordwell and Thompson, 2010, p. 281f]

However in some movies the rhythms are differentiated. This is often done to emphasize spoken text and to help let the cuts go unnoticed. [Bordwell and Thompson, 2010, p. 282]

- **Fidelity** In film sound terms, *fidelity* does not refer to a recording's quality but to how credible it is. Bordwell and Thompson describe it as the "extent to which the sound is faithful to the source as we conceive it". Furthermore, they remark that when the image of a dog barking is accompanied by a barking noise the sound keeps fidelity, no matter how it was created in the first place. Similar to other conventions, the awareness of a sound's non-fidelity may be typically used in humorous scenes. [Bordwell and Thompson, 2010, p. 283]

- **Space** The spatial dimension of a sound is determined by its *source*. Filmmakers distinguish between *diegetic* — coming from a source in the movie world — and *nondiegetic sound* — having an external source. Typical diegetic sounds are speech and sounds from characters as well as noises and music coming from objects and instruments within the film. The most common nondiegetic sound is the music in the soundtrack, and sometimes an omniscient narrator. [Bordwell and Thompson, 2010, p. 284]

  Diegetic sound may have an *onscreen* or an *offscreen* source, depending on the visibility of the source in the current image. Offscreen sound may save costs and time as it allows for simulating larger space and settings. For example typical airplane sounds create the illusion of an in-flight scene, although only a few characters and the seats might be visible. It can also simulate a larger space and may provide knowledge quite economically. [Bordwell and Thompson,

2010, p. 285]

Some movies combine point-of-view shots with offscreen sound to emphasize a subjective narration, i.e., to let the audience see and hear what the character is perceiving. Similarly, diegetic sound telling the thoughts of a character may be used to provide a narrator. Here filmmakers differentiate between *external diegetic sound* and *internal diegetic sound*. While external diegetic sound is perceived as having its source in the shot, internal diegetic sound originates from a character's mind. Such narrations, as well as the nondiegetic sound, are known as *sound over*. [Bordwell and Thompson, 2010, p. 285, p. 289f]

Especially since the introduction of multichannel sound, filmmakers make use of the *sound perspective*. Similar to visual depth cues (see chapter 2.1.7.1), sound may have a spatial position in scenery. Louder sounds are perceived to be closer and softer than the ones more remote. Reflected and direct sound causes a timbre which provides hints on the environment as well as the distance, e.g., echoes. [Bordwell and Thompson, 2010, p. 292f]

However such a sound perspective does not have to be realistic, as the voice of a character in a long shot is usually much clearer than it would be in reality. Furthermore voices will not necessarily change their perspective in a conversation shot, even if the camera moves. [Bordwell and Thompson, 2010, p. 292f]

A special case of sound perspective is a telephone conversation, known as the *telephone split*. The participants are usually shown alternating and the currently invisible character is heard "more coarsely rendered and more reverberant, carrying low pitches and providing little ambient sound". [Bordwell and Thompson, 2010, p. 293]

- **Time** Film sound has time-related aspects as well, since its timings may correlate with the visual. In the final movie sound can be either *synchronous sound*, meaning that the sound is perceived at the same time the corresponding event is shown, or it might be *asynchronous sound*. Sound may become asynchronous due to technical problems during playback but some filmmakers use it on purpose to create distraction or comic effects. [Bordwell and Thompson, 2010, p. 294]

  Sound may also influence the time of story and plot. Although most sound in movies is *simultaneous sound*, "when characters speak onscreen, the words we hear are occurring at the same moment in the plot's action as in story time" [Bordwell and Thompson, 2010, p. 295], some sounds might be *nonsimultane-*

ous. Such nonsimultaneous sound can be a sonic flashback or a flash-forward in the visual track, i.e., the sound is audible in story time before the corresponding image is visible. A common filmic device is the *sound bridge* where the sound from the previous scene is overlapping the image of the next one. [Bordwell and Thompson, 2010, p. 295f]

The sound might also be later in story time than the image. This can often be found in trial dramas in which a witness can be heard giving a testimony while images from the past give a corresponding illustration. [Bordwell and Thompson, 2010, p. 296]

2.1.8 Movie Genres

In this work an automatic trailer generation system which covers the action genre is extended to also handle the genres horror and comedy. In order to paint a picture of the scientific understanding of genres, an insight is provided in the following part.

This section is based mainly on Grant [2007, chapter 1]. Movie "genres are a way to characterize a film" [Bordwell and Thompson, 2010, p. 329] and as such are central for reviews and television coverage of movies [Bordwell and Thompson, 2010, p. 329]. According to Altman [1999, p. 14] the purposes of genres are:

- to give blueprints for industrial production,
- to give a structure on which individual films are based,
- to provide labeling categories easing the communication between distributors and exhibitors,
- to provide an implied contract to the audience for expectations.

Furthermore Altman describes the term as "a structure through which material flows from producers to directors and from industry to distributors, exhibitors, audiences and their friends" [Altman, 1999, p. 15]. Genres are "the product of audience and studio interaction" [Schatz, 1981, p. 16]. Diduck said that "genre signifies a defined or contained cinematic space where something familiar will happen", enriched by an unpredicted "redeployment of generic conventions" [Diduck, 2008, p. 33].

Research and History Filmic genre names are often taken from the field of literature. One of the first mentioned genre theorists was Aristotle, who described different endings in comedy. In the 1980s the topic of movie genre studies emerged as its own field of research. [Altman, 1999, p. 4, p. 13], [Grant, 2007, p. 4]

With the introduction of sound movies, first classifications emerged. According to the amount of spoken lines the films were labeled as being *0%-,50%- or all-talkie* ones. Later on, first genres as we know them were established, among these western and musical films. Genres developed over time. Before the genre of western movies was established, such movies were promoted as being related to chase-, railroad- and crime films. *The Great Train Robbery* (1903, by Edwin S. Porter) has been one of the first Western movies promoted as such. [Grant, 2007, p. 4, p. 6, p. 18]

In the so called *Studio Era* (roughly 1920s — 1950s) the production of movies was organized similarly to the industrial production of cars. Actors and crew were employed directly by the studios in an effort to supply genre movies as "dependable

products" [Grant, 2007, p. 7] and to maximize profits. In order to maximize acceptance, producers used "repetition and variation of commercially successful formulas". [Grant, 2007, p. 7f]

This "Hollywood style" promised "crisp and seamless flow of the story combined with high production values" [Grant, 2007, p. 7]. An example for such standardized movies given by Grant is the James Bond series. All individual movies of this series may have different directors, writers and actors, but contain the basic formula: "lots of action, fancy gadgets, beautiful women and colourful villains". [Grant, 2007, p. 7f]

During the studio era the "Hollywood studios had achieved vertical integration" [Grant, 2007, p. 8] and controlled "distribution and exhibition as well as production" [Grant, 2007, p. 8]. In 1948 the US Supreme Court forced the major studios to split their monopolistic structure and to "divest their exhibition chains" [Grant, 2007, p. 9]. Although the system limited the creative potential of filmmakers it provided a solid and consistent environment. [Grant, 2007, p. 8f]

Today film genres are an essential part of the business. [Bordwell and Thompson, 2010, p. 346]

Genre Elements Films of a specific genre share common elements. These can be grouped into *conventions*, *icons*, *setting*, *stories and themes*, *characters, actors and stars*, and *viewers and audiences*. [Grant, 2007, p. 9]

- **Conventions and Iconography** *Conventions* are often used techniques and elements typical for specific genres. Among these are "bits of dialogues, musical figures or styles and patterns of mise-en-scène" [Grant, 2007, p. 10]. Conventions do not have to be realistic though. They "function as implied agreement between makers and consumers to accept certain artificialities" [Grant, 2007, p. 10]. The acceptance of such artificialities depend on the context. An example is the inclusion of dance and performances choreographed exclusively for the camera in musicals. These interrupt the narrative and are thus very unrealistic but common and accepted for this genre. [Grant, 2007, p. 10]

  Other examples for conventions are low-key lighting and narrative flashbacks in film noir as well as tight framing in horror movies. Melodramas are often quite stylized by means of their mise-en-scène. Some genres may even have characteristic fonts in the opening credits which form their own graphical style. An example is the "Wild West font" used in many western movies. [Grant, 2007, p. 11]

  Conventions are also present in the soundtrack. Specific locations and characters are often accompanied by specific *Wagnerian leitmotifs*, i.e., specific

reoccurring musical themes connected to figures or locations, to support the narrative by means of emotional effects. Specific genres also feature typical forms of music, like sweeping strings in romantic melodramas and electronic music in futuristic science-fiction movies. Common conventions can also be used as a parody. Their utilization may lead the audience to false beliefs, and they can be used for disturbing purposes. [Grant, 2007, p. 11]

Icons are "symbolically charged objects and events" [Grant, 2007, p. 11], according to historian Erwin Panofsky. They are also described by genre critics, for example Lawrence Alloway, as *familiar symbols*. Such symbols "have a cultural meaning beyond the individual work in which they appear" [Grant, 2007, p. 11f]. Their meaning is "symbolic because of their use across a number of similar previous texts", and in this case movies. [Grant, 2007, p. 11f]

Buscombe describes an inner and outer form of films. Inner forms are referring to the theme of a film while outer ones are referring to the objects typically found in genre movies. Examples for such objects are horses, wagons, buildings, clothes and weapons in western. [Buscombe, 2003, p. 14f]

Iconography refers "to particular objects, archetypal characters and even specific actors" [Grant, 2007, p. 12] in genre films. Following the example of the western genre, a villainous gunfighter might be portrayed in a black dress and with two guns. This is an "iconographic wardrobe of a generic type" [Grant, 2007, p. 12], not paying much attention to historical reality. [Grant, 2007, p. 12]

The general mise-en-scène (see chapter 2.1.7.1) of a genre may also be related to iconography. For example horror films with low-key lighting and Gothic design or visually excessive melodramas. In western movies, typical icons are "gunbelts, Stetson and spurs", according to Grant. Film noir typically features *chiaroscuro lighting* iconographically. Furthermore Grant describes "pinstripe suits, dark shirts and white ties" as being defining for "which side of the law characters are on in gangster films". Similarly, in westerns the color of hats typically allows the audience to tell apart hero and villain. [Grant, 2007, p. 12]

- **Setting** Some genres are quite fixed to their story's space and time while others are not. On one hand, westerns are usually restricted to the era of the wild west and to the landscape situated west of the Mississippi River. Musicals, on the other hand, may take place anywhere. Their locations can be the "streets and docks of New York City in *West Side Story*" (1961, by Jerome Robbins and Robert Wise)[Grant, 2007, p. 14] or a supernatural village as in

Brigadoon (1954, by Vincente Minnelli). Common settings for gangster films are cities and urban areas. [Grant, 2007, p. 14]

A movie's story may unfold over a longer period of time, like some romantic comedies and dramas do. Some science fiction films use contemporary architecture while taking place in the future. This yields to "a more disturbing continuity between the present and the future" [Grant, 2007, p. 14]. An example is the utilization of subway systems in *THX 1138* (1971, by George Lucas). Isolated and rural settings are frequently used for horror movies as well as basements of mysterious old and dark houses. Some science-fiction films blend with the horror genre, such as *Alien* (1979, by Ridley Scott). In this film "the rusting spaceship became the futuristic equivalent of the old dark house full of unseen dangers". [Grant, 2007, p. 14]

- **Characters and Actors** The English novelist Forster describes *flat* and *round* characters. He states that characters are first flat and "built around one idea or quality" [Forster, 2010, p. 103]. By adding additional attributes they begin "to curve towards the round" [Forster, 2010, p. 103] and thus gain depth in personality. [Forster, 2010, p. 103f]

  Grant describes the characters in genre movies as "recognizable types rather then psychologically complex characters" [Grant, 2007, p. 17]. He mentions black hats (bad guys) and white hats (good guys) in Western movies, the femme fatal in film noir, the comic sidekick and the schoolmarm[3] as well as gunfighters in westerns as conventional characters. Ethnic stereotypes are often used in genre movies like "the Italian mobster, the black drug dealer, the Arab terrorist" [Grant, 2007, p. 17f] and "the cross-section of soldiers" [Grant, 2007, p. 18]. Although such flat characters seem to show a lack in originality, their use in genre movies is often sufficient. [Grant, 2007, p. 17f]

  Types of characters often "provide similar kinds of actions and purposes within a story" [Grant, 2007, p. 18] in genre movies, Propp, Scott, and Wagner described their appearance and acts as "functions" with respect to their "significance for the course of the action" [Propp, Scott, and Wagner, 2010, p. 21]. For example the function of a "donor" offering both a test and a help for the hero. [Grant, 2007, p. 17f]

  Often certain actors become associated with a specific kind of character because of their repeated casting in films of similar types and genres, like John

[3] schoolmarm: a (strict) female teacher [Stevenson, 2010, p. 1592]

Wayne and Clint Eastwood in western movies. Such associations also increase the depth of characters played by the same actor in different genre movies. This connection between actors and genres led the actors to become part of the iconography of the particular genres. [Grant, 2007, p. 18f]

Some filmmakers also developed *typage*, a special kind of casting where actors are selected according to their physical appearance. A prominent example is the soviet filmmaker Sergei Eisenstein, who cast "fat cats" for the role of capitalists. Such iconographic casting is still existing in contemporary mainstream cinema, as the physical appearance of an actor makes him particularly suitable for specific genres. An example is Arnold Schwarzenegger, whose muscular body suggests that he stars in action films but also suits playing in other movies for comedic purposes. [Grant, 2007, p. 19]

- **Stories and Audience** As described in section 2.1.1 most genre movies are based on a common dramatic structure in which an individual hero "must overcome obstacles to achieve a goal" [Grant, 2007, p. 15f]. These obstacles are often the result of a disturbance in the diegetic world of the movie which has to be resolved during the plot. [Grant, 2007, p. 15f]

  Bordwell and Thompson note that this primary narrative arc is accompanied by a romantic desire of the hero. Grant presents the example of a hero who defeats a monster with "masculine prowness" and the help of technology, and who finally "wins the scientist's lovely assistant or daughter, along with paternal blessing of the elderly scientist". [Bordwell, Staiger, and Thompson, 2003, p. 16], [Grant, 2007, p. 16]

  Schatz claims that "all film genres treat some form of threat — violent or otherwise — to the social order" [Schatz, 1981, p. 26], like gangsters in a crime movie or monsters in a science fiction film. [Schatz, 1981, p. 26]

  Neale says that "the existence of genres means that the spectator, precisely, will always know that everything will be 'made right in the end', that everything will cohere, that any threat or danger in the narrative process itself will always be contained" [Neale, 1980, p. 28]. In horror movies the monster is finally defeated, in Westerns the cavalry comes for rescue. [Grant, 2007, p. 17]

  Was it not for the audience, genres would not be meaningful. Even the first movies have been promoted by their affiliation to a specific genre. Such genre-specific promotion indicates the kind of story as well as which "kind of pleasure they are likely to offer" [Grant, 2007, p. 20], e.g., frightening and violent horror

> movies. These promotions help the potential audience to select which films to see. [Grant, 2007, p. 20]
>
> Viewers and makers of genre movies implicitly agree on a contract which gives the viewers an idea of what to expect. These expectations are based on previous conventional familiarities. Although alteration may be wanted in genre movies, "originality is to be welcomed only in the degree that it intensifies the expected experience without fundamentally altering it" [Warshow, 1962, p. 100].

Genre Purity and Definition Early movies were considered to belong to only a single genre which was preserved even after the affiliation of movies with multiple genres had been established. This was because of the understanding that such a categorization is fixed, "once generically identified by the industry, films are typed for life". [Altman, 1999, p. 18f]. On the other hand, Staiger said "Hollywood films have never been 'pure"' [Staiger, 2003, p. 185].

The main problem while defining genres is the selection of movies which constitutes their corpus. According to Staiger there are four ways to approach this [Staiger, 2003, p. 187]:

- idealist method — "judges films against a predetermined standard, is proscriptive in that certain films are privileged over others to the extent that they remain close to the chosen model"
- empiricist method — "involves circular logic in that the films selected already have been chosen as representing the genre"
- a priori method — "common generic elements are selected in advance"
- social convention method — "is problematic in how cultural census is determined"

A larger issue concerning the definition of genres is the "empiricist dilemma" as described by Tudor [1974, p. 138])

> To take a *genre* such as a 'Western', analyse it, and list its principle characteristics, is to beg the question that we must first isolate the body of films that are 'Westerns'. But they can only be isolated on the basis of the 'principle characteristics' which can only be discovered *from the films themselves* after they have been isolated. [Tudor, 1974, p. 135]

Genres can be differentiated on several levels. Some genres can be defined by the setting they take place in and the kind of narration. Among these are crime, science

fiction and westerns. Other genres work by having emotional effects on the audience. These are genres such as horror, pornography and comedy. [Grant, 2007, p. 23]

Williams describes the genres horror, melodrama and porn as "body genres" because of the strong physical reaction they provoke (fear, tears and sexual arousal). A common way of judging the quality of genre movies is the extent to which they can trigger such emotional responses. [Williams, 2003, p. 142f], [Grant, 2007, p. 23]

Generic categorizations such as narrative film, documentary, abstract and experimental are certainly of use for comprehension. However, they are considered not detailed enough to be applicable to genre criticism [Grant, 2007, p. 23]

Kaminsky and Leitch agree on "the difficulty of defining the genre of crime film, since it includes gangster films, detective and mystery films, action films, police films and heist movies" [Grant, 2007, p. 23]. [Kaminsky, 1974, p. 9], [Leitch, 2002, p. 1-18]

Collins described an "ironic hybridization" of genres since the 1980s, for example the combination of western and science fiction in movies like *Blade Runner* and *Back to the Future 3* [Collins, 1993, p. 242f]. Another example for a difficult genre definition is the monster- or creature movie. Sobchack argues that such a film "sits [..] between horror and SF [science fiction]" [Sobchack, 1980, p. 47]. Some of these films combine the setting of science fiction, such as spaceships and laboratories, with the elements of horror movies. [Grant, 2007, p. 24]

In this chapter an insight in narration, film production and movie genre studies was presented. This provides a background for understanding the domains and establishes terms and processes of film making. The editing techniques for example are relevant for the creation of trailers. Certain characteristics of specific genres also assist the following manual analysis of such trailers and the following definition of genre specific abstraction models.

2.2 Trailers

2.2.1 Overview

In early serial films, excerpts of coming episodes were shown at the end. This utilization of short fragments trailing the main film gave rise to the nowadays common term *trailer*. [Hediger, 2001, p. 61]

Scientific research about movie trailers covers not a large field. It is mostly treated as a subtopic of film advertising. An overview of the development of trailers and an analysis of common structures is given by Hediger [2001]. He describes an analysis of different kinds of trailers, for which he used a corpus of 2039 examples. [Diduck, 2008, p. 25f], [Hediger, 2001, p. 32]

2.2.2 Trailer Types

In the movie industry, different types of trailers are used in marketing campaigns. The important trailer types are *Trailer*, *TV-Spot*, *Teaser* and *Featurettes*. In the following these types are described in detail. A focus is put on the US market since Hollywood trailers form the domain of this work. [Goldberg, 1991, p. 33f], [Hediger, 2001, p. 32]

Theatrical Trailer The theatrical trailer — also known as *regular trailer* — is the most prominent in Hediger's book, and the main objective of this thesis. Such trailers "consist of cuts or samples of the movie that have been skilfully edited, scored, narrated, and titled" [Goldberg, 1991, p. 42]. Most footage is taken from the movie, although it is sometimes supplemented with material from additional shooting or special graphic effects. Goldberg mentions a digital countdown in the trailer for the Bond movie *License to Kill* (1989, by John Glen) as an example for such an effect. [Goldberg, 1991, p. 34, p. 42], Hediger [2001, p. 32]

According to Goldberg, the typical length of a theatrical trailer is between 2 and 3.5 minutes while the ideal length is 3 minutes, although Hediger notes that the actual average length of trailers before the 1960s was 144 seconds. Since then it decreased to about 120 seconds. The length suggested by guidelines is 90 to 120 seconds, although trailers for the international market are often a little bit longer. [Goldberg, 1991, p. 142], [Hediger, 2001, p. 46].

The length of a trailer also depends on its content: "if the audience can 'feel the length', it's too long" [Goldberg, 1991, p. 42]. Also a trailer with a reasonable length

has a better chance to get on the platter[4], so a "good rule of thumb is the shorter the better". [Goldberg, 1991, p. 142]

A trailer should be planned similarly to a movie: "a good trailer is built around an idea and a point of view" [Goldberg, 1991, p. 42]. According to Goldberg, it is better to start with a script or storyboard instead of simply arranging footage. [Goldberg, 1991, p. 42]

Trailers are used to build "awareness for the movie" [Goldberg, 1991, p. 42] as they are "the first impression that potential consumers have of the movie" [Goldberg, 1991, p. 42]. A trailer creates expectations about what to expect of the film and thus "should position the movie" [Goldberg, 1991, p. 42] and "influence the target audience favorably" [Goldberg, 1991, p. 42]. [Hediger, 2001, p. 27]

Occasionally multiple trailers for a film are produced. Sometimes a trailer is edited to address a specific target audience and certain interesting footage is not used. An example is a trailer for *Rocky* which includes ice-skating sequences in order to promise "a simple love story" instead of fighting scenes. This was done in order to make the movie look more interesting for women. [Goldberg, 1991, p. 42f, p. 44f]

In the beginning trailers were produced and distributed by a specialized company, the National Screen Service (NSS). This company also rented posters and other advertising material. All major film studios had exclusive trailer makers among the employees of the National Screen Service. About 35 feature films were produced by a single studio per year. Trailer makers like Max Weinberg were able to deliver a raw cut trailer within two days. "He did this for more than fifty films each year." [Goldberg, 1991, p. 43]

As movies were not released as quickly in that time period, trailer makers were able to work with material from the completed film. Before cutting, a script was made in advance, and raw versions of the trailer were sent for approval to advertising directors. Suggested changes were incorporated and the final trailer was released by the National Screen Service. [Goldberg, 1991, p. 43f]

Nowadays, trailers are no longer delivered by the National Screen Service. Instead, they are created and shipped by film distribution companies. The production of trailers as well as television and radio spots is generally done by specialized creative agencies who prepare other audiovisual material as well. Such agencies are mostly located in Los Angeles. Some of these creative agencies are specialized to specific genres. Trailers are also created internally by audiovisual units of their respective distribution companies. [Goldberg, 1991, p. 44]

[4]Platter: "the large reel that the feature is mounted on"[Goldberg, 1991, p. 42]

It is not unusual that more than one team is working on a trailer for a specific movie — sometimes referred to as *trailer derby* — "to provide a choice of trailers" [Goldberg, 1991, p. 44f] for the distributors. As nowadays the time between post-production and release of a film is much shorter, "most trailers are created from a rough cut of the movie" [Goldberg, 1991, p. 45]. [Goldberg, 1991, p. 43ff]

Teaser Trailer Goldberg describes teasers as trailers with a runtime of less than 90 seconds. Their purpose is to *tease* the audience, to inform about and raise interest in upcoming movies. [Goldberg, 1991, p. 33]

Teaser trailers are used to *position* a film, and to give a first impression of what to expect. Such trailers are also called *advance trailers*, as they are released several months in advance of the theatrical trailers, and usually six to eight months before the movie's release date. [Goldberg, 1991, p. 33ff], [Hediger, 2001, p. 297]

The footage used in teaser trailers can include "still photographs, key scenes from the movie, special footage, actual footage from the movie, and just graphics" [Goldberg, 1991, p. 34]. Special footage might be a shot exclusively made for the teaser, for example an address held by a character, as done in the teaser for *The Pink Panther Strikes Again* by Henry Kissinger. [Goldberg, 1991, p. 34, p. 37ff]

First teaser trailers were used by MGM in the 1930s and became more common since around 1960. [Hediger, 2001, p. 297]

TV Spot A television spot is a short trailer with a length of 30 or 60 seconds. Besides the personal appearance of a star in a late night television show, it was one of the few ways to get broad attention for a movie around the 1950s. [Goldberg, 1991, p. 51], [Hediger, 2001, p. 165]

Most TV spots use mainly footage from the movie and "are almost always a short version of the trailer" [Goldberg, 1991, p. 143]. The style should match that of the adjacent television program, meaning that a spot scheduled to air during an action program might not be appropriate during a comedy show. [Goldberg, 1991, p. 51, p. 143]

Featurette A featurette, also known as *making-of trailer*, covers not only the scope of the film but also its production by presenting "behind-the-scenes" footage. It is often used on television to fill up timeslots not fully occupied by movies, e.g., a feature film with a length of 100 minutes was shown on television in a two hour slot. After subtracting the time used for commercials, at least 12 minutes were paid for

but left unused. This unused time was filled with material about movies and their stars, leading to the introduction of featurettes. [Goldberg, 1991, p. 50ff]

Featurettes also served as appetizers in foreign territories with state-owned television stations. As these stations did not show commercials, there was no way of buying air time for showing trailers. However, featurettes were aired if they were reasonably interesting. [Goldberg, 1991, p. 52]

A good featurette should be based on an idea or a point of view and should be focused on a specific event during the shooting of a film. This is done to minimize costs and shooting time. A good example are the underwater scenes in *The Abyss*. The featurette for *The Abyss* showed technical equipment like cameras and lighting booms together with the film crew in the water. The director James Cameron explains the difficulty and novelty of this special shooting. [Goldberg, 1991, p. 52f]

According to Goldberg the length of a typical featurette varies between 8 and 15 minutes. [Goldberg, 1991, p. 54]

2.2.3 History of Trailers

The remainder of this chapter focuses on theatrical trailers. Hediger claims that there are two main trailer types: the *classical* and the *contemporary* one. In each type he identifies a common structure without major differences between film studios. Within these two trailer types, he spotted two different plot kinds, the *mystery plot (Rätselplot)* and the *suspense plot (Spannungsplot)*. [Hediger, 2001, p. 56]

Around 90% of the trailers produced by Hollywood studios follow either the classical, the contemporary or a mixed form. [Hediger, 2001, Apendix 4 p. 35]

Classical Trailer The era of the classical trailer dates back to approximately 1933 and lasted until the 1960s, although the first trailers appeared around 1912. [Hediger, 2001, p. 37, p. 61]

Classical trailers didn't reveal much information about the plot. Instead they concentrated on the starring actors and a few key scenes. Often they featured general phrases like "a story of love and passion" and lists of cast members. Especially trailers for thrillers and detective films asked questions like "Who was the mysterious woman behind the curtain? Why did she reveal the secret of xy to him?". [Hediger, 2001, p. 39]

Hediger refers to these trailers as having a *mystery plot*, because of the questions they ask and the focus on causing curiosity about the story of the movie. Such trailers have a structure very different from common narratively closed and chronological concepts. This means that such an excerpt from the movie not only raises

the question 'what happens next?' but also forms a desire to get to know how the story had developed to this point. As Brewer and Lichtenstein note, "the author leaves some significant event out of the discourse, but lets the reader know that the information is missing" [Brewer and Lichtenstein, 1980, p. 6]. Between 1928 and 1960 almost every trailer (93%) was produced utilizing this scheme. [Hediger, 2001, p. 39]

Almost two-thirds of the trailers from this exhibit a common structure of four elements in which the first two may be interchanged [Hediger, 2001, p. 42]:

(Intro/Title)/Development/Credits

- Intro — introduction of the basic content of the movie
- Title — announcement of the movie's main title
- Development — showing the intro's theme in more detail, often combined with cast, some excerpts from the movie and more titles
- Credits — final display of the title

Since the 1960s a different kind of plot has emerged: the *suspense plot*. Instead of invoking the feeling of curiosity towards the story, the focus was put on raising suspense in the audience. The structure of the mystery trailer changed into a highly concentrated three-act one (see chapter 2.1.1). It consists of the *exposition* and *introduction* of the problem, and the *confrontation*. Such a trailer ends with a cliffhanger, often at the point with the highest suspense. [Hediger, 2001, p. 39f, p. 44]

These trailers outline whether the hero will succeed or fail, and concentrate on summarizing the first two thirds of the story. Instead of posing numerous questions, they reflect the style and plot-structure of the movie. Hediger calls these trailers *two-third*-structured and centered on the protagonist. [Hediger, 2001, p. 39ff, p. 44]

The Structure of Contemporary Trailers In between approximately 1961 and 1980 a mixed form of trailer emerged. Such trailers combined the classical structure with a suspense plot. Starting in 1981, the use of the contemporary trailer became dominant. [Hediger, 2001, p. 45f, p. 55]

The classical structure evolved into a model of the film itself [Hediger, 2001, p. 44]. The introduction became the exposition, the title the inciting incident, the development the confrontation. The title remains and is sometimes supplemented by the *button*. The button is a final gag or fragment of a scene shown in the end

after the title which aims to help the audience to keep the film in mind to finally trigger the decision to watch it. [Hediger, 2001, p. 44f, p. 48, p. 292]

Although the structure and plot form changed, the typical length of a trailer stayed more or less the same. The average length of trailers designated for the U.S. market is 144 seconds before 1960, and about 120 seconds since then. He also noted that trailers produced for an international audience are often a little longer. [Hediger, 2001, p. 46]

The classical and the contemporary trailers also differ in their formal parameters. While classical trailers featured text overlays on footage and narrative voice-overs, contemporary trailers use intertitles, often containing texts which can as well be spoken by an anonymous voice-over. Such anonymous narration is quite common in modern trailers. Hediger found such narrations in 82.3 % of the modern trailers in his analysis. [Hediger, 2001, p. 47]

Classical trailers feature wipes and lists of attractions which are no longer used in contemporary trailers. In classical trailers, names of stars appeared within the first 30 seconds of the trailer, often during the introduction or the development. In current trailers, the names of the actors and the movie title are mostly shown towards the end. Unique features of contemporary trailers are the separation of audio and image, as well as the use of the button. [Hediger, 2001, p. 48]

References to literary sources and previous works of the producers were quite common in classical trailers. In contemporary trailers, however, their use decreased. The repertoire of rhetorical parameters used in trailers was reduced over time. Hediger notes that *endorsements*, direct approaches of the audience by actors, are no longer used. Similarly, references to the shooting and the film locations almost disappeared. Trailers moved away from presenting a huge *knowledgeability*, *self-consciousness* and *communicativeness*. Contemporary trailers no longer feature a discourse *about* the film but rather provide a discourse *through* the film. [Hediger, 2001, p. 48ff]

Instead of promoting the novelty of the movie and employing pseudo-journalistic methods, modern trailers work by the principle of *storytelling as selling* and embedding the excerpts of the movie in the suspense plot's narrative context [Hediger, 2001, p. 50].

In his analysis Hediger concluded that the ratio of the average shot length between a movie and the corresponding trailer varied. Starting from 1:1.35 in 1914, it reached a maximum of 1:2.93 in the 1970s and settled at around 1:2.76 in the 1990s. The average shot length in trailers changed from 4.8 seconds in 1914-1927 to approximately 2 seconds (1971-1990), and 1.27 seconds in the 1990s. The shortening of shots is a result of new styles of montage which separate audio from the corresponding image. [Hediger, 2001, p. 51]

In classical trailers shots were often condensed. However, mostly sound and image were not isolated. The sound was subordinated to the image. This changed with the introduction of contemporary trailers. The bond between soundtrack and image dissolved and the order changed: the dialogues and narrative voice-overs formed the course of the story, while the image supplements the soundtrack in an illustrating way. Scenes became fragmented and sequences reorganized. [Hediger, 2001, p. 51]

An important pattern in the montage of modern trailers is the *grid cut*, a special kind of montage of visual and audible footage, loosening their link. On the *dominant layer* a scene is used as basis, while selected shots on the *assigned layer* supplement the main scene. The soundtrack is usually taken from the dominant layer, and the visual footage of the assigned layer is arranged in a grid pattern "overlapping" the dominant layer. The visual is thus alternating between the two layers and allows the inclusion of additional visual elements in an existing scene.[Hediger, 2001, p. 55, p. 52]

According to Hediger [2001, p. 52] there are three different kinds of relation between the layers:

- analytical — the dominant layer shows a situation and the assigned layer the events causing it
- anticipatory — the dominant layer shows an event and the assigned layer its consequences
- illustrative — sequences in the assigned layer are illustrating an event on the dominant layer without a specific temporal relation

The relation between sound and image is mostly semantical in contemporary trailers. The length of shots is adjusted to fit the length of sound events, motions are matched to the music, and audible effects are combined with visual transitions. White dissolves are often used together with *tutti*[5] chords in order to symbolize intensity and actuality. [Hediger, 2001, p. 53]

In classical trailers, music is mostly used in a narrative context, and seldom in a rhetorical one. In contemporary trailers these modes are combined. A sudden change of music, cutting rate and narrative voice-over may indicate that a film is actually not a comedy but a thriller. Thus the music is used rhetorically by indicating the change of tone and narratively by explaining the mood in a subliminal manner. [Hediger, 2001, p. 54]

[5] "The term *tutti* (all) designated performance by the combined solo and orchestral groups" [Roeder, 1994, p. 24].

2.2.4 Contemporary Trailer Structure

According to Hediger a contemporary trailer has the following characteristics [Hediger, 2001, Apendix 4, p. 29ff, p. 66], [Hediger, 2001, p. 47f]:

- the length is between 120 and 150 seconds
- the average shot length is 1.4 seconds, about one-third compared to the movie
- a suspense plot is used, without the classical structure
- most trailers (82.3%) feature an anonymous narrator
- *key art*[6] is part of 63.6% of the trailers
- sound and image are often (60.7%) uncoupled by using grid cuts
- text overlays are used in every second trailer (53.4%)
- most trailers feature the title towards the end (90.5%)
- actor names at the end (64.6%)
- a button at the end is used in 22% of the trailers
- references to directors 9.5%

Contemporary trailers form a model of the film. The image track is adjusted to match the sound track. Extracts of dialogues and narrative voice-overs form the course of the story, while visual sequences illustrate the sound track. [Hediger, 2001, p. 50-53]

The background information described in this chapter provides a historical background and theoretical aspects about different trailer types as well as their essential structure. This information about the generic structure of trailers provides knowledge for the in-detail analysis of trailers in chapter 4 and formal parameters for the definition of abstracting models in chapter 5.5.

[6] key art describes graphical signets and icons used in advertising, for example the *Batman* logo

Chapter 3

Automatic Video Abstracting

Several researchers and projects have dealt with video abstracting and automatic trailer generation in particular. In this chapter a choice of relevant approaches is presented.

3.1 Video Abstracting

A systematic overview of research in the field of video abstracting is given by Truong and Venkatesh [2007]. They introduce two general approaches on video abstracting: the extraction of *keyframes* - also known as *representative frames* - and *videoskimming*. Furthermore, they describe basic concepts and they collected corresponding research literature for these approaches. While keyframe systems extract representative frames from a video source, video skimming results in a video consisting of several excerpts joined by cuts or other transition types.

3.1.1 Keyframe Systems

The simplest approach to keyframe generation would be to just extract frames at given intervals, without paying attention to the content [Truong and Venkatesh, 2007]. However, such methods could potentially omit important parts of the video due to missing content analysis. In their publication, Truong and Venkatesh [2007] therefore focus on abstraction approaches which take the structure and content of the source into account. For keyframe extractions, they focus on four attributes: *number of keyframes*, *unit*, *representation scope* and *underlying computational mechanism*.

Number of Keyframes The number of keyframes, which should be extracted from a source video, can be *a priori*, *a posteriori*, or *determined*.

- **A Priori** A number of keyframes is set before the extraction starts. It might either be a fixed number or a ratio with respect to the length of the source.

- **A Posteriori** A number of keyframes is not known before the extraction process is finished. Such an approach can assign more keyframes to visually complex parts and less to those parts with little changes.
- **Determined** An algorithm with a determined number of keyframes is similar to an a posteriori one, with the exception that the number of keyframes is determined prior to the actual extraction. An example are cluster-based approaches, where the number of clusters may be determined prior to the final clustering process [Hammoud and Mohr, 2000, Ferman and Tekalp, 2003, Yu et al., 2004].

Unit While extracting keyframes, a crucial question is which unit is represented by a keyframe. Typically, these units are either a *shot* or a broader defined *clip*. Truong and Venkatesh [2007] use a different definition of clip than the SVP team (see chapter 3.2.7). Here, clip refers to video excerpts longer than a shot, such as scenes or story units.

Basic keyframe extraction algorithms may select terminal frames as keyframes, as well as a middle frame. Other methods choose keyframes depending on the visual dynamics. Shot-based keyframe extraction depends on a reliable detection of shot borders, for which excellent results have been reported [Smeaton et al., 2004].

Clip-based approaches do not depend on a shot detection [Girgensohn and Boreczky, 1999, Yu et al., 2004]. Shot-based keyframe extraction may result in redundant keyframes being extracted for similar shots. For long videos, a shot-based method may also result in a large number of keyframes, whereas a clip-based approach might allow for less keyframes while not loosing much information. [Truong and Venkatesh, 2007]

Representation Scope Truong and Venkatesh [2007] distinguish between two scopes of video material that a keyframe can represent, either the local neighbourhood [Xiong et al., 1997] or a noncontiguous segment. An algorithm that focuses on the local neighbourhood will keep the temporal sequence in a comprehensible way and may help to understand the source video's progression. Methods based on clustering however [Zhuang et al., 1998, Yu et al., 2004] might result in keyframes representing the content of a cluster on a global scope.

Underlying Computational Mechanism Several mechanisms for the selection of keyframes have been developed. Truong and Venkatesh [2007] have organized them into eight categories:

- Sufficient Content Change - A new keyframe is created if the change in visual content compared to the previous keyframe is significant [Yeung and Liu, 1995, Zhang et al., 1997, Kang et al., 1999, Zhang et al., 2003, Kim and Hwang, 2002, Xiong et al., 1997, Rasheed and Shah, 2003]
- Equal Temporal Variance - A number or keyframes is set a priori and their distribution tries to balance the visual changes between them [Sun and Kankanhalli, 2000, Lee and Kim, 2002, Divakaran et al., 2002, Fauvet et al., 2004]
- Maximum Frame Coverage - Keyframes are extracted in a way that they provide good coverage for the corresponding clip and a low coverage for the whole video [Chang et al., 1999, Yahiaoui et al., 2001, Rong et al., 2004, Cooper and Foote, 2005]
- Clustering - All video frames are treated as points and standard clustering algorithms are used to determine representative frames, with the disadvantage of loosing the temporal progression [Gibson et al., 2002, Yu et al., 2004, Xiong et al., 1997, Girgensohn and Boreczky, 1999, Zhuang et al., 1998, Uchihashi et al., 1999]
- Minimum Correlation Among Keyframes - It is desired to have the least correlation between the extracted keyframes in order to select dissimilar keyframes [Doulamis et al., 1998, 2000, Liu and Kender, 2002a, Porter et al., 2003]
- Sequence Reconstruction Error (SRE) - Keyframes are selected in a way which allows for an assumed interpolation function to recreate the source video with the smallest error rate [Liu and Kender, 2002b, Lee and Kim, 2003, Liu et al., 2004a, Hanjalic et al., 1998, Lee and Kim, 2002, Lie and Lai, 2005]
- Curve Simplification - All video frames are treated as points in a feature space. A curve, connecting all points in temporal order, is to approximate a simpler curve with the least variation while using the least amount of points [DeMenthon et al., 1998, Latecki et al., 2001, Calic and Izuierdo, 2002]
- "Interesting" Events - Focus on semantically important keyframes, often based on surrounding motion patterns [Liu et al., 2003, Han and Kweon, 2005, Dufaux, 2004, Liu and Kender, 2002c]

Visualization Methods Besides the extraction of keyframes, another field of research is the way in which the results are presented. According to Truong and Venkatesh [2007], *storyboard display* and *dynamic slideshow* are the most common used techniques. In several works, hierarchical presentation models with different levels of detail are described [Lee et al., 2000, Yeung et al., 2001]. Other works use

different layout settings and dimensions of keyframes to display further information [Yeung and Yeo, 1997, Uchihashi et al., 1999].

In order to keep spacial relations between keyframes, mosaic-based approaches allow for showing all of a scene's static components [Xiong et al., 1997, Teodosio and Bender, 1993, Taniguchi et al., 1997]. However, such mosaic approaches depend on the knowledge of camera parameters and matching frame alignment and are thus problematic in real-world scenarios.

3.1.2 Skimming Systems

Basic video skim generation can be achieved by extracting excerpts at predefined points from the source video. Another easy way is to simply increase the playback speed [Omoigui et al., 1999, Peker et al., 2001, Divakaran et al., 2003, Peker and Divakaran, 2004, Christel et al., 1998]. However, Truong and Venkatesh [2007] and Christel et al. [1998] claim that such systems "would seriously degrade coherence" and violate their definition of video skim. They focus on systems which keep the frame rate by extracting excerpts with interesting content. An approach described by Wu et al. [2006, p. 116] is to simply join the surrounding of keyframes into a video.

Similarly to their categorization of attributes of keyframe extraction methods, Truong and Venkatesh [2007] formulated criteria for skimming: *skim length*, *target video domain*, *skim generation process*, *underlying mechanisms* and *features used*.

Skim Length The length of a video skim may be set in advance (a priori) or remain to be determined by the generation method (a posteriori), corresponding to the extraction of keyframes.

- **A Priori** A determination of the video skim's length to be extracted may be set as absolute time duration or as a ratio of the source's length, similar to keyframe extraction. An a priori definition of the skim's length is utilized by He et al. [1999] and Ma et al. [2002].
- **A Posteriori** The final length of a shot depends on the source video's characteristics. Some works use hierarchical generation methods in which the final length is dependent on the individual created units. [Truong and Venkatesh, 2007]

Target Video Domain Almost all published works concentrate on a specific video domain. The following list shows some examples.

- Sports [Babaguchi et al., 2001, Babaguchi, 2000, Marlow et al., 2002, Xiong et al., 2003a, Ariki et al., 2003, Hanjalic, 2003, Coldefy and Bouthemy, 2004, Coldefy et al., 2004]
- News [Kanade, 1998]
- Documentaries [Christel et al., 1998]
- Movies [Pfeiffer et al., 1996, Lienhart et al., 1997, Hanjalic et al., 1999, Sundaram and Chang, 2001, Sundaram and fu Chang, 2002]
- Home Videos [Lienhart et al., 1997, Zhao et al., 2003, Yu et al., 2003]
- Lecture Recordings [He et al., 1999]

Skim Generation Process Although several different works deal with the generation of video skims, Truong and Venkatesh [2007] identified a generic process for the generation. *Excerpt segmentation*, *excerpt selection*, *excerpt shortening*, *multimodal integration* and *excerpt assembly* are the five steps of this process.

Not all of these steps are used in every work. Some may vary, skip or mix them.

- **Excerpt Segmentation** During *excerpt segmentation* the source video is divided into smaller units. These may be shots, scenes, etc. Some researchers claim that this process is an a priori one which is performed separately from the skim generation. A basic approach on video segmentation is the detection of speech pauses via an extracted transcript, as done by Taskiran et al. [2006b]. Other approaches focus on the detection of interesting and important events in the video, as applied by Rui et al. [2000] (in the audio track) and Ariki et al. [2003] (multimodal). Different approaches pursued by Peyrard and Bouthemy [2003] use image motion changes for segmentation, while Cooper and Foote [2002] use the factorization of a self-similarity-matrix.

- **Excerpt Selection** Determining what parts of the source video will be covered by the video skim is done in the step of *excerpt selection*. One approach is to perform a clustering of shots in order to select the longest ones, as done by Gong and Liu [2003]. Other works select the parts to be included into the excerpt by looking for events [Ariki et al., 2003] and also consider factors like a desired skim length for the selection [Babaguchi, 2000]. Another way of creating a video skim is by removing parts until the desired content is left, as done by Ngo et al. [2003]. In Miura et al. [2002], summaries of cooking videos are created by removing shots containing faces.

- **Excerpt Shortening** According to Truong and Venkatesh [2007], the step of *excerpt shortening* aims to result in concise excerpts and avoid a noticeable

information loss. Furthermore, the cut points should be appropriate and avoid, e.g., the middle of a spoken sentence. The viewer should not be irritated and the coherence of the video should stay intact. A simple technique is to define a fixed fragment of each excerpt, like a specific length [Gong and Liu, 2001, 2003, Lee et al., 2004] or a certain time after specific events (slide changes in case of He et al. [1999]). Other approaches select a part which shares the greatest similarity with the complete excerpt by using a self-similarity matrix [Cooper and Foote, 2002]. Sundaram and Chang [2001] used the visual complexity of a shot to determine the time required to comprehend it and thus specify its length. Ma et al. [2002] use an attention curve and associated keyframes to select their surrounding segments for the video skim. In summarizing cooking videos, Miura et al. [2002] focus on parts containing cooking motion or food. User preferences for different semantic categories can be taken into account while shortening, if excerpts are classified accordingly [Zhao et al., 2003]. Other works drop muted parts of an excerpt or remove frames and modify the audio track [Li et al., 2003].

- **Multimodal Integration** Most of the previously described works concentrate on one modality, even if a *multimodal integration* helps to strengthen the skim's coverage, context and coherence [Truong and Venkatesh, 2007]. This can be achieved by two types of combinations of audio, image and text: *modal synchronization* and *modal asynchronization*. Video skims with synchronized modalities have the audio and video stream played in parallel, as in movies, dramas and talk shows. They may be visual-, audio-, or textual-centric, deriving their common timecode from the corresponding track. Truong and Venkatesh [2007] named two methods for assembling modal asynchronous generated excerpts into video skims. They may feature the OR operation, combining all segments that are either audio- or visual-centric [Erol et al., 2003, Agnihotri et al., 2005], or the AND operation. This is especially handy in case of television news and documentaries. For this purpose a modal asynchronous summary may improve the comprehensiveness. Information not present in one stream can be supplemented by other streams, as [Smith and Kanade, 1997] do by adding visual segments into an audio-centric skim. Some works which focus on news summaries only use the anchor audio to better understand the headlines and enrich them with visual overlays [Lie and Lai, 2005, Xie et al., 2004]. When the excerpt borders are based on visual information, the audio stream might start or stop on undesired points, especially in the middle of a spoken word or sentence. Then the borders of the excerpt segments need to

be aligned to the audio stream which can be done by using a transcript [Kim et al., 2003] or detecting breaks in the speech [Taskiran et al., 2006a, Ma et al., 2002].

- **Excerpt Assembly** Joining excerpts based on the temporal order is the most common approach for *excerpt assembly* [Truong and Venkatesh, 2007]. However, an exception is made by Lienhart et al. [1997], who classified excerpts and use those classes to generate movie trailers. A new temporal arrangement may be applied to improve the quality of the generated skims. Several works tried to improve the skim's quality by applying transitions, such as fade and wipe, to join excerpts [Lienhart et al., 1997, Erol et al., 2003] or by utilizing gradual transitions [Pan et al., 2001].

Perspective As a video skim cannot preserve all the information present in the source video, a focus has to be on certain perspectives. Truong and Venkatesh [2007] found three general perspectives in the literature during their research: *information coverage*, *interesting/important events* and *query context and personalization.*

- **Information Coverage** Video skims following the *information coverage* approach are generated to reflect the complete source video by shortening and redundancy elimination, without altering the audience's understanding of the video. This approach is often employed to the domain of information-centric video, such as news broadcasts and instructional videos and focuses on the whole video instead of specific parts. Works following this technique are Lienhart et al. [1997], He et al. [1999], Hanjalic et al. [1999], Sundaram and Chang [2001], Ma et al. [2002], Gong and Liu [2003], Gong [2003].
- **Interesting/Important Events** The *interesting/important events* approach aims to create video skims focusing on certain types of events. This is of special interest for sports videos. Such skims may focus on semantic labels or spatial and temporal attributes, for example goal scoring in soccer [Assfalg et al., 2003], and events around the goal [Wan and Xu, 2004b]. Other skim generation techniques concentrate on evoked reactions, such as cheering and applauding by an audience [Xiong et al., 2003b], and excited speech by narrators or commentators [Marlow et al., 2002, Coldefy and Bouthemy, 2004]. Slow-motion footage is used for detecting interesting events [Pan et al., 2001] and whistling as well as the sound of excitement coming from a crowd in [Tjondronegoro et al., 2004b]. The repetition of events and footage, for example in news, also hints at important parts [Bagga et al., 2002]. Interesting parts of a

video may also be found by scanning for rapid motion and colors intended to attract the viewer's attention [Ma et al., 2002]. Xiong et al. [2003a] identify interesting events through specific patterns in motion, while Radhakrishnan et al. [2004] do so with temporal patterns. Other works incorporate viewing patterns by users [Masumitsu and Echigo, 2000, Yu et al., 2003] or externally captured additional data, such as α-brainwaves [Aizawa et al., 2001].

- **Query Context and Personalization** Video skims may also be generated for *query context and personalization*. These approaches consider user preferences or queries during the generation, such as Christel et al. [1998] with a focus on query matching and audio transcript. Other works let the user apply weights to certain features to reflect their preferences [Lu et al., 2003, Zhao et al., 2003]. This was done especially for sports videos in Babaguchi et al. [2001]. Jaimes et al. [2002] performed automatic learning of user preferences to generate corresponding video skims. Agnihotri et al. [2005] dealt with the creation of adaptive skims that take personality traits into account.

Underlying Mechanisms Among the methods for the generation of video skims, Truong and Venkatesh [2007] identified three basic underlying mechanisms: *redundancy elimination*, *event/highlight detection* and *skimming curve formulation*.

- **Redundancy Elimination** By employing *redundancy elimination*, already known information is dropped from the source video. Only the information required for the viewer's comprehension is preserved [Sundaram and Chang, 2001, Sundaram and fu Chang, 2002]. Among such methods are the selection of interesting portions [Ma et al., 2002] and via clustering [Hanjalic et al., 1999, Gong and Liu, 2001, Gong, 2003, Ngo et al., 2003]. Redundancy elimination from the audio stream can be achieved by removing silent portions and parts containing only background noise. However, redundancy elimination may try to eliminate different parts in the audio and video channel which poses a crucial problem. [Truong and Venkatesh, 2007]

- **Event/Highlight Detection** Video skim generation with focus on *event/highlight detection* not only needs to identify desired parts but also the surrounding context [Truong and Venkatesh, 2007]. Chang et al. [2002] categorize highlights of baseball games via HMMs (Hidden Markov models) into seven different classes. Dagtas and Abdel-Mottaleb [2004] locate events based on keywords and high audio energy as well as on transition from a greenish image (grass) to frames without. Closed-captions and shot template matching

are used by Babaguchi et al. [2001] to detect score changes and corresponding events. The work of Peyrard and Bouthemy [2003] uses only motion to detect sports events. Other approaches use audio, visual and cinematic effects to detect game events, like replay and slow motion [Gu et al., 2000, Pan et al., 2001, quan Ouyang et al., 2003, Tjondronegoro et al., 2004b]. Audible responses to game events, such as excited speech by commentators, or applause and cheering are used as hints in Xiong et al. [2003a, 2004], Coldefy and Bouthemy [2004], Rui et al. [2000].

- **Skimming Curve Formulation** The *skimming curve formulation* approach generates a value describing the probability for each base unit — such as a frame, a shot, a fixed-size audio clip, or even a higher semantical unit — to be used in a skim with a given perspective p [Truong and Venkatesh, 2007]. Li et al. [2003] define story events, such as two-speaker-dialog, multispeaker dialog and hybrid events, as base units. After the curve is calculated, a threshold is set and those parts of the source video which score above this threshold are integrated into the video skim. Some works also search for local maxima of the curve to select excerpts [Ma et al., 2002, Xiong et al., 2003a]. Skims of predefined length can easily be generated using this approach by selecting the excerpts with the best calculated value until the desired length is reached [Truong and Venkatesh, 2007]. This reveals the problem of how to rate the interestingness of a unit. Sports related videos may use the loudness and extent of spectator sounds like cheering, applause and excited speech by commentators as score for interestingness. A more generic approach is to select several features from the base unit and determine its impact on the unit's interestingness via generic models. Such video skims may be personalized by applying weights to certain features.[Truong and Venkatesh, 2007] The overall motion activity, the cutting rate and the audio track's energy are used to determine the interestingness of videos by Hanjalic [2003]. Other works use manual annotated descriptions [Lu et al., 2005] or information about user interaction with the video [Masumitsu and Echigo, 2000, Yu et al., 2003]. Aizawa et al. [2001] directly measure the interestingness by capturing the filming person's α-brainwaves to model the skimming curve. Ng et al. [2002] use a similar approach but chose an SVM classifier to categorize excerpts as being interesting or not.

Features Used The features used for finding useful video excerpts can be grouped into the following categories: *visual*, *text*, *audio*, *visual dynamics*, *camera motion*

and *mid-level semantics* [Truong and Venkatesh, 2007].

- **Visual** Among the *visual* features are dominant colors, edges and textures. Especially in the domain of sports video their spatial relations are important to identify desired excerpts [Chang et al., 2002, Han et al., 2002, Ariki et al., 2003, Assfalg et al., 2003, Shih and Huang, 2004, ching Chen et al., 2004, Dagtas and Abdel-Mottaleb, 2004, Tjondronegoro et al., 2004b]. The similarity between images is often calculated using the color histogram which is especially useful for redundancy elimination [Gong and Liu, 2001, Gong, 2003, Gong and Liu, 2003, Ngo et al., 2003, Lu et al., 2004a,b, Lee and Hayes, 2004]. The human attention model is simulated by Ma et al. [2002] by means of contrast in color, intensity and orientation, while the movie summaries from Pfeiffer et al. [1996] use scenes with high contrast.

- **Text** The inclusion of *text* into the range of features allows for a simple extraction of semantical concepts, in contrast to video and audio [Truong and Venkatesh, 2007]. Text can be gained from the video stream as captions, overlays or inserts [Pfeiffer et al., 1996, Lienhart et al., 1997, Lienhart, 1999]. It may also be taken from separate closed caption streams [Han et al., 2002, Miyauchi et al., 2003, Dagtas and Abdel-Mottaleb, 2004] or it can be extracted using speech recognition [Kanade, 1998, Ariki et al., 2003, Gong, 2003] or taken from external transcripts [Babaguchi, 2000, Babaguchi et al., 2001]. Textual transcriptions allow for keyword spotting [Babaguchi et al., 2001, Babaguchi, 2000, Han et al., 2002, Miyauchi et al., 2003, Dagtas and Abdel-Mottaleb, 2004], or the use of text summarizing methods for video skims [Kanade, 1998, Gong, 2003].

- **Audio** Among the *audio* features used in video skim generation are simple ones found in the audio bitstream, such as scale factor [Marlow et al., 2002], or the audio's energy, sound type, and duration [Ma et al., 2002]. Marlow et al. [2002], Dagtas and Abdel-Mottaleb [2004] derived the interestingness in sports videos detecting increases in audio energy and amplitude. A more common approach is the combination of several low-level features to form models, such as Hidden Markov models or Gaussian mixture models, to detect more generic sounds like excitement [Han et al., 2002, Petkovic et al., 2002, Cai et al., 2003, Xiong et al., 2003a,b, 2004, Coldefy et al., 2004, Coldefy and Bouthemy, 2004, Sugano et al., 2004, Wang et al., 2004, Wan and Xu, 2004a, Tjondronegoro et al., 2004b]. Other works search for more specific sounds [Wang et al., 2004, Ariki et al., 2003, Tjondronegoro et al., 2004a], such as hits in a baseball game.

The identification of speech via primitive audio features and pitch tracking is done by Aoyagi et al. [2003], He et al. [1999]. Another way of finding interesting excerpts in sports videos is to look for unusual sound patterns [Radhakrishnan et al., 2004].

- **Visual Dynamics** Quite often the amount of *visual dynamics* is used to find interesting parts of a video. This was done for movies [Pfeiffer et al., 1996], for sports [Petkovic et al., 2002, Peyrard and Bouthemy, 2003, Hanjalic, 2003, Xiong et al., 2003a, Wang et al., 2004], and for cooking videos [Miura et al., 2002]. In the skimming curve methods, visual dynamics are also used as they are found to direct the attention of viewers [Ma et al., 2002, Ngo et al., 2003]. Visual changes are used to find concepts like slow-motion in replays [Pan et al., 2001, Kobla et al., 1999, Ekin and Tekalp, 2002].

- **Camera Motion** The *camera motion* is of special interest in the domain of sports. It correlates with player movement and thus enables the detection of specific game actions and highlights [Han et al., 2002, Assfalg et al., 2003, Coldefy et al., 2004, Shih and Huang, 2004]. Camera movement patterns can also support user attention models [Ma et al., 2002].

- **Mid-Level Semantics** Besides the features described so far, some approaches do not use those features directly. In Miura et al. [2002], shots containing faces are excluded from the summary as they are unwanted in the source cooking videos. The presence of faces in a shot is used in Lienhart et al. [1997] to group movie sequences into dialog and non-dialog. Furthermore, the presence of faces can be used to model the attention of viewers [Ma et al., 2002]. Access patterns of viewers extracted from logged data are features used in He et al. [1999], Yu et al. [2003]. Other works, such as Jaimes et al. [2002], Takahashi et al. [2005], use MPEG7 event metadata contained in soccer videos. It is also possible to take data from the production into account, such as the spatial source of sounds for meeting videos [Erol et al., 2003].

3.1.3 Evaluation Methods

Evaluation is substantial for measuring progress in video abstracting, but there is no consistent framework to do so. Truong and Venkatesh [2007] point out that every team has its own evaluation system. One reason is that such an evaluation requires an objective ground truth. Furthermore humans have difficulties in comparing the quality of video abstracts. Another reason is that systems developed in other works

are often not accessible or difficult to reproduce which hinders a comparative evaluation. [Truong and Venkatesh, 2007]

In their overview of video abstracting systems, Truong and Venkatesh [2007] grouped the evaluation techniques used so far into *result description*, *objective metrics* and *user studies*.

Result Description An evaluation based on the description of the results is very common and simple, because a comparison with other approaches is not done. The technique is often tested on some videos and the resulting abstracts are demonstrated and described to express if the output is adequate [Truong and Venkatesh, 2007]. Such evaluations usually also feature a discussion of the used parameters or visual dynamics and their impact on the generated results [Zhuang et al., 1998, Hanjalic et al., 1999, Zhang et al., 2003, Yu et al., 2004]. The differences and advantages of some approaches are also sometimes described in comparison to other existing ones [Joshi et al., 1998, Vermaak et al., 2002]. Truong and Venkatesh [2007] argue that such evaluations are not adequate, as they are quite subjective and do not provide much experimental proof on how the method will work in general.

Objective Metrics Regarding keyframe extraction, the fidelity function generated from both the original video and the extracted keyframes is often used as an objective metric. By using this metric, keyframe sets extracted by different parameter sets or different techniques can be compared. Truong and Venkatesh [2007] remark however, that such evaluations are biased or strongly connected to specific viewpoints and techniques and that there is a lack of justification if such a metric corresponds to human judgement. An example for such an evaluation with SRE-based and Semi-Hausdoff methods (see chapter 3.1.1) us described in Liu et al. [2004b], who compared their approach to several others [Wolf, 1996, Zhang et al., 1997, 2003, Liu et al., 2000, Lee and Kim, 2002]. Some works plotted the number of extracted keyframes against the fidelity level [Chang et al., 1999] or "well-distributed" keyframes [Liu and Kender, 2002a]. A groundtruth object mask is used by Kim and Hwang [2002] to determine if important events related to objects are covered by the extracted keyframes.

Keyframes selected by subjects are used as groundtruth to rate the temporal matching of extracted keyframes [de Silva et al., 2005] or prove the superiority of a GMM modeling based technique [Kang and Hua, 2005].

If a skimming system is based on the detection of interesting or important events (as described in chapter 3.1.2), common metrics of information retrieval like precision

and recall can be used for rating the system. The required groundtruth may be constructed by the authors or extracted from skims generated by independent users. [Truong and Venkatesh, 2007]

User Studies The third method of evaluating video abstracts is the use of *user studies*. This requires a group of independent users to determine the quality of a created video abstract. Truong and Venkatesh [2007] consider this to be "probably the most useful and realistic form of evaluation" but also note that user studies are not easy to set up. Such an evaluation for keyframe extraction techniques was performed by Drew and Au [2000], Dufaux [2004], Liu et al. [2003].

For the evaluation of video skims, similar methods have been used. One way is to let the viewers simply judge the quality of summaries and how satisfied they are, or how good they helped in realizing certain tasks (like browsing, searching and content identification) [Lienhart, 1999, Christel et al., 1998, Sundaram and Chang, 2001, Ma et al., 2002, Li et al., 2003, Ngo et al., 2003, Agnihotri et al., 2004]. Ngo et al. [2003] also compare the satisfaction of viewers with skims of different a priori defined durations. Several works [Kanade, 1998, Erol et al., 2003, Yu et al., 2003, Lu et al., 2004b] measured the performance of users while identifying content.

Truong and Venkatesh [2007] conclude that most approaches of video abstracting have achieved satisfactory user acceptance but still lack coherence and smoothness. However, Li et al. [2003] claimed very good results in creating movie abstracts for browsing purposes and Zhao et al. [2003] for home videos.

3.2 Trailer Generation Systems

Among the general video abstraction systems described so far, a few have dealt with this special form of skim generation.

Lienhart et al. [1998] have argued that a trailer is a different kind of media than the previously described video abstracts because the purpose is to attract and advertise, rather than to summarize the source's content. However, Ionescu et al. [2006] defined two different types of video skimming, or in their words, *movie skimming*: One is a *summary sequence*, which tries to cover the entire video, whereas a *movie highlight* focuses on the parts which are most interesting. In their work, a movie trailer is described as a specific case of movie highlights.

The generation of movie trailers bears special difficulties. On the one hand, as Xu and Zhang [2013] noticed, most work in video abstracting has been done for news and sports video which "have well-defined structures and characteristics" [Xu and

Zhang, 2013] easing the summarizing task. In movie trailers, the features need to be on a more semantical level which is why low-level features (color, motion) would not suffice. On the other hand the focus on scenes which are especially exciting, funny, or in another way spectacular can show certain patterns. Such patterns might allow for an automatic production of trailers [Xu and Zhang, 2013]. However, Smeaton et al. [2006] argued that a fully automatic generation of a trailer is hardly possible without a complete understanding of the grammar of movies and trailers.

In the following some works explicitly dealing with the generation of trailers are discussed.

3.2.1 Automatic Generated Recommendation for Movie Trailers

Xu and Zhang [2013] developed an assisting system for selecting clips that can be used in the creation of movie trailers. The creative process of arranging the clips is left to the user. Xu and Zhang [2013] argued that trailers are very subjective and different from other video skims as they require knowledge about the storyline which they do not believe to be automatically extractable.

This system is based on the assumption that a trailer contains some of "the most exciting, funny, or otherwise noteworthy parts of the film" [Xu and Zhang, 2013]. A supervised machine learning algorithm which was trained on professional trailers and corresponding films is used to extract potential trailer clips from movies.

First, they performed a shot boundary detection and extracted keyframes using the method by Wolf [1996] in order to reduce the computational complexity. For the training phase, keyframes from the trailer and from the movie were matched on shot basis using SURF (Speeded Up Robust Features) [Bay et al., 2006]. These matches served as positive training samples. In the experiments performed, about 5-15% of the keyframes were labeled as positive examples and the remaining as negatives. A support vector machine with an RBF kernel was used for the classification of the movie shots. The output of the classification process was a list of the top 100 clips recommended for the trailer generation. This list of clips is presented to the user who can select the ones suiting the desired trailer.

The classification uses several features. A motion analysis based on optical flow is performed to find action sequences with high motion. The system distinguishes between camera and object motion. Faces are detected by means of an Adaboost classifier and Haar features. For the recognition of faces a PCA (Principle Component Analysis) is used on manual selected prototype faces. A sound volume detection calculates the average and the peak sound volumes. A combination of high volume and high frequency is considered to be a scream or hint for an accident. Using zero-

crossing rate and the percentage of low energy frame speech and music are detected. These and other visual and audio features are utilized by Xu and Zhang [2013] for the machine learning algorithm.

An evaluation was performed by means of a statistical analysis using 4 out of 8 movies for cross-validation. The proposed clips were compared to a random selection of clips and showed to be about twice as good regarding precision. The recall was computed on the first 100 candidates and was 15% compared to 1% on randomly selected clips. Furthermore, a manual analysis was performed. The recommended clips are mostly continuous clip sequences, such as gunfire and car driving scenes. Xu and Zhang [2013] explained that such scenes may contain several key features. The system has difficulties in finding so called "flash-and-off" clips and scenes introducing supporting characters. Finally, a user study with 10 participants was performed. In this study the participants were presented with a pool of clips proposed by the system and randomly selected clips. A concatenation of the best scored clips was also performed to create a trailer. The participants of the user study gave higher rating for aspects like "key characters introduction", "magnificent scene" and "overall impression". However, aspects like "support character introduction" got lower ratings, which Xu and Zhang [2013] explained with the focus on the main character. A lack of "plot introduction" was also reported.

The system by Xu and Zhang [2013] aims to support the trailer creation process by proposing potential video clips for inclusion. It requires human input in several steps during the generation process.

3.2.2 Automatically Selecting Shots for Action Movie Trailers

The work of Smeaton et al. [2006] deals with the extraction of audiovisual features and a support vector machine for the selection of relevant shots for trailers. The focus is put on action films to get a homogeneous data base.

Smeaton et al. [2006] explicitly use proven low-level features, e.g. the length of shots, motion intensity and percentages of speech, music, silence, other audio, speech and background music and the percentage of shots containing camera motion. Other mid- and high-level features like face detection, indoor/outdoor or location detectors, as well as color, texture or edge detection algorithms are not used. In a first step, the trailers and movies are segmented into shots using color histogram changes as hints. The audio track is then analyzed by means of High Zero Crossing Rate Ratio, Silence Ratio, Short Term Energy and Short-Term Energy Variation as features using a support vector machine for every category. Similarly, the motion intensity and percentage of camera movement are also computed.

A support vector machine was trained on the automatically segmented trailers. The results of the classification were evaluated using a leave-one-out cross validation. By doing so, one movie out of 6 was chosen as validation subset and the other 5 were used for training. This was repeated with all remaining 5 movies as validation candidates. The results were measured using an *R-Precision* metric which relates the precision to the expected number of results. The average R-Precision in the experiments was 9.21. However these results rely on an accurate shot detection. This is because according to Smeaton et al. [2006] a trailer contains mostly sub-shots of movie shots and the validation treats a selected shot as correct if the groundtruth trailer contains a corresponding sub-shot. If, in an extreme case, the whole movie would be detected as only one shot every selected shot would be a sub-shot and precision and recall would both be 100%.

3.2.3 Automated Production of TV Program Trailer using Electronic Program Guide

Kawai et al. [2007] describe two methods for the automatic generation of trailers for TV programs. This approach uses closed caption (CC) streams, which contain a transcript of the spoken text, and, if available, the introductory text of Electronic Program Guides (EPG) of digital television. The first method requires the EPG text to be present. It tries to locate sentences with the most similarity to the introductory text in the closed captions. The second method tries to find sentences in the closed captions which share the same textual features as a common introductory EPG text and may be used if no EPG is available. Corresponding video segments are then joined to form a trailer.

The EPG introductory text is a manually created textual advertisement for the program, describing its highlights with appealing expressions. By means of a Bayesian belief network, the similarity of the sentences contained in the EPG text and the sentences extracted from the closed captions are computed. The CC sentences with the highest similarity are chosen and the corresponding time codes extracted. Using these time codes, it is possible to select the corresponding clips from the video and audio stream to generate the trailer.

The second approach by Kawai et al. [2007] works without EPG text for the actual TV program. By utilizing a machine learning approach using the AdaBoost algorithm, a classifier is generated. This classifier computes a likelihood for each CC sentence to be part of an EPG introductory text or not. The textual features used are the *total numbers of morphemes*, *the existence of particular parts of speech*, *the existence of particular terms* and *the existence of particular named entities*.

The two methods developed in this work were tested using 10 episodes of a documentary about nature, each of them with a duration of about 45 minutes. The lines extracted from the CCs were preprocessed and the terms with only one occurrence removed. Using the first method, the rarity of terms was calculated by analyzing about 13.000 existing broadcasts from different genres. Foreign terms, names of places and persons and technical terms ranked among high rarity scores, while periods had the lowest ones. The generated trailers using this method were between 21 and 37 seconds long and resembled the story quite well.

For the second method about 10.000 sentences were used as positive and another 10.000 sentences as negative training examples, each obtained from about 500 programs. To set a length for the generated trailers, five sentences with the highest likelihood were chosen from the output of the AdaBoost algorithm. This resulted in trailers with a length of 33 to 36 seconds.

Compared to trailers generated by the first method, these trailers contain no story. However, it is possible to create trailers for programs without EPG data. Kawai et al. [2007] argued that a combination of both methods would be desirable to get a basic story line and enrich a trailer with additional interesting footage.

Furthermore, a comparison of trailers generated by the system with actual broadcasting trailers was done. Precision was calculated by dividing the number of shots present in both the generated and in the broadcast trailer by the number of extracted shots, while recall was, again, the amount of shots present in both trailers divided by the amount of shots in the broadcast trailer. The shots were considered to be the same if their content matches roughly, however different camera angles were tolerated. The values achieved by the first method were 42% for precision and 58% for recall, while the second method reached a precision of 20% and a recall of 32%.

3.2.4 Animation Movies Trailer Computation

In their publication, Ionescu et al. [2006] present a method for the automatic generation of trailers for the specific group of animated movies. A large database of such movies is available from "The International Animated Film Festival"[1]. These animated movies are quite different to other movies: they have a very individual color distribution, they use artistic concepts, and object motion dominates. Ionescu et al. [2006] point out that the understanding of content in these animation movies is especially difficult as about 30% of such films from their corpus have no logical

[1] "Centre International du Cinema d'Animation", http://www.annecy.org

meaning. For the trailer generation, they performed an analysis on the shot and frame levels.

First, the movie is segmented into shots by means of transition detection and short color change (SCC), a specific color effect of animation movies used for thunder or explosions. Based on this segmentation, a *video transition annotation* showing the distribution of transitions throughout the whole movie is computed. By means of an inter-shot analysis, action segments are detected. By performing experimental tests the most attractive shots relate to fast and repetitive shot changes. Corresponding patterns can be found in the previously described video transition annotation as regions with a high density of perpendicular lines.

This inter-shot analysis is followed by an intra-shot analysis, since until now only global action information is obtained. In order to analyze the shot content, a cumulative inter-frame distance histogram is used. In order to reduce computational complexity, every second frame is dropped, the frame resolution is reduced and the color palette decreased. The similarity between histograms is computed by the Manhattan distance. Through an analysis of cumulative histograms on sample animation films, four basic patterns of histograms were obtained. *Small distance histograms* relate to shots with only little changes in color. *Histograms with both small and high distances* indicate shots with a dominant color similarity but also some important changes in color. *Multi-modal histograms* hint on multiple similar color clusters and connecting camera motion, and *single-modal histograms* correspond to many color changes in the respective frames.

The trailer generation focuses on highlighted action segments of the movie, taking into account their corresponding spatial activity. For the evaluation, a user study with 27 participants was performed. The generated trailers rating was very satisfying. The trailers were said to include most of the action parts and have a correct duration.

3.2.5 Video Abstracting

The work of Lienhart et al. [1997] describes the MoCA (movie content analysis) project, a system which is, among other things, capable of generating trailers for feature films.

The approach to video abstracting consists of three steps. First, the movie is segmented and analyzed. In the second step, clips for the generation of the trailer are chosen. For the third and final step, the temporal order and transitions of the video clips are determined and the trailer is assembled.

A basic segmentation into shots is performed by means of the edge-change-ratio

parameter. An additional segmentation in larger units — such as scenes — is done by looking for content similar in color and accompanying audio, and non-matching video and audio cuts. Audio cuts are detected by calculating frequency and intensity spectrums and detecting abrupt changes.

Apart from the segmentation, the movie is scanned for several special events. Among these are the faces of actors in order to locate the protagonists and dialogues. This is achieved by using a proven face detection algorithm in conjunction with a neural network. The neural network was trained on about 1000 manually annotated example face images. In order to increase processing speed, only image regions with skin-like colors in every third frame were analyzed. The detected faces were classified into groups of similar faces in order to get clusters for each character, a so called *face-based class*. By further analyzing the appearances of the actors, shot/reverse-shot patterns and multi-person dialogues can be detected.

Another kind of special event is the appearance of the movie title and the names of actors. Utilizing a text segmentation and recognition algorithm the size of text is determined. Based on the assumption that the title is centered and has the largest font or the longest line, an extraction via OCR is possible. The actor's names are identified by comparing text occurrences with the presence of face-based classes.

Besides these visual features, Lienhart et al. also incorporated an audio feature detection to detect gunfire and explosions. Audio parameters such as loudness, frequencies, pitch, fundamental frequency, onset, offset and frequency transition are analyzed and their distribution is compared against a database in order to find such events.

The generation of the video abstract focuses on the following aspects: *important objects and people*, *action*, *dialogue* and *title text and title music*. In order to not reveal the story's end, only the first 80% of the footage are used for the trailer. A desired maximum target length may be set. The system first selects scenes and then clips as a subset of these scenes. Two mechanisms are used for the selection of scenes and clips. First, special events and texts are extracted from the video. The user may specify the ratio of special events for the abstract with the default being a share of 50%. The title is always part of the abstract. In a second step, filler clips are chosen from portions that are least present in the abstract so far. This is done in order to achieve a uniform distribution of the footage. The clips of a trailer should be much shorter than the movie scenes. The extraction of clips out of scenes is done by two heuristics. The first approach uses shots with most action (motion) and the same color composition as the movie in average. The second approach is to retrieve genre clips characteristics from a database. This data is extracted from previous experiments by the authors and used to select clips matching the desired

genre.

In the final trailer assembly stage, the clips need to be arranged in an order and joined with transitions. The ordering of clips is organized in four classes, in which the temporal order is preserved. These classes are the *event class*, containing special events (gunfire and explosions), *dialogues*, *filler clips* and finally the *texts* extracted from the movie (title and optionally actor names). Clips from the first two classes are put together to form *edited groups*. Using filler clips, the gaps between these edited groups are bridged. The title is shown in the trailer for one second and the names of actors may be added as well.

Transitions between clips can be either hard cuts, dissolves or wipes. Clips of the special event type are joined with other clips using hard cuts, while dissolves and wipes are used to connect calmer clips like dialogues.

Arranging the sound track of the trailer requires more attention, especially during a dialogue. In the MoCA system certain rules are implemented: The audio stream of special event clips is preserved. In case of dialogue sequences, the duration of the audio is adjusted to fill the gaps between special event clips. Between different audio tracks, dissolves are mainly used as transitions. The movie title music track is used as background music and the volume is lowered during dialogue and special events.

In order to evaluate the trailer generation system, some experiments using German television sequences were performed. Similar to other works, Lienhart et al. note that the evaluation of such an abstract is quite difficult. Furthermore, they note that the purpose of the abstract plays an important role. Movie trailers will not reveal the ending but focus on thrill and action while a preview for a documentary would seek to summarize the source as thoroughly as possible. During comparison with professional commercial abstracts, the authors claimed to have found no big differences in respect to quality, however they noted that the professional trailers feature additional music. While summarizing previous episodes of television shows, the tool performed similar to broadcast summaries. A user study showed similar results, as the participants had difficulties deciding which trailers they perceived to be better.

Besides the trailer generation, the system is also capable of generating a website containing an interactive video abstract.

3.2.6 Automatic Trailer Generation

Irie et al. [2010] developed a system for automatic trailer generation called *Vid2Trailer (V2T)*. It follows the basic assumption that a movie must contain *symbols*, such as

the title and the musical theme, and impressive scenes.

The V2T-system performs the trailer generation in three steps: first, the mentioned symbols are extracted. Then impressive segments of the movie are located and finally the trailer is constructed. The title logo extraction algorithm works under the assumption that the title uses the largest font compared to other captions and that the title will appear in the first 10% of the playtime. By looking for the music segment with the longest duration, the theme music of the movie is extracted. Additionally, the melody of the theme is assumed to be repeated throughout the rest of the movie. Employing a melody matching algorithm, the most frequently used melody is detected.

Video sequences containing impressive content are analyzed via *affective content analysis*. By means of features as color, brightness and motion intensity, the affectiveness of the shots is calculated using an eight-dimensional probability vector and a model combining LDA (latent Dirichlet allocation) and a CPT (conditional probability table). About 6 hours and 20 minutes of movie footage were labeled and used for training the model. By employing clustering via *affinity propagation (AP)*, visual affective shots are selected.

Audible information is processed to find occurrences of impressive speech. An assumption made by Irie et al. [2010] is that impressive speech is highly emotional and thus has quite different features compared to "average" speech. A detection is performed by first computing an *average speech model* and then performing a detection of outliers. The parts with the highest outlier scores are then selected, given that their duration is not longer than the chosen theme music.

The generation of the trailer is performed with a focus on maximizing the impression on the audience. Irie et al. [2010] formulate the rearrangement of shots in form of an optimization problem by suggesting a method for assessing the impact of a sequence of shots. This method uses a framework which calculates surprise based on visual information, called the *Bayesian Surprise(BS)*. The selection and distribution of shots is performed in order to maximize the affect impact.

Finally, the title logo, the theme music and the previously extracted impressive speech segments are inserted. The title logo is placed at the end of the trailer, while the title music accompanies the last part of the trailer. In regular intervals, the segments containing impressive speech are laid over the theme music.

The V2T system was evaluated using different methods. The shot arrangement performance was evaluated by first extracting seven to ten shots from 16 movies by means of automatic shot extraction. These shots are then arranged by the V2T system and for comparison by a control group of 12 test subjects who aimed to maximize the impact of the result. Those two resulting sequences were then compared by

average rank correlation, together with a baseline consisting of randomly arranged shots. The results of the V2T system were better than those of the other two.

Additionally, a user study was performed. In this study, trailers generated by different methods were compared. Among these methods were clustering-based video summarization (CVS), attention-based video summarization (AVS), trailers made by V2T as well as real movie trailers. 20 participants judged the quality of four different trailers for each method. Aspects of the evaluation were *appropriateness*, *impact* and *interest* which were obtained using Likert scales. Along all questions, V2T performed better than CVS and AVS but not as good as professionally made trailers. Irie et al. argued that one reason for this is the lack of a narrator in automatically generated trailers.

3.2.7 Semantic Video Patterns in Action Movies

In the student project *Semantic Video Patterns* [Brachmann et al., 2006], a prototypical system for automatic creation of Hollywood-like movie trailers for the action genre was developed. This genre was chosen because dialogue and story information are less prominent in such trailers, compared e.g., to thriller and drama. In a first phase eleven trailers for action movies were manually analyzed to gather knowledge about their structure. Based on this analysis, a *trailer grammar* was defined, describing the syntactical elements of a trailer structure and the rules connecting them.

In a second phase, a prototypical software system was developed to automatically create trailers. This system is combined from two parts. The first part consists of a set of analyzing tools which extract various syntactical features from the video and audio track. The second part, the generator, performs two major tasks. For the first one, it combines previously extracted low-level features to higher semantic ones, and performs a categorization of the shots. As for the second one, the generator infers a specific trailer model and creates a corresponding video file.

This chapter is based on the project report [Brachmann et al., 2006] as well as internal documents and the analysis of the developed software system.

3.2.7.1 Syntax Analysis

Automatic trailer generation covers several fields of expertise. At first, it requires knowledge about how a trailer is constructed, what its elements are and how they are combined. To fill the gap between a film researcher's view on films which is often rather interpretive, and the more low-level and technical analysis by a computer scientist working in image processing, a detailed manual analysis of trailers was done.

The team performed a manual shot-by-shot analysis of the following eleven trailers: *Aliens* (1986, by James Cameron), *Predator* (1987, by John McTiernan), *Die Hard* (1988, by John McTiernan), *Terminator 2: Judgment Day* (1991, by James Cameron), *Leon – The Professional* (1994, by Luc Besson), *Bad Boys* (1995, by Michael Bay), *Blade* (1998, by Stephen Norrington), *The Mummy* (1999, by Stephen Sommers), *Charlie's Angels* (2000, by McG), *Pirates of the Caribbean: The Curse of the Black Pearl* (2003, by Gore Verbinski) and *The Bourne Supremacy* (2004, by Paul Greengrass). Performing the analysis, information about five main questions concerning trailers was supposed to be gathered.

1. What is the source of the trailer's footage? Is it taken from the movie? If

so, from which parts of the movie? Is the trailer a linear short version of the movie? What is the mapping between trailer and movie?

2. What type of content is used? Is it footage, animated text, company logos, a blank screen or credits? Is it additional footage which is not part of the movie? What is happening during the shot?
3. Which characters are shown? Is one of the five most important ones present or not? How much screentime does the hero have? What is the relation to the screentime of the other protagonists?
4. What kinds of transitions are used? Which shot length (see chapter 2.1.7.2) is used? What is the overall lighting?
5. Does the shot contain speech? If so, what is its content? Is it narration or dialogue? Does the shot contain music? What is the loudness?

The data collected was stored in a database. For the evaluation of this database, statistics were calculated and visualized.

By the end of the analysis several common properties of the trailers were found. Most of the shots have a duration of less than 3 seconds, some shots are as short as 1 or 2 frames. The pace of the trailer accelerates from the beginning to the end, meaning that shots become shorter towards the end. The last third of a trailer contains as many shots as the first two thirds.

In all the trailers a similar basic structure of phases was found. Most trailers focus on one or two protagonists which are often shown in close-up shots.

Two tendencies can be observed which correspond to the findings of Hediger [2001] (see chapter 2.2): older trailers tend to have a rather linear structure, while newer ones have a more shattered mapping of shots. In general, however, the mapping tends to be linear. Newer trailers also feature faster cuts and a more complex selection of shots.

The transitions between shots are mostly hard-cuts, especially towards the climax. Transitions to textual animations are combined with special sound effects, in particular when the animation is fast. Sometimes a bright white flash is used for surprise and shock effects.

Most trailers contain a long quote, acting as *break*, before the action starts around the second half of the trailer. In the second half only short cuts are used and the trailer speeds up until the climax. Then the title is shown, slower cuts and a posing sequence follow. Finally the end credits are shown, using a much smaller font than the other text inserts.

A general structure found in all trailers allows a grouping into five phases:

- Intro – introduction of locations, the main characters and a conflict through slow shots
- Story – action and dialogue summarizing the task, medium fast shots
- Break – a dramatic or significant long comment by a protagonist, mostly without accompanying music
- Action – fast action sequences and close-ups of the protagonists, loud sounds
- Outro – calm ending, showing the title, credits and a release date. Some trailers feature a final close-up with a comic or tough comment of a main actor.

This pattern of five phases corresponds to the trailer structure described by Hediger [2001] (see chapter 2.2.3). The *exposition* in his trailer structure corresponds to the *intro* phase. The *story* phase covers the *inciting incident*. The *break* leads over to the *confrontation* which is shown in the *action* phase of the trailer. Finally the film's *title* is shown in the *outro* phase and in some trailers the *button* can be found as well. The results of the SVP analysis also showed the focus on the protagonists, similar to Hediger [2001, p. 36ff]'s findings.

The video segments in movies and trailers are called shots, while the categorized ones used for trailer generation are referred to as *clips*, as defined by Lienhart et al. [1997]. Besides the initial pattern analysis, the next task was to order the footage used in trailers. This was done by defining a set of categories for the shots. Additional categories were defined for the generation of trailers. They cover the leading greenscreen, title- and tagline animations as well as the credits. A list of shot categories used in the SVP system is shown in table 3.1

Footage Clip			Animation Clip
Character1CUSilent	PersonSpeaking	SlowAction	ActorName
Character1CUSpeaking	Quote	Spectacular	CompanyName
Character1Silent	QuoteLong	Shout	Credits
Character1Speaking	Explosion	Scream	DirectorProducer
PersonCUSilent	Fire	Setting	Greenscreen
PersonCUSpeaking	Gunshot		Tagline
PersonSilent	FastAction		Title

Table 3.1 Footage categories defined by the SVP team

These categories reflect aspects found during the analysis and focus on the shot types found in trailers. On one hand, they should fulfill the requirements of an action

trailer and allow building up a convincing structure. On the other hand, they should be technically feasible to perform an automatic categorization given state-of-the-art media processing algorithms.

A trailer can be described by a formal language. The *symbols* of this language are the clips and transitions and the *grammar* describes the rules how they may be combined. A trailer would then be a *word* made up of clips and transitions. The SVP system uses additional *non-terminal* symbols in three layers. First, a trailer is made up of *phases*. These phases are made up of *sequences* which in turn consist of *clip/transition pairs*. A clip/transition pair finally consists of the *terminal* symbols clip and transition. Such sequences allow a more detailed description of the trailer and provide a way to model and reuse subpatterns consisting of several shots and transitions.

The generation requires further information regarding the clips. This is realized by the *properties* of a clip in the SVP systems:

- Category: the category the clip belongs to
- Speed: how fast the clip should be played back (to achieve either slow-motion or speed-up effects)
- Volume: how the sound track of the clip should be processed
- Location: describes the location of the clip in the corresponding movie

Additionally, the start- and endframe are used to precisely address the clip.

A transition can be a *(hard-)cut*, a *fade-black*, or a *flash-white*. Optionally, a transition can be linked to a sound effect.

To summarize, the trailer grammar consists of syntactical and semantical elements. The syntactical ones are the clips — including their categories — and the transitions. The semantical elements are the rules describing the possible patterns of combination and are modeled using a hierarchical trailer pattern tree structure. They can be understood to constitute a trailer mereology.

```
<?xml version="1.0"?>
<movie>
    <general>
        <movie filename="hitch-xvid.avi" checksum="" title="Hitch" tagline="The cure for
        the common man." releaseDate="10 February 2005" frameRate="25"
        totalFrames="170235" genre="Comedy/Romance" company="" awards="Teen Choice
        Award/BMI Film Music Award/Blimp Award/" width="720" height="292"/>
        <director name="Andy Tennant" lastMovie="Sweet Home Alabama (2002)"
        secondLastMovie="The American Embassy (2002)"/>
        <character number="1" name="Will Smith" characterName="Alex 'Hitch' Hitchens"/>
        <character number="2" name="Eva Mendes" characterName="Sara Melas"/>
        <character number="3" name="Kevin James" characterName="Albert Brennaman"/>
        <character number="4" name="Amber Valletta" characterName="Allegra Cole"/>
        <character number="5" name="Julie Ann Emery" characterName="Casey Sedgewick"/>
        <quote number="1" actor="0" quote="Life is not the amount of breaths you take,
        it's the moments that take your breath away."/>
        <quote number="2" actor="2" quote="What should we toast to?"/>
        <quote number="2" actor="0" quote="Never lie, steal, cheat, or drink. But if you
        must lie, lie in the arms of the one you love. If you must steal, steal away
        from bad company. If you must cheat, cheat death. And if you must drink, drink
        in the moments that take your breath away."/>
        <quote number="3" actor="0" quote="No, I was told that you help guys get in
        there."/>
        <quote number="3" actor="0" quote="Right, but see, here's the thing - my clients
        actually *like* women. &quot;Hit it and quit it&quot; is not my thing."/>
```

Figure 3.1 XML description showing the results of the analyzing step

3.2.7.2 Analyzer

The automatic generation of trailers depends mainly on a good categorization of the footage. In the SVP system, this process consists of two steps. First a group of *detectors* analyses the movie file — its video and audio part — and extracts low-level features. In a second step, low-level features are used to model higher-level semantical categories to describe the shots. The detectors are individual programs and return their results as an XML description. An excerpt from such a file for the movie "Hitch" is shown in figure 3.1.

This chapter describes the available detectors used in the first step and lists the features these modules can provide. For each detector, the provided features and a description of the algorithm is given. The input for the whole system is a video file in AVI format with an MPEG4 (Xvid[2]) video track and a stereo MP3 audio track. Additionally, the movie title is required by the general movie information module.

[2] https://www.xvid.com/

General Movie Information

Feature Name	Type	Comment
Title	String	
Tagline	String	
Release Date	String	
Frame Rate	Float	*frames per second*
Frame Amount	Integer	
Genre	String	*according to the IMDb*
Company	String	*according to the IMDb*
Awards	String	*according to the IMDb*
Width	Integer	*in pixels*
Height	Integer	*in pixels*
Director	String	*according to the IMDb*
Last Movies	String	*two last movies from this director*
Character Id	Integer	*for the five main characters, 1-5*
Character Name	String	
Actor Name	String	
Quote Id	Integer	
Quote Character	String	*0 means unknown*
Quote Transcript	String	

This detection module provides metadata about the movie. As the system is intended to work automatically, dependencies on manual data input should be reduced to a minimum. This is achieved by utilizing the *Internet Movie Database (IMDb)*[3] for querying various information about the movie, the director, actors as well as famous quotes from the movie. This module is realized as Python script and uses the IMDbPY[4] package to access the IMDb.

The script requires the movie name as parameter and returns a list of possible matches found. A confirmation is required before the movie data is finally retrieved.

Besides the information from the IMDb, this module extracts technical aspects of the video file. Among these are the file name, the frame rate, the total amount of frames as well as the width and height of the picture. This data is provided by the face detection/recognition module (see chapter 3.2.7.2) and merged into the

[3]Internet Movie Database: http://www.imdb.com/
[4]http://imdbpy.sourceforge.net/

resulting XML description.

Shot Detection

Feature Name	Type	Comment
Shot Number	Integer	
Start Frame	Integer	
End Frame	Integer	
Transition Type	String	
Transition Duration	Integer	duration of the transition in frames
Transition Histogram Change	Float	total change in grey histogram
Average Color	Integer	*red, green and blue values*
Color Variance	Float	
Average Brightness	String	
Brightness Variance	Float	

The shot detection module is used to segment the movie into shots. Furthermore it calculates color values of a shot.

This module incorporates a tool previously developed at the TZI by Miene et al. [2001]. It calculates the differences of the Gray Histogram X^2 between adjacent frames and comparing them with given thresholds to detect shot boundaries. This tool was modified to provide the desired output. By performing a batch run with different thresholds and a comparison to a manual shot annotation of the movie *The Transporter* the two thresholds ($T_{diff} = 7$ and $T_{conc} = 7$) were determined. The shot detection module first runs the TZI tool with these thresholds. Then the average colors and standard deviations are calculated. All these results are then added to an XML description.

In an evaluation against the manual annotation of *The Transporter*, precision and recall were both at 0.93. In the current implementation the module is only capable to detect hard-cuts but this was decided to be sufficient as 99% of the cuts in *The Transporter* are hard-cuts as well.

Movement Detection

Feature Name	Type	Comment
Start Frame	Integer	
End Frame	Integer	
Direction	String	
Magnitude Absolute	Float	
Magnitude Normalized	Float	*normalized regarding the current movie*

The movement detection module segments a film based on movement. Frame ranges sharing a similar intensity of movement will be grouped together. Furthermore the information about the movement of the magnitude (see table 3.2) is normalized and supplemented by the average movement direction.

magnitudeAbsolute	magnitudeClass
0.000000 - 0.000001	no movement
0.000100 - 0.005000	nearly no movement
0.005000 - 0.040000	very little movement
0.040000 - 4.000000	little movement
4.000000 - 10.00000	medium movement
10.00000 - 30.00000	medium-fast movement
30.00000 - 60.00000	fast movement
60.00000 - MAX	very fast movement (only very few frame ranges)

Table 3.2 Absolute Movement Magnitude Ranges (Source:[Brachmann et al., 2006])

The motion is identified by using an optical flow based method. The open source image processing library *OpenCV*[5] provides several functions for these purposes. A pyramidal Lucas Kanade method [Bouguet, 1999] was chosen to calculate the optical flow by finding and tracking strong corners in adjacent shots. The average of the movement of each feature is then used as indicator for low or high movement. This allows grouping frames with similar movement magnitudes together. Frame ranges shorter than a pre-defined threshold are combined into one range.

This module was not evaluated against a manual annotation. Instead, a visual overlay onto a movie file was produced, showing the detection motion. The comparison of this representation with the actual film showed satisfying results.

[5]Open Source Computer Vision Library: http://opencv.org/

Face Detection/Recognition

Feature Name	Type	Comment
Start Frame	Integer	
End Frame	Integer	
Actor Id	Integer	
Size	Float	
Face Absolute	Float	
Face Normalized	Float	
Actor Absolute	Float	
Actor Normalized	Float	

This module covers two aspects. First, it detects frame ranges containing faces. Second, a recognition is performed on the faces found.

The face detection is based on the work on Haar-like features by Viola and Jones [2001] and Lienhart and Maydt [2002]. Such methods use multiple sub-classifiers for the classification of objects. The face detection module uses an implementation provided by the OpenCV Library. This implementation returns rectangular regions of the image that might contain faces.

In order to improve the results, these rectangles are tracked over several frames. This is done because it can be assumed that a character will most likely be shown for more than a single movie frame. A focus was put on larger faces, as close-up shots are typically found in a trailer. Additionally, the face detection was optimized on delivering a better recall and not so much on precision. Compared with a manual annotation of *The Transporter*, the face detection module has a precision of 0.8.

To be of better use for automatic trailer generation, this face detection module has been extended to be able to recognize faces that were previously detected. This is motivated by the wish to determine the importance of the character and to identify the protagonists. The approach of the SVP team was to utilize a Principle Component Analysis (PCA) in order to select proper facial features from the analysis and perform a clustering via k-means. The results of this module were rated as satisfying, although the results highly depend on the individual movie and training data.

In conclusion, the face detection is considered to be quite robust, however the clustering is not very stable.

Sudden Volume Change

Feature Name	Type	Comment
Start Frame	Integer	
End Frame	Integer	
Intensity Absolute	Float	
Increase	Float	

The soundtrack of a movie may contain information about interesting shots as well. An abrupt increase in volume can hint at a surprising element in the film. Such an event is defined by a time span with a loud sound level that is preceded by a quite one.

This module uses the soundtrack and performs a Fast Fourier Transformation (FFT) in order to calculate the energies in the result array. The processing is done frame-wise, resulting in the use of sound fragments with a length of 0.04 seconds.

When a raise in volume is identified, the average volume before and after this raise is calculated. If two criteria are fulfilled, the raise is counted as volume change. These are:

- the average volume after the raise has to be above a certain threshold
- the raise must be bigger than another given threshold

This simple module performs quite well in detecting spectacular events, like crashes and explosions, although sometimes other non-interesting events are found as well.

Sound Volumes Detection

Feature Name	Type	Comment
Start Frame	Integer	
End Frame	Integer	
Level Absolute	Float	
Level Normalized	Float	

The purpose of this module is to calculate the overall volume of a movie and to segment it into regions of similar loudness. The audio is processed in chunks of the duration of one frame. Then the individual energy is calculated using an FFT. A logarithmic scale is applied to estimate human perception of sound. Furthermore

the results are smoothed to get an estimated volume curve. A new region is started when:

- a turning point is reached by the smoothed curve
- a new and different stability point is reached, compared to the previous level
- the current region exceeds a certain length

The borders between the regions are set in a way that the difference between smoothed curve and actual intensity is smallest.

This modules calculates the volume of the movie's soundtrack in a way that is comparable to human perception. It is robust against short, loud clicks and high amplitude infra-sound frequencies. The volume regions have a length between 1 and 10 seconds and the smoothed adaptive determination of the region's borders prevent fragmentation.

Music Detection

Feature Name	Type	Comment
Start Frame	Integer	
End Frame	Integer	
Probability Absolute	Float	
Is Disturbing	Float	

An important factor while using movie footage to generate a trailer is the detection of background music. Since a trailer features its own mixed score, sequences containing disturbing music must be marked as such in order to not mix existing music from the movie with the newly created soundtrack. Such disturbing music should have a rhythm (beats or melody), higher volume than other sounds and contain a lot of energy in frequency bands where human speech is normally not found in. Furthermore, speech should be clearly audible. Thus it is important to know if those shots contain music or not.

The approach by Hawley [1993] and Minami et al. [1998] analyzes the power spectrum for peaks of stable frequency. Such peaks are visible as horizontal lines which are detected by using image processing methods. The module evaluates frequencies between 0 and 4000 Hz and uses chunks of 0.1 seconds. By using a Fast Fourier Transformation, the spectral images are calculated. The FFT uses a window length of 2048 samples and overlap of 512 samples. The spectrum image is drawn with about 10 seconds of audio and contains 371 frequency bands.

In order to find the horizontal lines, the image is heavily blurred horizontally beforehand. Following the proposal of Minami et al. [1998], a Canny edge detection algorithm is used after this preprocessing. In contrast to the approach by Minami et al. [1998], a binary image is used for further calculation.

Two thresholds are used, *Tlines* for a minimum length of a line and *Tint* for the minimum line intensity of the music. In an evaluation the optimal values of 450 pixels and 16 pixels have been determined. When speech masks music, the line intensity drops.

Finally, the position regarding the movie timeline is calculated along with the disturbance factor of the music. When the value *isDisturbing* is greater than 1.0, the music is considered disturbing with respect to trailer generation.

Character Audio Detection

Feature Name	Type	Comment
Start Frame	Integer	
End Frame	Integer	
Actor Id	Integer	*currently not implemented*
Speech Absolute	Integer	
Speech Normalized	Float	
Actor Absolute	Float	*probability, currently not implemented*
Actor Normalized	Float	*probability, currently not implemented*

Similar to the music detection, the purpose of this module is to find the frame ranges that contain speech in the audio track. The actual words spoken are not important. The result is a segmentation of the movie into parts with speech and parts without.

This is achieved by utilizing the CMU Sphinx speech recognition[6] together with the pre-trained HUB4 acoustic models and the AN4 3-gram language model. First the format of the soundtrack is adjusted to 16 kHz to match the model and segmented into parts with a length of 2 minutes. A FFT with a window of 25 ms length is used to extract Mel-frequency cepstra vectors. This window is then shifted by 10 ms. [Huang et al., 2001]. Finally, frame ranges shorter than 18 frames are removed.

The character audio detection was evaluated with *The Transporter* and provided a precision of 0.79 and a recall of 0.77. Problematic content contains singing and rap which are close to normal speech and can be detected as speech. It is also possible

[6]CMU - Carnegie Mellon University Sphinx: http://cmusphinx.sourceforge.net

to identify dialogue structures by looking for breaks with a length shorter than 50 frames between two speech parts.

Shout Audio Detection

Feature Name	Type	Comment
Start Frame	Integer	
End Frame	Integer	
Probability	Float	

This module is used to find parts of the movie in which a character is shouting. This is done by combining the output of the sound volumes detector (see chapter 3.2.7.2) and the character audio detector (see chapter 3.2.7.2).

Therefore, the shout detection depends on these two modules to be run in advance. It takes the two resulting files containing the XML descriptions as input and compares their found ranges. Every character audio event is checked if it is inside a certain sound event frame range. If the normalized sound level is above a certain threshold, in this case 0.5, the event is considered a shout and marked accordingly.

Applied to *The Transporter*, the shout detection achieved a precision of 0.5 and a recall of 0.15. Screams are often classified as shouts, resulting in half of the false positives being screams. However, even humans have sometimes difficulties to differentiate screams and shouts. The SVP team defined shouts as utterance caused by anger, by giving orders or having emotional outbreaks in contrast to screams which are caused by fear of scared characters.

Sound Detection

Feature Name	Type	Comment
Type	String	*each type has its own section in the xml output*
Start Frame	Integer	
End Frame	Integer	
Probability	Float	

The sound detection is able to find different kinds of sound events in an audio track. Currently, such events can be gunshots, explosions, crashes and screams. The approach is similar to the idea of Hoiem et al. [2005]. However, a support vector

machine (SVM) is used instead of a decision tree classifier and the feature set is different to increase robustness against volume changes.

Depending on the event of interest, the length of the audio chunk to be processed in one step may vary. The module uses 2000 ms for explosions, 1200 ms for screams, crashes and machine gun sounds and 800 ms for individual gunshots. First, features are extracted. For this step a simplified spectral representation is calculated. A sample is taken every 11.5ms, resulting in 550 samples per spectrum and approximately 87 slices per second. Every slice is divided into 17 bands between a frequency of 50Hz and 9550Hz. On this representation 63 descriptive features are calculated to describe the energy distribution across the 17 bands. This results in a vector of 63 float values which are then used for the classification.

For each sound about 10 to 20 training samples are used to prepare the SVM classifier, along with hundreds of negative samples. After an evaluation of the results, the epsilon-SVR algorithm was chosen, with gamma values ranging from 0.01 to 0.045 and cost values of 100. To provide realistic training data, sample sounds were cut manually from the movie's audio track.

In order to improve the results, the threshold of the predicted vectors has been lowered. This yields in more results being found, among the positives are now negative results up to -0.4 and lower. To prevent too many false positives, an additional filtering was introduced which uses the overall loudness of the movie as support. By combining these features, only sound events with a certain volume are detected and the overall detection rate is improved. The detection of explosions is supported by the visual explosion detection module (see chapter 3.2.7.2).

Overall, this module delivers good results. Compared to the manual annotation of *The Transporter*, a precision and recall of both 0.4 were achieved by the gunshot and machine gun shot detection. The detection of screams had a precision of 0.33 but only a 0.1 recall rate. Explosions were detected with a precision of 0.82 and a recall of 0.18. Here most false positives consisted of either music with drums or machine gun fire. Other forms of destruction were also found frequently. The detection of crashes was not very reliable, suffering from very distinct crash sounds in the training set. The differences between a car crash, breaking glass, fighting and dark rumbling were quite big. From the 95 positives reported for *The Transporter* only very few can be considered crashes. Most were music beats, fight scenes, explosions or gunfire.

A crucial factor is the selection of training sounds. The manual selection and preparation is a complex process but further training is likely to improve the detection rate.

Visual Explosion Detection

Feature Name	Type	Comment
Start Frame	Integer	
End Frame	Integer	
Probability	Float	

The automatic detection of explosions is quite difficult when based on audible features only. An additional feature was implemented to further refine these results. This module is based on the approach described in Rasheed et al. [2003]. In contrast to several other approaches they added visual features to the detection process. However, the approach by Rasheed et al. [2003] turned out to have either good precision and bad recall, or vice versa. The SVP team thus refined the algorithm. They performed an histogram analysis in different color models (RGB, L*a*b* and HLS) and looked for characteristic patterns of explosions.

While Rasheed et al. [2003] used gray level histograms, this module uses a histogram for every color channel of each color model. This method improved the results of the approach by Rasheed et al. [2003] on gradual brightness changes or camera movement.

Although the improved algorithm performs better, it alone is not suitable for the detection of explosions. Shots with reddish and saturated colors are often detected as false positives. However, if used in post-processing the output of the sound detection (see chapter 3.2.7.2) can successfully filter out false positives.

Text Detection

Feature Name	Type	Comment
Start Frame	Integer	
End Frame	Integer	
Transcript	String	*not implemented*
Probability	Float	

The presence of text in a shot can be disturbing during the generation of trailers. Especially shots showing company logos, actor names, or the credits can interfere with newly made text animations.

This module uses an existing tool created by Wilkens [2003] to mark the frame ranges containing text. A disturbing text was defined as text containing at least three

characters. After an evaluation, a 3x3 edge filter was chosen. As the adaptation of the module became troublesome, the SVP team wrote a converter for the output of the existing tool and used these two programs successively.

Tested on *The Transporter*, the module produced a precision of 0.92 and a recall of 0.78. However, the team noted that these values might suffer from an imperfect manual annotation.

Quote Detection

Feature Name	Type	Comment
Quote Id	Integer	
Start Frame	Integer	
End Frame	Integer	
Actor	Integer	
Transcript	String	
Probability Absolute	Integer	
Probability Normalized	Float	

Important parts of a trailer are quotes from the protagonists (see chapter 3.2.7.1). Among other data, the general movie information module (see chapter 3.2.7.2) retrieves a list of such quotes and the characters speaking them. The quote detection performs keyword spotting on the audio track of the movie to find these quotes.

This module processes audio in windows with a length of 20 seconds and shifts the windows by 2 seconds. The previously retrieved quotes are used to build a language model using the CMU-Cambridge Statistical Language Modelling toolkit [Rosenfeld and Clarkson, 1997]. The text-to-phone program addttp4 [Fisher, 2000] is then used to build a word-phoneme dictionary.

The audio track of the movie is down-sampled to 16kHz to be compatible to the acoustic models and split into the 20 second windows. On these 20 second audio chunks a Fourier-transformation is performed and 39-dimensional Mel-frequency cepstra coefficient (MFCC) vector is extracted to perform the speech recognition. The CMU Sphinx 3.5 speech recognition is used together with the HUB4 acoustic models as well as the previously created language model and dictionary to spot the quotes in the audio track. Finally, the most probable frame ranges for each quote are written into the output file.

The output of this module was evaluated on *The Transporter* as well. Both precision and recall were at 0.67. The module performs quite well, even false positives

turned out to be usable for the intended purpose in a trailer. However, the detection rate depends on the quality of the quotes from the IMDb.

3.2.7.3 Generator

After the analysis step, the features and metadata are stored in one XML file per movie. This requires the time-consuming feature extraction to be run only once per movie even if new semantical concepts would be introduced later on. In the next step the XML file and the corresponding video file of the movie are used as input for the automatic generation of trailers. This is done by the generator module.

The generator performs two main tasks. First, the extracted features of the movies are used to *categorize* the footage into the predefined shot categories. Second, these clips[7] are arranged according to an inferred trailer model and combined with textual animations, sound effects and music tracks into a video file, the resulting trailer.

The module core is written in C++ and consists of a controller part and several submodules. Besides this, the knowledge about categorization and the trailer grammar (see chapter 3.2.7.1) is separated from the actual C++ implementation and stored in a knowledge-based system. The expert system is implemented using the CLIPS[8] software and a formal grammar to store knowledge. Additionally, external applications are used for 3D-animated text and video processing.

Categorization In the categorization step, the footage is grouped into pre-defined categories. This is done using the features and metadata gathered in the analyzing step (see chapter 3.2.7.2).

The categories are defined using the CLIPS framework and the Protégé[9] editor. A *category* is defined by several *category parameters*. One category parameter defines valid ranges for one feature. Only if all the category parameters match a certain shot's attributes the shot is added to the corresponding category. Figure 3.2 illustrates an example: Each feature is represented on a track along the movie timeline. In this case the category "explosion" has five category parameters: the feature *Explosion_audio* should have a value between 0.7 and 1.0, the feature *Explosion_visual* between 0.5 and 1.0, the feature *Text* should be -1 (no text visible), the feature *Music* should be between 0.0 and 0.2 (almost no music audible) and the result should

[7] For better understanding, the *categorized* shots of the movie are called *clips* [Lienhart et al., 1997].

[8] CLIPS: http://clipsrules.sourceforge.net/

[9] Protégé 3.5: http://protege.stanford.edu/

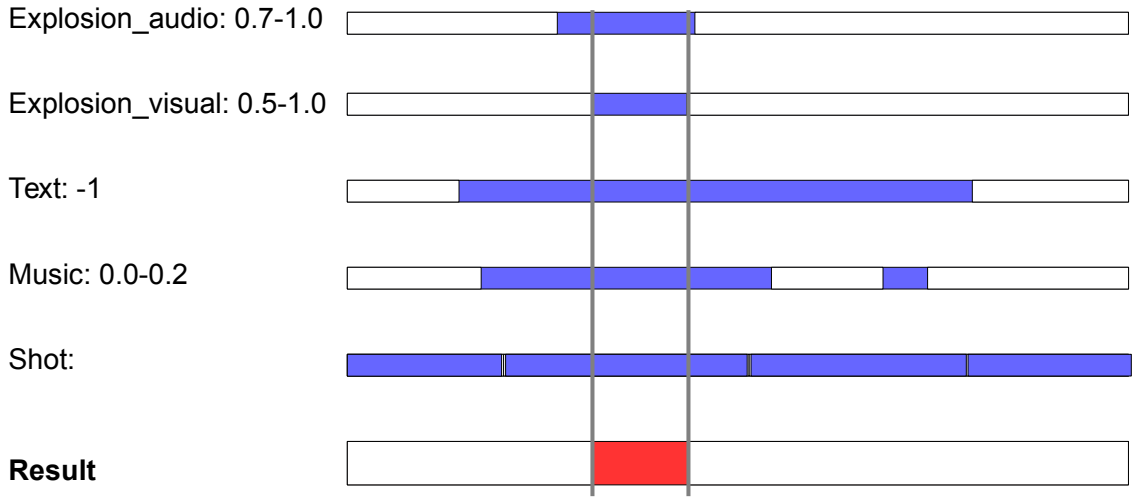

Figure 3.2 Categorization Example, frame ranges containing the feature are marked blue, the categorized frame range is marked red

be within a shot. In the example, all conditions are met, so the corresponding shot is categorized as being a clip of the category *explosion*.

However, a clip doesn't have to last as long as a shot. In this case, only the frame range containing the desired features is used to determine the start and end of the clip. A track may also be inverted, meaning that a frame range should be excluded if the category parameter matches.

At the end of the categorization step, the pre-defined categories (see table 3.1) should be filled with the corresponding footage from the movie.

Generation Once the movie footage has been prepared by the categorization, the generation process can start. This process is structured as follows: First, a trailer model is derived from the hierarchically structured trailer grammar. Then, corresponding clips are selected from the categorized movie footage. Additional animated clips, such as the leading greenscreen, actor and director names as well as the credits are rendered. Sound effects accompanying the animations and a music score are arranged to complete the trailer. Finally, all these trailer elements are combined and rendered into a video file to create the actual trailer.

Trailer Grammar As described in chapter 3.2.7.1, each trailer has a basic structure consisting of intro, story, break, action and outro. The trailer itself is treated as a formal language. The terminal symbols of this language are video clips and transitions, while the non-terminal symbols are phases, sequences and clip/transition pairs. The trailer grammar describes the rules which define the possible arrangements of these symbols to form words, respectively a trailer.

The trailer grammar is implemented by the SVP team as a formal grammar.

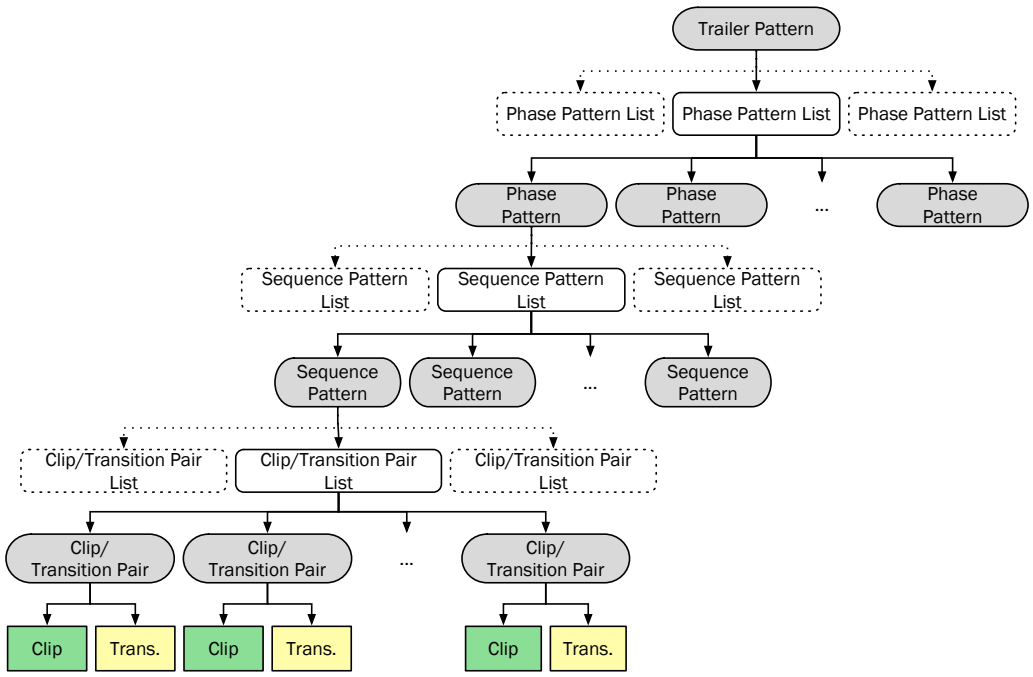

Figure 3.3 Hierarchical Trailer Model of the SVP System, graphic taken from Brachmann et al. [2006]

The formal grammar has the following non-terminal symbols, numbered according to their hierarchy (see figure 3.3 as well):

1. TrailerPattern
2. PhasePattern
3. SequencePattern
4. ClipTransitionPair

The starting symbol is the trailer pattern. A trailer pattern might be an *action trailer*. It may consist of a phase pattern containing the phases intro, story, break, action and outro. These *phases* can consist of several *sequence patterns*, e.g., *setting introduction*, *character introduction* etc. Such a sequence pattern may in turn contain several *clips* with *transitions* in between, e.g., a clip from the category "setting", followed by a clip from the category "Character1CUSilent".

In order to allow variation and some randomness, a list of possible subpatterns can be inserted at each hierarchical level. This allows to have, e.g., two different "setting introduction" sequence patterns. During the creation of a trailer model, one of these alternatives is selected randomly from the list. This allows the system to create slightly different trailers in every run.

Upon request, the knowledge-based system generates a *specific trailer model* according to the rules stored in the trailer grammar. This trailer model is a list of desired clips and transitions.

Clip Selection After a trailer model is created, corresponding footage clips must be chosen. The trailer model contains information about what kind of clip is desired. Additionally, a preferred location may be requested by the trailer model.

This module also keeps a black list of already used clips. This is required to avoid using a clip twice in the same trailer.

3D Text Animation As shown in chapter 3.2.7.1, a trailer contains more than just footage from the movie. Additionally, artistically animated text is used. For example the movie title, famous actors, or the director's name and previous renowned works.

To successfully generate a Hollywood movie trailer, these animations must be included. The textual content required for the animations is already available (see chapter 3.2.7.2). The animations are done using the 3D modelling and animation software Blender[10]. A submodule of the generation software gathers the desired text and chooses either an assigned or selects an *animation style* by chance. Four of such animation styles were created by the SVP team: *fast zoom out*, *white bomb*, *metal flash* and *soft zoom in*.

The animation module determines the desired width, heights and length of the requested animations and starts Blender with these values, the chosen style and the desired text to create the actual animation. The resulting clip is then treated in the same way as clips based on movie footage.

[10] http://www.blender.org/

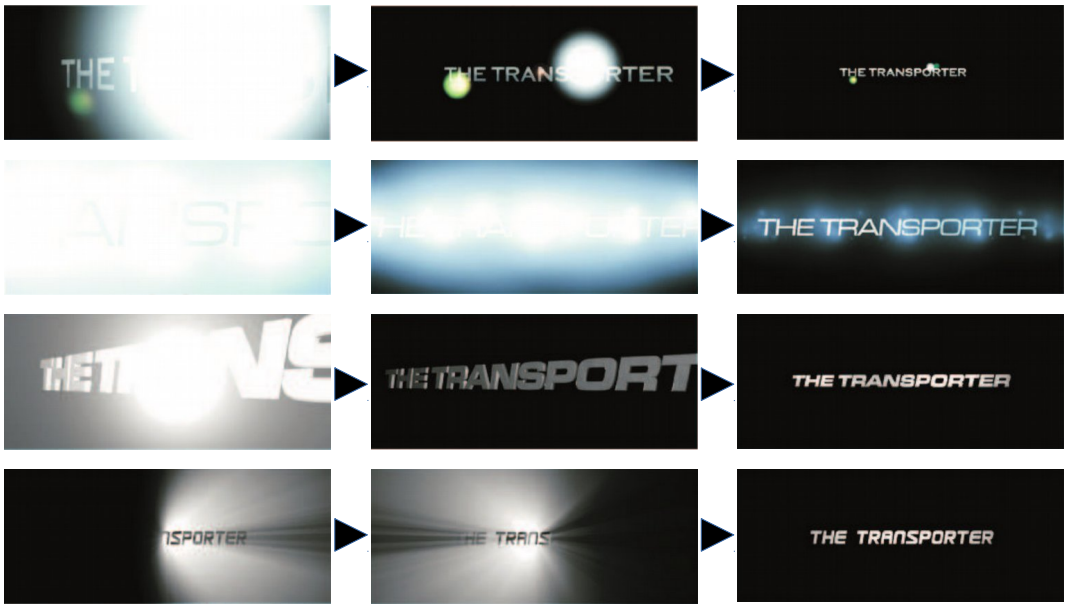

Figure 3.4 Different animation styles used in the SVP system, image taken from Brachmann et al. [2006])

Sound Effects and Music Besides the movie footage and visual animations, an important part of a trailer is the audio track. In addition to the sound from the movie, sound effects — accompanying the animations or cuts — and typical trailer music is necessary.

The SVP system comes with a sound archive containing both types. Three different kinds of sound effects are used:

- "woosh" – used for text animations, simulates high-speed passes
- "wooshbang" – used for title animations, adds tension and strong beat
- "boom" – may be used with a white flash, has a surprising effect

The SVP team first wanted to extract music from the movie but this approach did not work as special music is needed. Instead they selected typical trailer music and created a pool containing several snippets for each part of the trailer. They had certain requirements for the phases:

Phase	Shots & Cuts	Music
intro	slow cuts, long (often faded) shots	calm and cool mood, mystery feeling. slow beat, long and dark, vibrating sounds. high tension is often provoked through raising tones
story	motion and life, quotes, medium long shots, hard cuts	dynamic, non-disturbing music. medium fast drums. raising sounds and beginning tension as well
action	motion, explosions, gunshots, high dynamic	powerful, energetic and rough music. very fast beats, strong drum sounds, combination with guitar sounds
outro	climax of tension risen throughout the trailer, title is shown	tension should be brought down again. music starting with short and dramatic elements and long fall down, often orchestral flourishes, echoed and faded out. dark vibrating sounds, dark ambient sounds to prevent silence in the end

Metadata about the sound files is stored in an XML database. This data contains the exact duration of the sound, the allowed phases and the physical file name.

Some of the sound effects depend on time synchronicity with text animations or cuts. To achieve this, every sound effect file has it climax after 1 second. For

technical reasons, all audio files are stored as uncompressed wave-files with 48kHz and 16bit.

Rendering Up to this step, all footage and additional material for the trailer has been prepared. A list of clips is available, textual animations are rendered, sound effects and music snippets are selected.

The generator of the SVP system performs the final rendering of the trailer into a video file in two steps. First, a set of scripts for the external programs is written. Among these are AviSynth scripts controlling the conversion of the previously created 3D animations, as well as the main AviSynth script containing all the cutting and mixing instructions for the media files assembly. Other scripts control the video processing tool VirtualDub[11] which is used in conjunction with AviSynth to create the final video file. Finally, these scripts are executed and the video file is generated, which is the resulting trailer.

Outcome and Conclusion The SVP system is able to generate Hollywood-like trailers. In order to determine their quality, an evaluation was performed. A total of 59 persons were invited to watch and judge a set of seven trailers. A paper-based questionnaire was used to collect their feedback.

The following trailers were shown:

1. Two professionally produced Hollywood movie trailers: *War Of The Worlds* (2005, by Steven Spielberg) and *Miami Vice* (2006, by Michael Mann)
2. A trailer produced using the automatic video generation software Muvee[12]: *The Transporter* (2002, by Corey Yuen and Louis Leterrier)
3. A trailer consisting of randomly selected frame ranges: *Bad Boys* (1995, by Michael Bay)
4. A trailer build by the SVP system using no knowledge about patterns: *Blade* (1998, by Stephen Norrington)
5. Two of the best trailers created by the SVP system: *The Transporter 2* (2005, by Louis Leterrier) and *Terminator 2* (1991, by James Cameron)

The test persons had no knowledge about the production process of the trailers except for the last one, as for this trailer questions specifically aiming at the generation quality were asked. First, all the trailers were shown to the participants, later on

[11] VirtualDub: http://www.virtualdub.org/
[12] Muvee: http://www.muvee.com/

they were able to watch them again if they desired. The following table shows the overall average score of the trailer evaluation (score possible between 1 and 10):

Trailer	Average Score	Source
War of the Worlds	7.86	*professional trailer*
Miami Vice	4.92	*professional trailer*
Transporter	3.97	*Muvee*
Bad Boys	3.61	*random shots*
Blade	3.52	*random patterns*
Transporter 2	7.29	*SVP trailer*
Terminator 2	7.26	*SVP trailer*

The results show that the trailers with a random shot selection were rated worst, and furthermore, the best automatically generated trailers by the SVP system could be competitive with professional Hollywood trailers.

Whether the participant has seen the corresponding movie or not had no or just little influence on the judgment. A more detailed analysis showed that the most problematic parts of the automatically generated trailers are the introduction of people and the missing storyline. However, it was also mentioned that the viewers made up their own storyline in their minds corresponding to the clips of the trailer.

Chapter 4

A Multi-Genre Approach to Movie Trailers

As the source code for the previously described SVP system (see chapter 3.2.7) was available, it was a rational choice to start the work by this existing system. Starting from the capability of generating action trailers, the aim in this thesis is to extend the system to be able to take different genres into account and also allow for a generalization to other types of video abstracts.

The first step includes the selection of further genres to experiment with and a manual analysis of the corresponding trailers. Further steps would have been the adaptation of the existing system for these additional genres. In order to simplify the process a new video abstracting application was developed. In comparison to the existing SVP system, the focus was set onto interaction to ease editing semantical concepts, add an optional video editing step as well as tuning the movie segmentation and abstract generation processes.

4.1 Analysis of Hollywood Trailers

Since the categorization of films into genres can be rather ambiguous (see chapter 2.1.8), the new genres to be chosen should be well known and accepted. In the ADDiCT project [Asaad et al., 2008], a selection of movies of different genres had been used. Among these genres were action, comedy, drama and horror.

4.1.1 Genre Selection

The genres *horror* and *comedy* were chosen for the following reasons:

- Movies and trailers of these genres were available and already manually analyzed to a certain degree thanks to the ADDiCT project
- Horror trailers seem to share certain characteristics

- Comedy movies and trailers depend more on dialogue and semantics, such as humour, than action and horror, and provide a further challenging context.

For each of these genres 15 movies and their corresponding trailers were used during the ADDiCT project The data collected consists of information about the movie, the trailer, the music in the movie and in the trailer as well as the emotions occurring while watching the movie.

As these annotations used a different set of properties and were not always conducted on the theatrical trailer, in order to get an appropriate and comparable basis for further analysis, a reannotation using the SVP tools was necessary.

4.1.2 Manual Annotation

In order to create trailers for the genres horror and comedy, knowledge about the structure and special characteristics of such trailers was necessary. This issue was addressed by performing a manual annotation and analysis of a small test set containing two horror and two comedy trailers.

For the annotation, the trailers were segmented into shots and properties of each shot were annotated in a database, utilizing the same annotation tool as the SVP project [Brachmann et al., 2006] (see chapter 3.2.7.1) and a custom made shot-based video player (see chapter 5.6.3). Although two trailers are insufficient for a general empirical study, the results provide a formal background for the creation of such trailer models.

4.1.2.1 Horror

The trailers chosen for the horror genre were the theatrical trailers for the movies *Dreamcatcher* (2003, by Lawrence Kasdan) and *The Texas Chainsaw Massacre* (2003, by Marcus Nispel).

The basic structure of trailers described by Hediger [2001] and the SVP team [Brachmann et al., 2006] (see chapter 3.2.7.1) applies to these horror trailers as well. The characters and the setting are established in the intro phase. The basic story of the movie is introduced in the following story phase. The climax of the trailer is comprised by spectacular and frightening scenes in the action phase. The horror trailers are concluded by a calm outro phase, bringing down the tension. In-between the following movie title and credit sequence, an optional button scene can be placed.

Figures 4.2 and 4.3 show the semantical structure of the two analyzed horror trailers. Each trailer can be split into the four phases. The first track *Speech* of each figure represents dialogue or general speech sequences which contain information

Shots from 'Bruce Almighty'

#	screenshot	trailer time (mm:ss.fraction)	movie time (hh:mm:ss)	footage	description	keywords	Bruce Nolan	God	Grace Connelly	Jack Baylor	Susan Ortega	transition (to next shot)	shot type	camera motion	lighting	speech	music	audio level
1		00:00.167 dur: 4.21 sec	00:00:00	legal information		-	—	—	—	—	—	hard cut		still	medium		—	silent
2		00:04.375 dur: 0.96 sec	00:00:00	blank screen		-	—	—	—	—	—	hard cut		still	medium		—	silent
3		00:05.333 dur: 3.08 sec	00:00:00	company logo		-	—	—	—	—	—	hard cut		still	medium	Diegetic sound	—	medium
4		00:08.417 dur: 2.08 sec	00:00:00	movie scene	Introduction Main Characters	-		—		—	—	hard cut		still	dark	Diegetic	—	medium
5		00:10.500 dur: 2.46 sec	00:00:00	company logo		-	—	—	—	—	—	hard cut		still	medium	Diegetic	✓	medium
6		00:12.958 dur: 6.38 sec	00:00:00	movie scene	Introduction Main Characters	-		—	—	—	—	hard cut		zooming out	dark	Diegetic	✓	medium

Figure 4.1 The shot-based manual annotation of the movie Bruce Almighty

about the story and are of semantical value. The second track *Filler* symbolizes additional footage clips with other content, such as setting or general action sequences. Those sequences mostly do not contain much background information but fill the gaps between the speech sequences and support the setting of the trailer. Finally, the third track *Animation* represents the arrangement of additional material with mostly textual animations. The horizontal position of the objects corresponds roughly to the temporal position in the trailer. In the following part the numbers shown in the figures are references to identify the semantic sequences.

Intro Phase A trailer may open with a green screen showing its rating. However, the two analyzed horror trailers both started with a black screen, followed by one or more film company logos (sequence A1 in figure 4.2 and sequence B1 in figure 4.3), separated by black frames with fading transitions.

In the following shots, either the main characters or the setting of the film is shown using establishing shots. The trailer for Dreamcatcher opens with shots showing one of the main characters as a young boy and a group of the other main characters giving a toast to him (sequence A2). Both scenes are accompanied by a diegetic voice-over originating in the second scene, connecting them via a sound bridge. In-between those two shots, a textual animation framed by a Native American's dreamcatcher (sequence A3) is displayed. The textual animation is read aloud by a non-diegetic

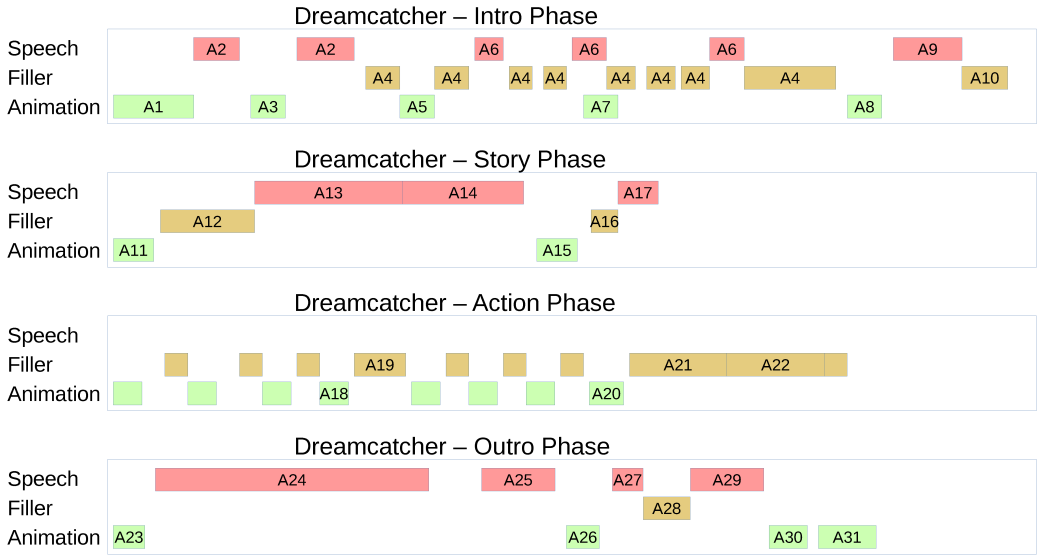

Figure 4.2 Semantical Sequences in the *Dreamcatcher* Trailer

narrator.

The trailer continues by alternating between two scenes. One scene establishes the location, a snowy forest, and shows a car driving down a road from various angles. The scene plays out by showing the car evade a man sitting in the street causing it to crash (sequence A4). In the other scene, two of the main characters are having a telephone conversation (sequence A5). The scene changes are stylized by showing about half a second of black screen. Additional textual animations are placed between the shots as well (sequence A5 & A7).

Another textual animation with a corresponding narrative voice over (sequence A8) is followed by a scene showing two main characters talking about the strange behavior of the animals in the forest (sequence A9). The intro phase of the Dreamcatcher trailer is concluded with two short shots showing another main character hiding in the snow (sequence A10). Throughout the whole phase, the trailer is accompanied by calm music, building up towards the end.

The trailer for *Texas Chainsaw Massacre* uses a different arrangement. After the company logo (sequence B1), a van is shown driving down a road (sequence B2). About one second of black screen separates this setting introduction from the next scene showing the main characters (sequence B3). The transitions are fades to and

from black. The main characters are a group of young people on a road trip. The following scene shows their relationships (sequence B5) and is broken by a textual animation stating the date when the story takes place (sequence B4). After another fade to black and another textual animation stating the location as *Trevis County, Texas* (sequence B6) the group comes across a young woman on the street (sequence B7) saying "They are all dead" (sequence B8). This marks the end of the intro phase of this trailer. A pop song is used as background music throughout the whole phase which is only lowered to emphasize dialogues in the trailer.

Story Phase The purpose of the story phase is to present an overview of the story once the setting and the characters have been introduced. In the Dreamcatcher trailer, the story phase starts at about 48 seconds with a textual animation and a narrative voice-over (sequence A11) revealing the basic plot of an *alien invasion*. The following shots show the arrival of military personnel and the antagonist's voice explains the story (sequence A12). This diegetic narration overlaps the next scene showing a main character (sequence A13). Another diegetic voice-over and the corresponding shot (sequence A14) supplement the narrative explanation. After a text insert (sequence A15) featuring the movie company names, three more shots show the ongoing story of the movie (sequence A16), combined with a narrative voice over originating in the second shot (sequence A17). The music is similar to the one used in the intro phase. Towards the transition to the next phase, the lingering background music turns into a more active and present piece of music.

In the trailer for Texas Chainsaw Massacre, the story phase starts at about 34 seconds. The first part of this phase consists of a repeating pattern of black frames and shots, each about half a second long and separated by a fading transition (sequence B9). Each shot is accompanied by a heartbeat-like sound effect. The first three shots show various objects of yet unknown meaning. Later shots show a house and the group of young people discovering it. After about fourteen iterations, the pattern changes as the protagonists realize that they are not alone (sequence B10). The trailer continues with three longer scenes (sequence B11), composed of multiple shots and showing further discoveries by the main characters. The soundtrack builds up tension towards the following action phase while the protagonists encounter an old man in a wheelchair who is shouting "What the hell are you doing in my house?" (sequence B12). The transition to the next phase is marked by the young woman, who the group met at the end of the intro phase, crying: "We are all gonna die"(sequence B13).

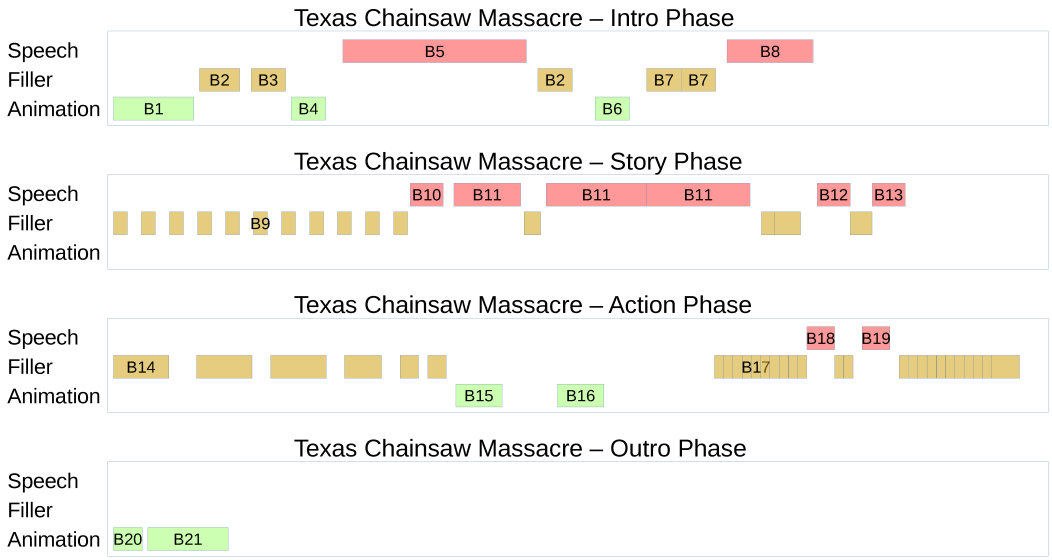

Figure 4.3 Semantical Sequences in the *Texas Chainsaw Massacre* Trailer

Action Phase The action phase of the trailer for Dreamcatcher starts at about 1 min 15 seconds and is arranged in a simple scheme. A textual insert featuring the name of the first actor (sequence A18) is followed by a few fast-paced shots (sequence A19). Some of them are separated by about 0.35 sec of black frames. The transitions between the shots are hard cuts. This sequence is followed by a textual insert showing the director's name (sequence A20) and a scene of helicopters launching rockets and soldiers firing guns (sequence A21). In this phase a dramatic music track is used as background soundtrack. At the turning point of the tension, a protagonist breaks through glass (sequence A22) and the music is reaching its climax.

In the Texas Chainsaw Massacre trailer, a similar pattern is used. It starts at about 1 minute 11 seconds. Two shots showing screaming and running people are followed by about 0.35 seconds of black screen. The transition between the shots and the black frames is a stylized fade to white, simulating a photograph being taken using a flash. A clicking and screeching sound is added to each transition. After five iterations, the flashlight itself is shown and blends into a text insert showing the director's name (sequence B15).

The following part of the action phase consists mainly of a black screen with a

short textual animation stating that the movie is inspired by a true story (sequence B16). During this black sequence, a running and screaming woman can be heard. Later on, her breathing gets louder and louder and is accompanied by footsteps.

Suddenly a chainsaw breaking through a wall can be seen and heard, accompanied by the return of the music. More fast-paced shots of running and screaming people follow and are interrupted by two diegetic voice-over narrations of a police officer (sequence B18) and an old lady (sequence B19). At the end of the action phase, a door is closed and the music also ends on it climax.

Outro Phase In the Dreamcatcher trailer, the outro phase starts at about 1 minute 38 seconds, marked with a text insert of a book reference (sequence A24) and a break in the soundtrack. A longer scene of several shots showing the protagonists trying to keep a door closed (sequence A24) and finally being confronted by someone off screen (sequence A25). These shots are mostly medium-close up shots showing the characters with a duration of 0.6-1.6 seconds. Every two or three shots, black frames with a duration between 0.6 and 3.0 seconds are inserted. After another textual animation showing a tagline (sequence A26), the scene closes with a final question asked by a protagonists towards the offscreen figure (sequence A27) and a falling match igniting a fire (sequence A28). The final shots of the trailer consist of a montage of medium-close up shots showing a protagonist in the snowy forest trying to defend himself with a torch (sequence A29). These shots are alternating with black frames, matching the shots length of about 0.2 to 1.2 second. The trailer concludes with a burning dreamcatcher blending into the movie title (sequence A30) closing with the credits (sequence A31). The music becomes calmer and slowly fades out.

The trailer for Texas Chainsaw Massacre however has a much shorter outro phase which starts at about 2 min 2 seconds with a black frame, followed by the movie title (sequence B20) and the credits (sequence B21).

4.1.2.2 Comedy

For the analysis of the comedy genre the theatrical trailers for the movies *Bruce Almighty* (2003, by Tom Shadyac) and *Hitch* (2005, by Andy Tennant) were selected.

Although the four phases are present in these trailers, they are far less differentiated than in action or horror trailers. Most content is transported via spoken language or depicted actions. Figures 4.4 and 4.5 illustrate the semantical structures of the two analyzed comedy trailers.

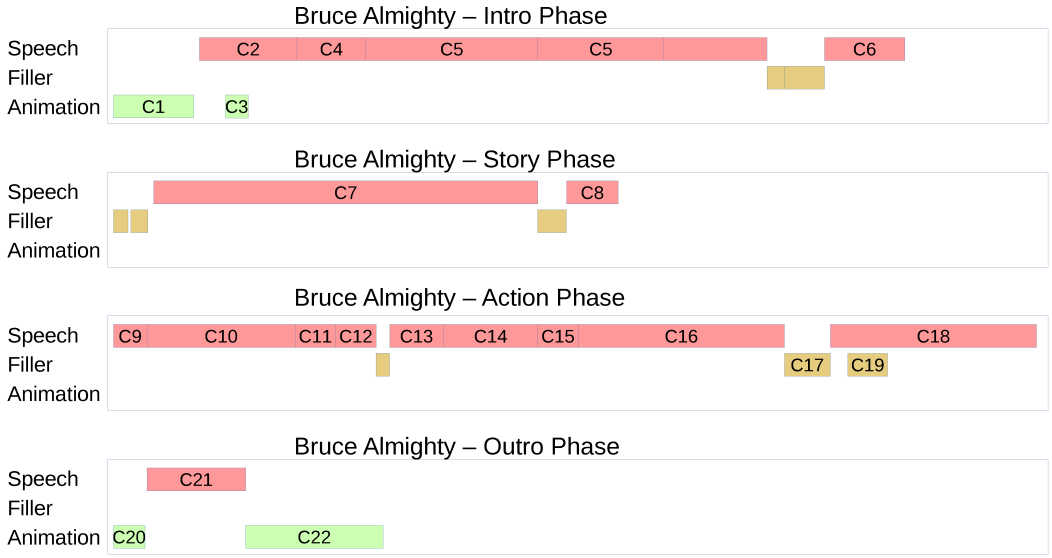

Figure 4.4 Semantical Units in the *Bruce Almighty* Trailer

Intro The trailer for the movie Bruce Almighty starts with a green screen showing the rating of the preview (sequence C1). Separated by one second of black frames, the animated film company logo starts off the actual trailer (sequence C1). Synchronously an alarm clock sound can be heard. It originates in the next shot which shows the protagonist Bruce waking up in the morning (sequence C2). A second film company logo follows two seconds later while the alarm continues (sequence C3). At this point, the first background music pop song starts and accompanies the whole phase. The rest of the intro consists of five more scenes, each containing several shots. These show the daily routine and problems of the protagonist as a television reporter. The first of these scenes shows his dog urinating in his flat (sequence C4). The next scenes show challenges in his job and the relations to his co-workers (sequence C5). A non-diegetic narrator is commenting the scenes, asking if the shown incidents are familiar to the audience. In the final scene of the intro phase, Bruce shouts towards god complaining that he is not doing “his job” (sequence C6).

The beginning of the Hitch trailer features a green screen as well. It is followed by the company logo and the start of the first music track (sequence D1). The first scene of the trailer shows a couple having an argument and how the protagonist Hitch helps to resolve it (sequence D3). In this scene, three textual animations are

inserted (sequence D2) in a grid-cut pattern, using the movie scene as base. After the last textual insert, another shot is added showing the protagonist introducing himself (sequence D4). Three fast-paced shots showing the skyline of New York are used to establish the location where the movie takes place and transitioning to the next scene (sequence D5). The next shots show Sara, the female main character and future love interest of Hitch, talking about him and giving further background information (sequence D6). In the middle of this three-shot scene, Hitch is shown while the diegetic voice from the previous scene continues, forming a sound bridge (sequence D7). Another short scene of three shots shows the protagonist "at work" having dialogues with other men (sequence D8). The diegetic voice of one of these shots is used as narrative voice-over during the whole scene. A narrative voice-over is providing further information about Hitch. Towards the end of the intro phase, a slapstick scene (sequence D9) introduces Albert, another main character and the "apprentice" of Hitch. In the first two shots of the scene the apprentice is stumbling and in the third and final shot Hitch and another man seem to comment the scene (sequence D10).

Story Phase The transition from the intro to the story phase in the trailer for Bruce Almighty takes place at about 48 seconds and is marked by a fast camera turn around Bruce with an accompanying sound effect. The background music changes as well. The story phase in this trailer is basically one large scene, showing a conversation between Bruce and God (sequence C7). During the conversation, God offers Bruce all of his powers to prove that he can do better than him. The scene is a typical dialogue and consists of 13 shots, alternating between Bruce and God. The remainder of the story phase shows Bruce trying out his new powers for the first time in a coffee shop while the voice of God continues (sequence C8).

In the trailer for Hitch, the beginning of the story phase starts after 38 seconds and is also marked by a change in the background music. In the first shots, Hitch introduces himself to Albert, supported by a voice-over narration (sequence D11). The next shots show how Hitch helps Albert to get the phone Number of his love interest in a comedic way (sequence D12). The transition to the following scene is realized through two fast-paced shots of night life and commented by the narrator (sequence D13). Thematically matching, the next scene shows Hitch giving dance lessons to Albert (sequence D14). This scene contains 13 shots, starting with a dialogue and alternating camera positions. During the dance part, the camera position changes into a bottom-up and later on a top-down view. The story phase of this trailer ends with a comedic comment by Hitch.

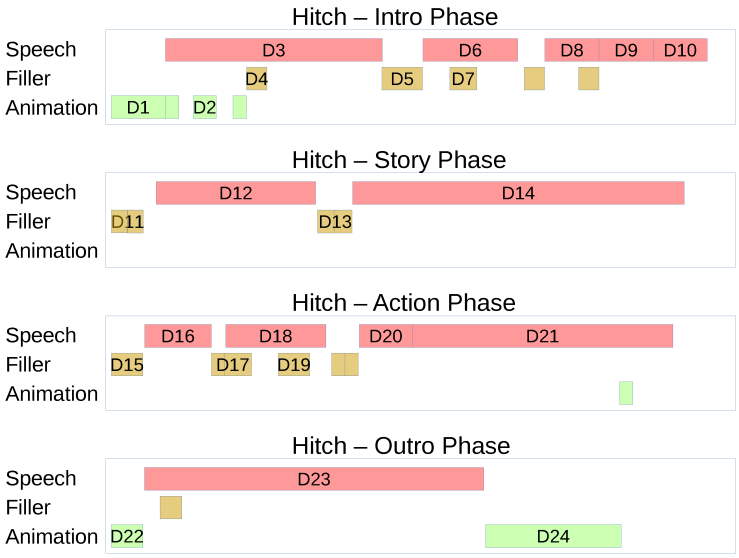

Figure 4.5 Semantical Units in the *Hitch* Trailer

Action Phase The action phase of the Bruce Almighty trailer starts at around 1 minute 16 seconds. It is accompanied with a matching music track, *The Power* (1990) by Snap!, and shows several shots of Bruce demonstrating his new powers. In the first shot he causes fountains of water to burst out of fire hydrants (sequence C9). The next scene shows a conversation with his girlfriend about the size of her breasts and Bruce uses his powers to enlarge them (sequence C10). In the following shot, God surprises Bruce and asks him if he "is having fun" (sequence C11). In the next scene God and Bruce are having a conversation while standing on water (sequence C12). This scene is commented by the non-diegetic narrator who mentions previous movies by the director. The voice of God can be heard over the next shots, as he reminds Bruce of his new responsibilities coming along with his new job. This sound bridge acts as transition towards the next humorous scene showing Bruce replying to God's e-mails (sequence C13). The action phase continues with several scenes of Bruce demonstrating his powers. He is pulling the moon closer for a romantic evening with his girlfriend (sequence C14) and makes his dog use the toilet instead of using the carpet as in the opening scene of this trailer (sequence C15). Another scene shows Bruce at work (sequence C16) where he makes a fool out of a colleague, who mocked him in the beginning of the trailer. The scenes are also commented by the narrator. In the following two scenes, Bruce parts the coffee in his cup and later

on the cars on the street to make room for his sports car (sequence C17). The final scene of the action phase is a dialogue between Bruce and God above the clouds in the sky (sequence C18). God is asking how many people Bruce has helped which is followed by a shot illustrating how Bruce misused his power for personal pleasure (sequence C19). The scene closes with a cliffhanger as God is mentioning that Bruce might be dead.

In the Hitch trailer the action phase starts at about 1 minute 17 seconds. It begins with a transitional shot combined with the voice-over narrator commenting on what Hitch is doing (sequence D15). The transitional shot is followed by a humorous scene in which Hitch takes Albert to a waxing studio (sequence D16). The following shots show the beginning of a romantic relationship between Hitch and Sara by using alternating medium close up shots (sequence D17). A shot showing Hitch talking to a friend about Sara is inserted in between (sequence D18). The whole scene is accompanied with the voice of Hitch taken from this shot. Two shots showing slapstick from a jet-ski tour of Hitch and Sara are also inserted into this scene (sequence D19). After two more shots that show the further development of the relationship, a slapstick scene shows Hitch getting his shirt caught in a cab door (sequence D20). The last scene of the action phase is about 17.5 seconds long and shows a dialogue between Hitch, Sara and her parents (sequence D21): During a meal preparation, Hitch has an allergic reaction to the food. A textual animation of the main actor's name is used as transition to a different setting, a pharmacy. The final shot of the scene shows Hitch's swollen face.

Outro In the trailer for Bruce Almighty, the outro phase starts at around 2 minutes 18 seconds with the textual animation of the movie title (sequence C20). The trailer ends with a button scene showing the dog reading the newspaper on the toilet (sequence C21) followed by the credits (sequence C22).

The outro phase of the Hitch trailer has a similar pattern. After about 1 minute and 55 seconds, the movie title is shown (sequence D22). The final humorous scene shows how Hitch tries to teach Albert how to kiss the right way and how Albert gets it wrong (sequence D23). The trailer ends with the credits (sequence D24).

4.1.3 Structure of Horror and Comedy Trailers

Although the two analyzed horror trailers share a common basic structure, some parts use very specific stylistic cut patterns. Such patterns, like the overlapping grid cut in the intro phase of the Dreamcatcher trailer, are very difficult to realize in an abstracting model.

Content	Comment
Intro Phase	**Duration: 33-48 Sec**
Company Logo(s)	
Character Introduction	or vice versa
Setting Introduction	
Character Relation Introduction	
Problem Introduction	
Story Phase	**Duration: 27-39 Sec**
Characters Telling Background	one of these, presenting story development
Montage Short Clips / Black frames	
Discoveries	characters make discoveries or meet antagonists
Action Phase	**Duration: 23-51 Sec**
Actor Names	repeated for main actors
Fast Shots	
Black Frames	
Director Name	
Montage of Fast Clips	combined with dramatic music
Spectacular Scene	
Outro Phase	**Duration: 13-39 Sec**
Slow and Frightening Scenes	optionally
Movie Title	
Credits	

Table 4.1 Generic Structure of a Horror Trailer

Table 4.1 shows the generic arrangement of scenes in the analyzed horror trailers. Scenes are often separated by black frames which fade in and out from the footage. The music in the intro and story phase is often mysterious and deceivingly calm.

Sometimes special effects are used in the arrangement as well as in the audio track. For instance, the intro phase of the Dreamcatcher trailer makes use of different overlaying scenes. Also, in the Texas Chainsaw Massacre trailer the optical effect of taking photographs is combined with the screeching sound of the flash (see chapter 4.1.2).

During the the end of the action phase, the music raises the tension towards the climax. The soundtrack of the outro phase is composed similarly to the intro phase with calm and damped sounds to complete the dramatic arc.

Content	Comment
Intro Phase	**Duration: 37-47 Sec**
Company Logo(s)	
Humorous Scene, Character Introduction	
Fast Scene Setting Introduction	optionally, used as transition between talking scenes
Humorous Scene, Character Background	multiple
Problem Introduction	
Story Phase	**Duration: 27-36 Sec**
Humorous Scenes, Story Background	2 times
Transitional Setting Scene	
Action Phase	**Duration: 36-61 Sec**
Humorous Scene	5-11 times
Fast Transitional Scene / Shot	
Main Actor Name	
Humorous Scene, Cliffhanger	
Outro Phase	**Duration: 8-23 Sec**
Movie Title	
Final Humorous Scene	
Credits	

Table 4.2 Generic Structure of a Comedy Trailer

The generic arrangement of a comedy trailer is illustrated in table 4.2. The individual phases of a comedy trailer are much less distinctive than those of action or horror trailers. The intro phase focuses on the introduction of the main characters, settings and relationships. The transition from the intro to the story phase is accompanied by a change in the protagonist's world (comparable to the first plot point in chapter 2.1.1). This lays the foundation for the development of the story phase. The action phase of a comedy trailer loosely continues telling the story but focuses on showcasing humorous scenes and slapstick. Near the end of the action phase the name of the main actor is either shown in a textual animation and/or announced by a voice-over narrator. An animation containing the movie title starts the outro phase. This is followed by a final humorous scene, the button (see chapter 2.2.3) and closed by the credits.

Compared to action and horror film trailers the comedy trailer relies more heavily on speech. Such speech can either be diegetic, i.e. from the footage, or from a non-diegetic narrator.

The soundtrack of a comedy trailer often features one or more pop songs. In some scenes however, calmer background music is used. During scenes containing speech or voice-over, the music is either lowered or stopped entirely.

4.2 Abstracting Application

In order to generate a trailer in a more professional way, some desirable functionality must be provided, e.g., to monitor the generation process in detail and to see intermediate results. Based on this desired functionality an entirely new video abstraction system was developed because of the prototypical character of the SVP system.

4.2.1 Shortcomings of the SVP System

In the course of extending the SVP action trailer model towards horror and comedy trailers, the following properties turned out to be improvable:

- The software is based on multiple external programs, programming languages and operating systems.
- A general lack of interaction and possibilities to adapt or correct manually
- No preview capabilities, only a batch run of the software is possible.
- Categories cannot be edited interactively
- The categorization process lacks visual output and ways to adjust filters

- Editing of trailer models is difficult and requires additional software.

The software developed by the SVP team is a proof-of-concept trailer generation system. Therefore, the focus was to create a working system rather than an efficient software architecture. The modules of the system are more like individual programs than an integrated application. Besides the analyzer modules (see chapter 3.2.7.2), which were implemented in various programming languages, the generator part depends on external tools.

The system runs iteratively using command-line parameters supplied on execution with no further interaction. While such behavior is suitable for the automatic trailer generation, the possibility to pause the process on several steps could be useful. The possibility to alter certain parameters for the definition of footage categories and trailer models in runtime would speed up and simplify the manual creation and optimization. Additionally, such improvements would open up other possible use cases for the system, such as content retrieval.

In the SVP system, the definition of categories and trailer models is done using an external modeling software. This modeling software is not specialized for the purposes of the SVP system and does not offer assistance for the tasks of editing.

4.2.2 Requirements for a new Application

Taking the previously identified shortcomings into account, the following requirements for the software were defined:

- Monolithic structure: the software should have fewer external dependencies and be easier to maintain, deploy and to run.
- More interaction and possibilities to intervene in the movie segmentation and trailer generation process: in order to examine and improve the whole abstracting process, more intermediate information and tools to influence the process are desirable.
- Easier editing of category definitions, trailer models and sound archives: the application should support the editing by showing the results of the categorization process, by visualizing the trailer model in a more comprehensible way and help the user by making suitable suggestions during the editing process.

For automatic video abstracting, three different steps should be considered:

- Categorization: provide an overview of the segmented footage in order to finetune the categorization parameters

- Model editing: provide an easy-to-use interface for the definition of the model describing the rules how an abstract is to be generated
- Abstract preview: provide a preview of the abstract, show the chosen clips and allow for alternatives to be selected

In the first step, the footage is segmented according to a set of defined categories. Each category has a set of category parameters which are used to filter the footage according to the previously extracted features. In order to create trailers for different genres, these categories may need to be adjusted. It is much easier to rate the categorization precision and to adjust the categories and their parameters if the categorized footage can be examined instantly. This speeds up the whole process as the need to generate a trailer every time a value is changed becomes obsolete.

The editing of a trailer model faces similar difficulties. Without a complete run of the software it is difficult to test and debug the changes. In the final video abstract generation step, a list of clips and transitions is derived from the trailer model and supplemented with footage and textual animations. Additional music and sound effects are added.

Besides the automatic abstract generation, another mode of compiling a video abstract might be of interest for testing as well as for improving the resulting videos. In this mode, a user is presented with a suggested arrangement and may select alternative clips, transitions, or effects.

Chapter 5

Implementation

In chapter 4.2.2 the requirements for a video abstracting software were defined. In this chapter the implementation of such a software application is described. Furthermore, the creation of abstracting models for the generation of horror and comedy trailers is detailed.

5.1 Preparational Work

The focus of the application is on the generator part of the video abstracting process. The extraction of features from video files is performed in a preceding step by the detector modules of the SVP system (see chapter 3.2.7.2). In order to simplify this task, a script starts the detector modules in the right order and compiles the results in a single XML file.

5.2 Application Design

The application requires a fast and powerful programming language, since a lot of data processing and multimedia operations need to be performed. Additionally, matured and proven libraries for graphical user interfaces, data storage, threading and multimedia playback and editing are required. By focusing on a few libraries and languages, platform-independence is easier to achieve. The *Qt* framework[1] is used in combination with the *C++* programming language used for the user interface, XML handling and the algorithm logic. Additionally, the *GStreamer* multimedia framework[2] and especially its *GStreamer Editing Services (GES)* module[3] provide methods for multimedia playback and editing.

[1] https://www.qt.io/
[2] https://gstreamer.freedesktop.org/
[3] https://gstreamer.freedesktop.org/data/doc/gstreamer/head/gstreamer-editing-services/html/ges-architecture.html

5.2.1 Video Abstracting Workflow

The purpose of the application is the automatic generation of video abstracts, exemplified in the domain of movie trailers. Besides the automatic generation, the application is supposed to allow for the definition and refining of the parameters and models used in this process.

From the perspective of a user, the video abstracting process is as follows:

1. Selection of movie (video file and extracted features), category and trailer model file.
2. Adjustment of categories, if necessary or desired and execution of categorization.
3. Selection and adjustment of a video abstract generation model, if necessary or desired.
4. Generation of the trailer structure, selection of footage clips, animations and effects by the system, adjustment by user if desired.
5. Compilation of the multimedia components into a video, adjustments of parts by user if necessary, save the result as video file.

This workflow is reflected in the design of the user interface.

5.2.2 User Interface

The user interface should include methods for managing the data files, such as *open* and *save* dialogue actions, and provide specific tools for the steps of the abstracting workflow. Among the data files serving as input are a video file and an XML file containing the corresponding features. The system is capable of dealing with multiple movies at once. Additionally, an XML file holding the description of the current categories and their corresponding category parameters is used. Another XML file stores the abstracting models. Finally, several *sound archives* are used in the generation process. Each contains an XML file with metadata and several audio files containing music and sound effects.

The main user interface is divided into three individual widgets, one for each of the main tasks: *categorization*, *abstract model editing* and *video abstract viewing*.

5.2.2.1 Categorizer

The purpose of the categorizer widget is the display, creation and modification of footage categories and the preview of the segmented clips in their respective category.

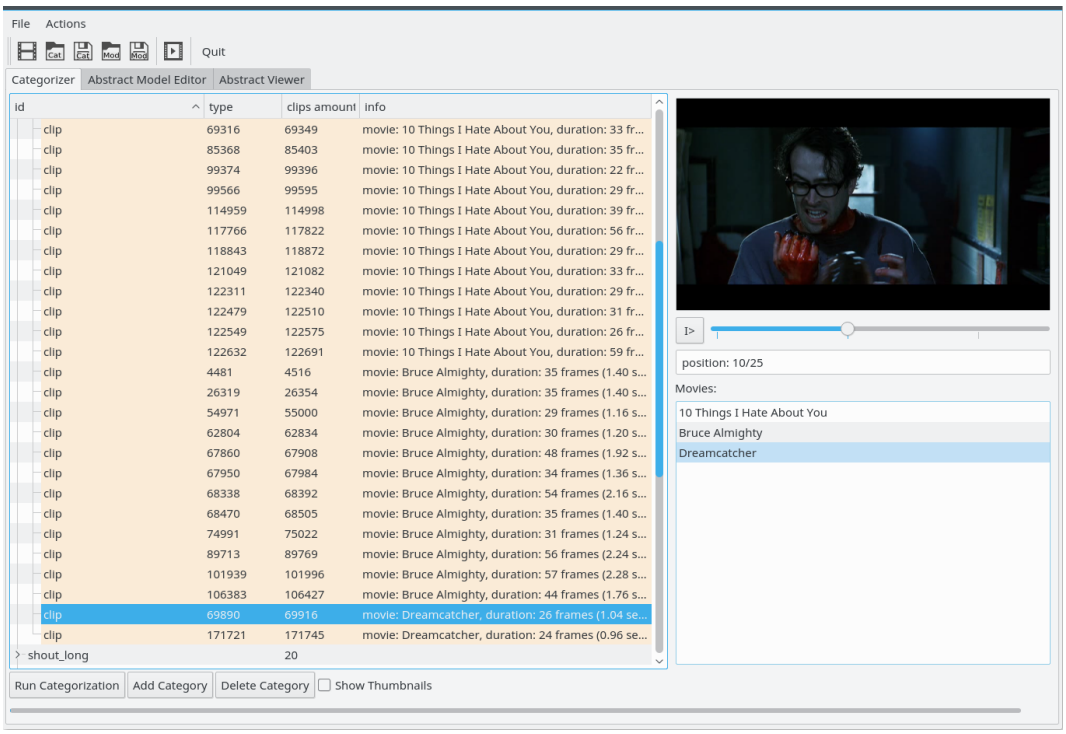

Figure 5.1 Categorizer Widget

The left part of this widget contains a list of categories. The right part holds a video player window for the playback of categorized video clips and a list of the movies currently available into the system (see figure 5.1). Once the categorization is completed, all corresponding video clips are shown as collapsible subitems in the list. A double click on a category opens the *category parameter dialogue* (see figure 5.2). It contains the *category id* and a list of *detector parameters*. An id can be assigned to each parameter. A single category parameter contains a value filtering for one specific feature. To do this, a *detector id* has to be chosen from the dropdown menu or entered manually. Then a feature has to be selected for which a minimum and maximum value has to be chosen to classify it as said category.

If a category contains more than one parameter, only footage fulfilling all set category parameters is assigned to that category. In this a case the *range adjust policy* which can be set to either *shrink* or *ignore*, is used to determine the handling of the current frame range. If *shrink* is selected, the current frame range is adjusted

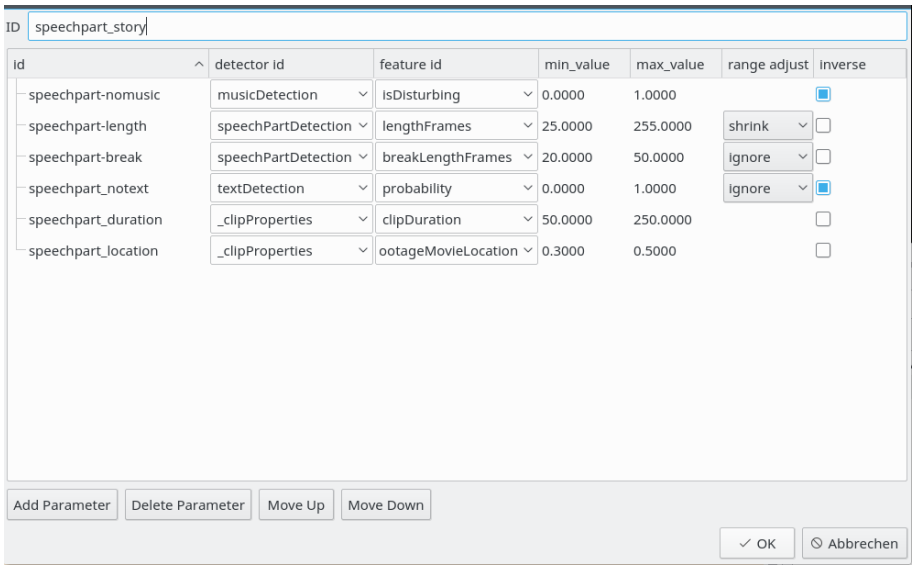

Figure 5.2 Category Parameter Dialogue

to not exceed the part where the feature is present.

Finally, a category parameter can be set to *inverse*. This means that the frame range of this feature is excluded from the category if the value of the feature is between the minimum and maximum defined.

A special detector named *_clipProperties* can be chosen in order to limit the maximum duration of a video clip via the feature id *clipDuration*. Furthermore, the feature *footageMovieLocation* allows for filtering the video clips according to their position in the movie timeline.

On the bottom of the category parameter dialogue buttons for adding and removing parameters, as well as for moving a selected parameter up and down in the list can be found.

Controls for the categorization process are placed below the categorizer widget (figure 5.1). The first button allows for starting the categorization process using the listed categories and the listed movies. Furthermore, additional buttons can be used to add a new or delete an existing category.

After the categorization process, a single click on a category will open a list of corresponding video clips, provided a syntactical movie analysis has been taken place beforehand. By double clicking a video clip entry, it will be played in the video player window on the right side of the categorizer widget. In order to allow for a quick

preview of the categorization results, a check box can be ticked to enable thumbnail previews of the categorized movie footage.

The categorizer widget features a video window with a play/pause button below. Furthermore, a progress slider and a status bar provide information about the current playback status. While the video playback is paused, the slider may be used to scroll through the current video clip.

Below the video player, a list of currently loaded movies is displayed. Movies can be added or removed from this list through the menu entry *Edit Movie Configuration* or the corresponding button below the menu bar.

A progress bar on the bottom of the categorizer widget informs the user about running background processes during categorization or thumbnail generation.

5.2.2.2 Video Abstracting Model Editor

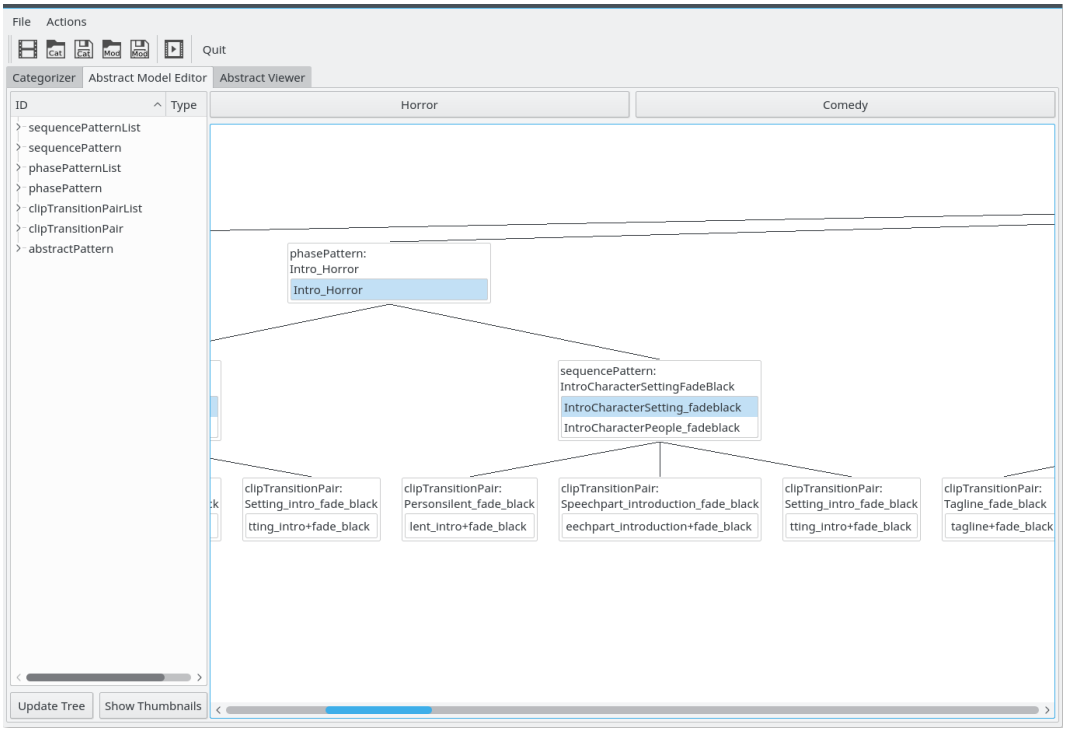

Figure 5.3 Model Viewer Widget

In the *model viewer widget* (see figure 5.3), the video abstracting models are

displayed and the definition of the rules describing the structure of a video abstract can be edited. On the left side of the widget a list of all elements in the video abstracting model is presented and the right side shows the hierarchical view.

On top of the hierarchical view, all currently loaded abstracting models are represented with a button. By clicking one of them, the corresponding model is drawn by iterating through the tree structure.

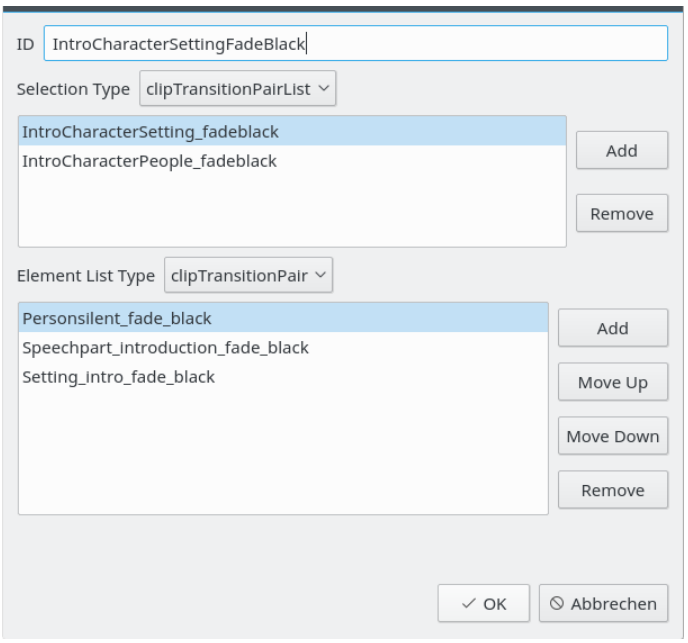

Figure 5.4 Node Edit Dialogue

Each node of the abstracting model may be edited by opening a corresponding *node editing dialogue* (see figure 5.4). This is done via double click in the model or by using the context-menu in the list on the left side. Besides the id of the node, two lists of child nodes can be edited here. The first list is called *Selection* and contains the children of the node, from which one child is selected in the generation process. This allows the system to generate non-deterministic video abstracts.

The second list called *Element List* contains the child nodes of the currently selected item of the first list. In comparison to the first list, all of the children are chosen in the corresponding order during abstract generation. For this reason, the items in this list can be moved up and down using the buttons on the right side. The video abstracting model and algorithm is explained in detail in chapter 5.3.2.

The node editing dialogue for nodes of types *phasePattern* has an additional drop-down menu which allows the selection of a music phase id. This id corresponds to the ones in the sound archives.

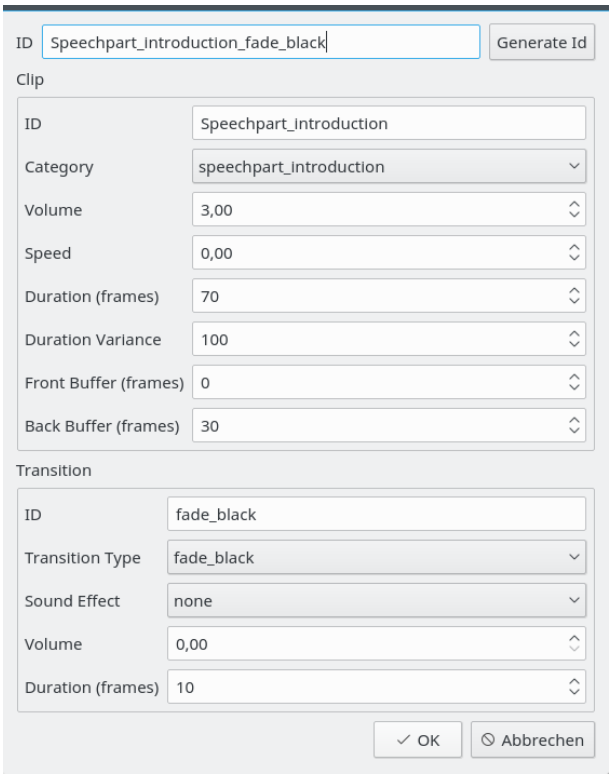

Figure 5.5 Clip/Transition Dialogue

A different dialogue is used for editing the bottom level nodes of the abstracting model, the *clip/transition pairs* (see figure 5.5). It consists of a generic part containing the id of the clip/transition pair which can be generated out of the clip and transition id using the button next to it. Below are two areas, one for options of the clip and one for the options of the transition.

In the clip area, an id and a category, from which the clip should be taken, can be selected. Furthermore, the volume of the audio track can be changed and the playback speed adjusted to achieve slow-motion or fast-motion effects. The desired duration can also be selected, as well as the allowed variance from the length.

The transition section allows for assigning an id and to select a transition type

which can be either *hardcut*, *fade black* or *flash white*. An accompanying sound effect can be chosen together with its volume set. Finally, the duration of the transition can be defined in frames.

5.2.2.3 Video Abstract Viewer

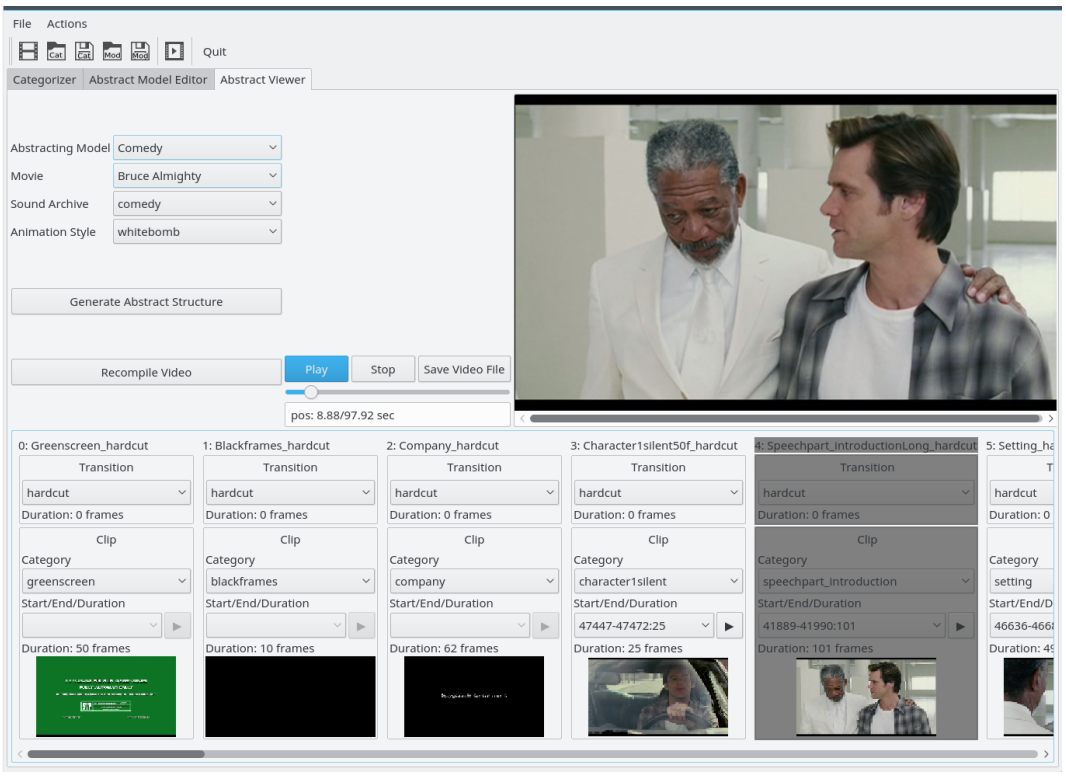

Figure 5.6 Abstract Viewer Widget

The third main widget of the application allows for the actual generation of the abstract. This is done in two steps. First, a list of clips and transitions is derived from the abstracting model, together with music tracks and sound effects. The text animations are rendered as well. Second, the media files are arranged and the actual playback of the video abstract started.

The abstract viewer widget is shown in figure 5.6. In the top half of the widget, several parameters for the generation process can be set on the left side. The final

video can be seen in the player window on the right side. In between there are controls for playback and saving the current video abstract to a video file. The bottom half shows the clip/transition pairs selected for the video abstract.

Among the parameters for the generation process are the abstracting model and the movie to be used. Additionally, a sound archive and an animation style can be selected. The two buttons *Generate Abstract Structure* and *Compile/Recompile Video* start the two steps described above: first the preparation of the structure and material, and second the actual media arrangement and playback.

The bottom area of this widget shows the structure of the currently generated video abstract. Initially empty, it is filled with a list of widgets representing the generated structure of the video abstract. Basic information is given about each clip/transition pair in these widgets. Additionally, it is possible to select alternative transitions, categories and/or alternative clips. A thumbnail preview of the selected clip helps to get a fast impression of the abstract to be generated. An individual preview of each clip can be played back by clicking the corresponding button above the thumbnail.

A click on the *Compile Trailer* button will arrange the media files and start the playback of the currently displayed abstract structure.

5.3 Video Abstracting Algorithms

5.3.1 Categorization

The categorization process segments the movie into video clips and groups these clips in semantic categories. The input of this algorithm consist of the extracted features (see chapter 3.2.7.2) and a list of categories. The movie features are organized according to the detectors used in the analyzing process. A detector may detect several different features. For example, the detector *movementDetection* delivers the features *direction*, *magnitudeAbsolute* and *magnitudeNormalized* in a given strength at a certain frame range. Each category parameter contains the detector id and feature id.

Frame Ranges versus Shot-based approach In the segmentation process of movies, the question of the smallest unit to be used needs to be answered (see chapter 3.1.1). For the system developed, either a shot-based or a feature-based approach may be used. While the first one would treat shots as the smallest unit, the second one would allow to use excerpts of shots as well.

Since the shots in a trailer are usually shorter than in a movie (as noted by Hediger, see chapter 2.2.4), a feature-based approach was chosen. Such an approach takes only the frame ranges containing the features requested by the category parameters as footage units, instead of the whole surrounding shot.

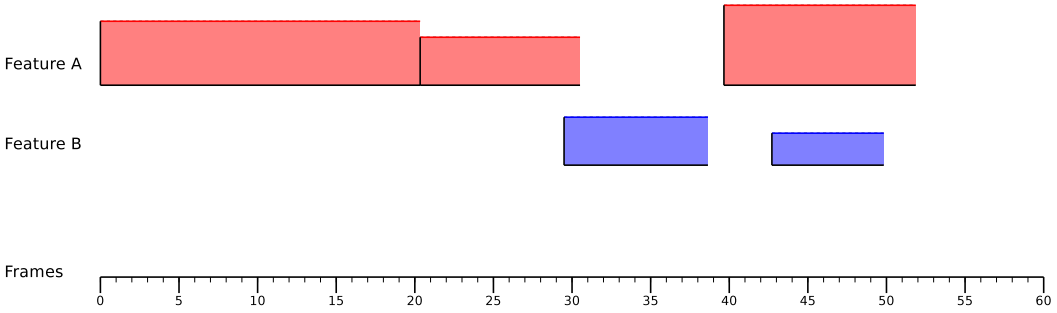

Figure 5.7 Feature Tracks used in the Categorizing Algorithm

Feature Tracks Overlap The categorizing algorithm loops through the category list and locates the required features by checking the detector ids and feature ids in each category parameter. For each of these features, the algorithm creates a track along the movie timeline. This is shown in figure 5.7, in which a timeline illustrates the value of two features along the temporal dimension of the movie.

The algorithm works as follows: In an initial step the first feature track is created. If only one category parameter is defined for the current category, all frame ranges of the feature track are added to this category. This only holds true if the value of the feature is between the minimum and maximum threshold of the category parameter in the current frame range. If the category parameter has the *inverse* flag set to *true*, all but the frame ranges contained in the first feature track are added to the category. Additionally, those frame ranges where the value is smaller than the minimum threshold of the category parameter or bigger than the maximum threshold the category parameter are assigned into the category as well.

When the category contains more than one category parameter, additional feature tracks are created and the remaining category parameters are checked against the respective feature tracks.

Depending on the state of the *range adjust policy* of the category parameter, two different mechanisms are used. If it is set to *ignore*, the start and end frame of the frame range are not adjusted and the frame range is added into the category if all feature tracks overlap at least partly. In the example shown in figure 5.7, the frame

ranges 20-31 and 40-52 would be added into the category if the range adjust policy of the category parameter corresponding to Feature B would be set to ignore.

If, however, the range adjust policy is set to *shrink*, the start and end of the current feature define the start and end of the frame range. This ensures that the feature is present in the whole frame range. If in the example in figure 5.7 Feature B's category parameter is set to *shrink*, only the frame ranges 30-32 and 43-50 would be added into the category.

An enabled *inverse* flag in the category parameter requires one of the following two conditions to be met in order to add the current frame range to the category. First, no occurrence of the current feature overlaps the current frame range, and second, if the value of the feature at the overlapping segment is below the allowed minimum or above the allowed maximum threshold.

This process is repeated for all combinations of movies and categories currently available in the application. At the end of the process, each category contains the excerpts of the movie satisfying all the constraints defined by the parameters of the category.

5.3.2 Abstract Building

The generation of a video abstract is performed in two stages: First, a list of template clip and transition pairs is derived from the abstracting model forming the structure of the abstract. Second, the corresponding footage is selected or additional material generated.

Create Template Clip List from Abstracting Model The data model and the algorithm used for the generation of a video abstract is based on the one used in the SVP system (see chapter 3.2.7.3). A hierarchical structure allows for model video abstracts at different abstraction levels. For trailer generation, a structure consisting of the following main layers is used:

1. Abstract pattern: the topmost hierarchical level representing the complete trailer
2. Phase pattern: a trailer pattern is composed of several phase patterns
3. Sequence pattern: a phase pattern itself consists of sequence patterns
4. Clip/transition pair: each sequence pattern is composed from clip/transition pairs

At the bottom level of this hierarchy, clips and transitions are the smallest unit composing the structure of the video. This abstracting model is encoded in a tree structure.

In addition to the layers described above, the tree structure employs additional layers to achieve non-determinism. On each of these layers, alternative sub-trees can be defined. In the process of deriving a template clip and transition list, one of these alternative sub-trees is selected.

To summarize: an *abstracting model* is a directed rooted and ordered tree. Two kinds of edges connect the nodes, *list edges* and *selection edges*. By traversing the tree structure of a selected video abstracting model, a non-deterministic chain of clips and transition pairs is formed. This is done by starting at the root node and taking a look at the child nodes. If the edge connecting root and child nodes is of type *selection*, one of the child nodes is chosen as start for the next iteration. If the edge is of type *list*, the next iteration is started for each of the child nodes.

- If child nodes are connected to their parent node via a list edge (marked as solid line in figure 5.8), the parent node is replaced by *all* the child nodes in the clip list creation process.
- If child nodes and the parent node are connected via a selection edge, only *one* of the children is selected to replace the parent node in the process.

On the lowest level, no children but clip/transition pairs represent the units of which the final video is constructed.

In the graphical user interface of the application, this tree is shown in a simplified form (see chapter 5.2.2 and figure 5.8). In this representation, only the list edges are shown in order to increase usability. The different alternative child nodes can be selected inside the node widgets and the tree structure is displayed accordingly.

Selection of corresponding Footage Once a chain of template clips and transitions is created, corresponding footage matching the requested template has to be selected. The template clips contain a category along with a desired duration and an allowed variance of this duration.

The process of footage selection first checks whether the corresponding category has footage clips left. If this is the case, these remaining footage clips are checked against a blacklist. This blacklist keeps track of the start and end frame of already used footage clips to avoid redundancies in the selection of footage. The duration of the footage clips not filtered out by the blacklist check is then compared to the duration requested by the current template clip in accordance with the allowed

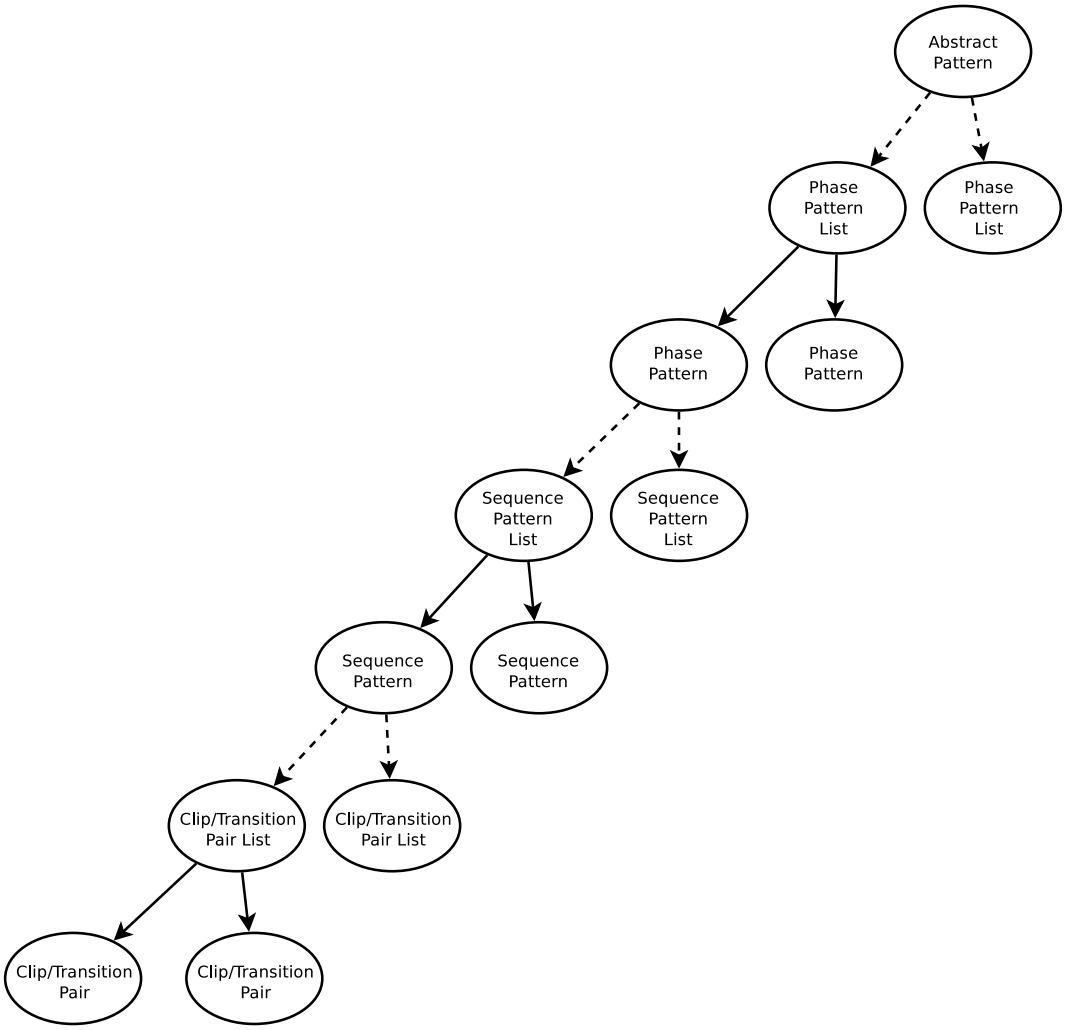

Figure 5.8 The Abstract Model: a solid edge means that all child nodes are to be chosen, while a dotted edge means that only one child node is selected.

variance. All footage clips passing this duration check are now possible candidates for the current template clip. The system then randomly chooses one of these footage clips and blacklists it.

5.4 Application Structure

This section describes the architecture of the application by means of common data structures and the organization in different modules.

5.4.1 Data Models

Two main data structures are used by the application for the categorization process and for the abstract generation. On the one hand these are the categories and the corresponding category parameters, and on the other hand the hierarchical abstracting model. Additionally, other data models relevant are described.

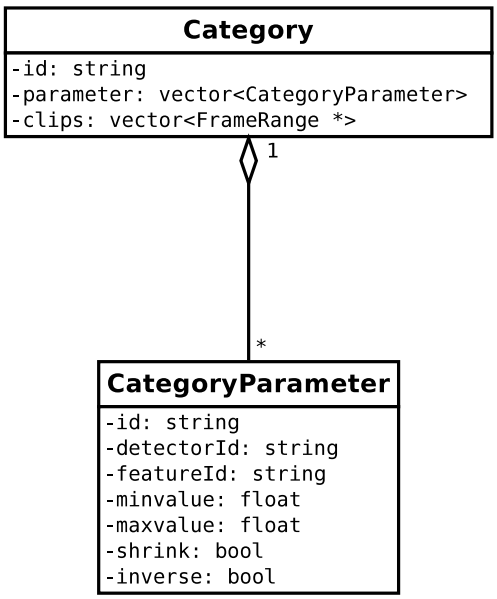

Figure 5.9 Category and Category Parameter Data Type

5.4.1.1 Categories

The categories are implemented as a vector of *Category* objects (see figure 5.9). This object consists of an id for the category, a vector of category parameters and a vector of pointers to frame ranges for storing the categorized movie footage. The category parameters are used to filter the footage using the extracted features handed over to the application. A *CategoryParameter* contains of an id, a detector id and a feature

id. The detector id and feature id are used to identify a specific feature. For example, the detector *movementDetection* provides a feature *magnitudeNormalized*. The CategoryParameter class also contains a minimum and a maximum threshold. Only those frame ranges of the movie containing the feature with a value between those thresholds are taken into account for the category (see chapter 5.3.1). Additionally, the category parameter data type contains two boolean variables *shrink* and *inverse*. These control the range adjust policy and inversion described in chapter 5.2.2.1.

5.4.1.2 Abstraction Model

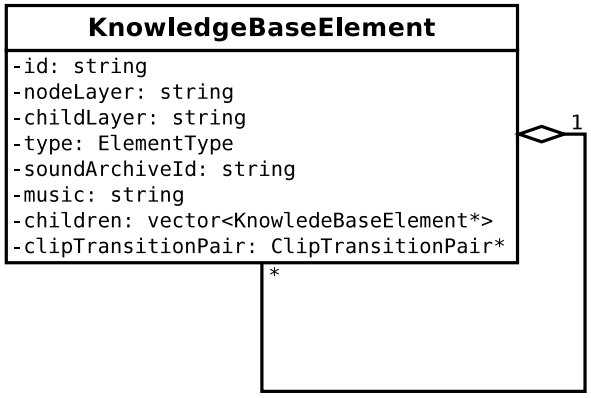

Figure 5.10 Abstracting Model Node Data Type

The abstraction model consists of a hierarchical tree structure described in chapter 5.3.2. The nodes of this structure consist of the data type displayed in figure 5.10. Each node is identified by an *id*. Additionally, the semantical layer of the node, e.g. *phasePattern*, and the layer type of the children is stored in the variables *nodeLayer* and *childLayer*. As the edges between the node and its children have two different meanings (see chapter 5.3.2), the type of these edges is stored in the enumeration variable called *type*.

A sound archive to be used in the abstract can be specified by the variable *soundArchiveId*, as well as a music phase in the variable *music*. The sound archive is used at the root node of the model, while the music phase is set at the phase pattern layer in the hierarchy of the abstract model.

At the bottom layer of the hierarchy, clip/transition pairs are used to represent the structure of the video abstract. Consequently, nodes of this layer have a pointer to a clip/transition pair object stored in the variable *clipTransitionPair*.

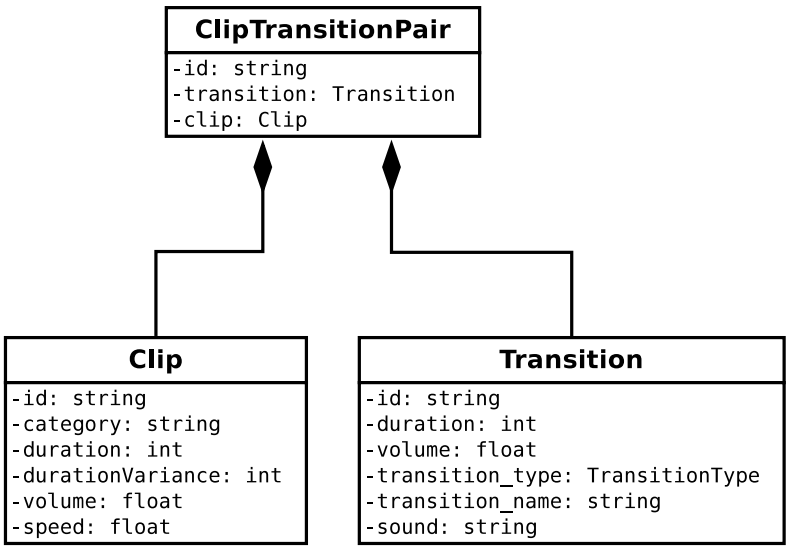

Figure 5.11 Clip/Transition Pair Data Type

A clip/transition pair (see figure 5.11) contains an *id*, a *transition* and a *clip*.

The clip data structure holds the values that an actual footage clip is required to have in order to be selected. The structure contains a variable for an *id* as well. Additionally, a *category* variable defines the category from which the footage clip should be taken. A target *duration* can be defined in frames, as well as the allowed deviation from this duration via the variable *durationVariance*. Finally, the variables *volume* and *speed* determine the volume and playback speed applied to the selected footage excerpt. This allows the system to keep speech audible and to achieve slow-motion effects.

The transition data structure contains an *id* and a *duration* in frames. The *transition_type* is set via an enumeration and can be *hardcut*, *fade_black* or *flash_white*. The transition may be accompanied by a sound effect. The id of the sound effect might be any of the ones set in the sound archive and its volume is controlled by the variable *volume*.

5.4.1.3 Movie Data

Each movie is represented by a video file and an XML file containing information about the movie and the extracted features. This XML file is parsed and the information stored in the data type *MovieData* (figure 5.12). Among the movie information

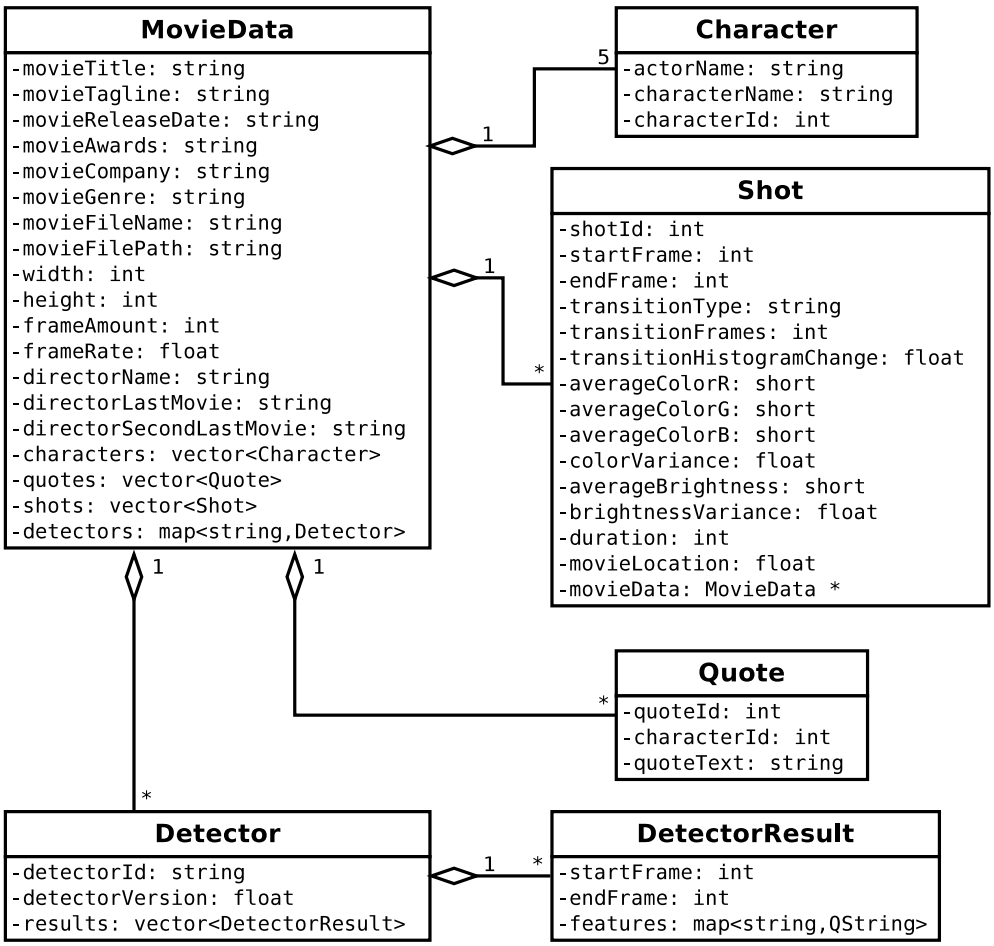

Figure 5.12 UML-Diagram of the Structure MovieData

are the title, the tagline, the release date, the awards won and the genre. Furthermore, information about the director and the five most important characters is stored here. Originally, this information has been retrieved from the Internet Movie Database (see also chapter 3.2.7.2). Important quotes from the movie are stored in the structure *Quote* which contains the id of the character and the text itself.

Technical parameters of the video file are saved, such as the file name and its full path, the resolution of the video and the duration in total amount of frames.

Besides this metadata about the movie, the extracted features are stored in a map

as a string-*Detector*-pair. The string contains the id of the corresponding detector which allows fast access. The *Detector* data structure contains a vector with all found occurrences of the current feature. Every *DetectorResult* in this vector consists of a start and an end frame and a map linking the id of the feature with its value. This map is stored in the variable *features*.

5.4.1.4 Additional Data

Additionally, the sound archive is used to store information about the audio files containing the individual music and sound effects, such as the duration, the file name, their proposed use and their source.

In order to avoid unnecessary recalculations, time-consuming processes use caching for their results. Among these processes are the categorization, the thumbnail extraction and the textual 3d-animations. The categorization cache saves the categories and the corresponding footage clips. A checksum algorithm is used to verify that the cached results are still valid.

Whenever the application requires thumbnail previews, the caching system checks whether the desired thumbnail already exists. If that is not the case, the extraction is started. For each video file, a folder of the same name is created in the *thumbnails* subdirectory. The absolute frame number is used as file name for each thumbnail.

The animation module calculates a hash value for every animation request based on the given parameters, such as the resolution, the animation style and the actual text. This hash is used as file name for the video. Whenever an animation is requested the system first looks for the hashed file name and only starts the actual animation process if the video file is not found.

Finally, the currently open categories and abstracting model files are stored in a preferences cache, as well as the current movie configuration. This allows the users to continue their work upon application restart.

5.4.2 Application Modules

The video abstracting application is divided into several modules. These modules and the data flow within the application are depicted in figure 5.13.

GUI The graphical user interface acts as controlling part for the whole application. All processes are initiated by this module at startup or through user interaction.

The module passes the file paths for the categories, the abstracting model and the movie- and sound archive configuration to the data handler module. In turn,

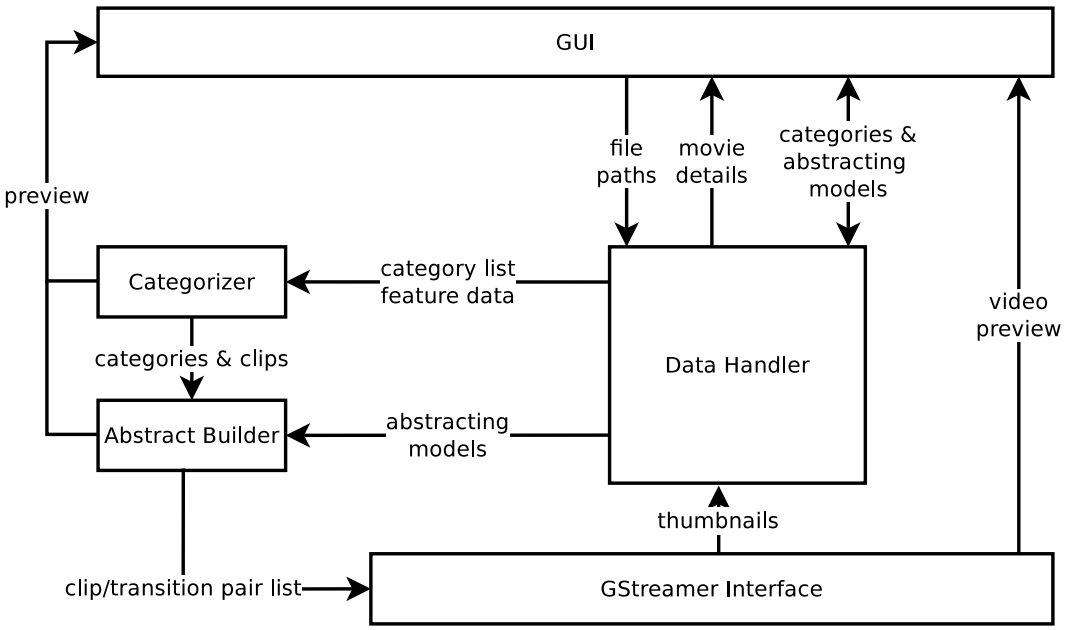

Figure 5.13 Overview of the Application Structure and Main Data Flow

the user interface requests the current movie list, the current categories and the abstracting models from the data handler and displays the information. Modified categories or abstracting models can be saved back to the data handler and stored into XML files as well.

Data Handler The module is responsible for parsing the XML files and holding the extracted data in instances of the data structures described in chapter 5.4.1. Among these data are the categories (see figure 5.9), the movie data and the extracted features (see figure 5.12), and the abstracting models (see figure 5.10). Furthermore, the data handler provides methods for requesting thumbnails and for saving the data instances back to XML files.

GStreamer Interface The GStreamer interfaces handles all media playback and mixing operations. It connects with the video playback widgets in the user interface to display previews of categorized clips in the categorizer widget (see chapter 5.2.2.1) and to show preliminary video abstracts in the video abstract viewer widget (see chapter 5.2.2.3). Furthermore, the data handler module utilizes the GStreamer

interface for the extraction of thumbnails from video files.

The GStreamer interface module uses the GStreamer library[4] which provides various media playback and modification methods. Additionally, the GStreamer Editing Services (GES)[5] are used to allow non-linear editing and arrangements of video and audio files. This framework provides various video and audio effects as well.

Categorizer The categorizer module performs the filtering of the footage utilizing the currently loaded category definition. The category list and the feature data of the current movie are retrieved from the data handler module. Using the algorithm described in chapter 5.3.1, the footage is categorized into the categories. At the end of this process, each category contains a list of video clips with corresponding footage.

This list of video clips is sent to the user interface and displayed in the corresponding category widget. Optionally, each clip can be represented by a thumbnail.

Abstract Builder In the abstract builder module, the steps for the generation of the video abstract are performed. This involves the derivation of a list of clip/transition pairs, the selection of actual footage video clips and audio segments and the generation of textual animations. Finally, all these media are assembled into a single video.

The clip/transition pair list describing the semantical abstract structure is generated by the algorithm described in chapter 5.3.2. The output of the categorizer module provides the abstract builder module with the categorized footage clips. For each clip in the clip/transition pair list, a video clip from the corresponding category with matching length is selected or a corresponding text animation requested. Each selected footage clip is blacklisted to avoid multiple selection of clips or overlaps of frame ranges.

If a requested animation is not cached already, the animation is rendered using the 3D modeling software Blender.

A list of sound effects accompanying transitions or animations is created along with a list of music parts. This list of clip/transition pairs, the list of sound effects and the list of music parts now point to actual media files and are handed over to the GStreamer interface for setting up the media arrangement.

[4] https://gstreamer.freedesktop.org/

[5] https://gstreamer.freedesktop.org/data/doc/gstreamer/head/gstreamer-editing-services/html/

5.5 Implementation of Video Abstracting Models

The development of abstracting models for horror and comedy trailers used ths information gained in the manual annotation of corresponding trailers (see chapter 4.1.2). Based on this shot-by-shot analysis, two abstracting models for the two genres horror and comedy have been implemented.

5.5.1 Action Trailer Model

The trailer model for action trailers developed in the SVP project [Brachmann et al., 2006] is shown in table 5.1. Due to its complex non-deterministic structure, the three alternative action phases are shown separately in table 5.2, table 5.3 and table 5.4.

Content	Category	Duration	Comment
Intro Phase			
Setting Introduction	setting	2.4-20s	repeated 2-3 times
Loud Quote	quote	1.6-40s	or vice versa
Protagonist Close-Up, Silent	character1closeupsilent	0.8-1.2s	
Person Close-Up, Silent	personcloseupsilent	0.8-1.2s	optional
Story Phase			
Tagline	tagline	1.32s	
Person Silent	personsilent	0.8-1.2s	
Slow Action, Silent	slowAction	0.6-1.0s	optionally slow-motion
Quote	quote	1.6-40s	
Setting Introduction	setting	2.4-20s	
Company Animation	company	1.32s	optional
Protagonist Close-Up, Silent	character1closeupsilent	0.8-1.2s	optionally slow-motion
Slow Action, Silent	slowAction	0.6-1.0s	optionally slow-motion
Protagonist Close-Up, Silent	character1closeupsilent	0.8-1.2s	optionally slow-motion
Quote	quote	1.6-40s	
Person Silent	personsilent	0.8-1.2s	
Protagonist, Silent	character1silent	0.8-1.2s	
Person Close-Up, Silent	personcloseupsilent	0.8-1.2s	
Slow Action, Silent	slowAction	0.6-1.0s	optionally slow-motion

Protagonist Close-Up, Silent	character1closeupsilent	0.8-1.2s	optionally slow-motion
Slow Action, Silent	slowAction	0.6-1.0s	optionally slow-motion
Director Animation	directorproducer	1.32s	optional
Setting Introduction	setting	2.4-20s	
Quote	quote	1.6-40s	
Slow Action, Silent	slowAction	0.6-1.0s	optionally slow-motion
Break			
LongQuote	quoteLong	2.4-40.0s	
Action Phase			
Action Phase A *or* Action Phase B1 or B2 Action Phase B1 or B2 *or* Action Phase C			see table 5.2 see table 5.3 see table 5.3 see table 5.4
Outro Phase			
Quote	quoteLong, quote	1.6-40.0s	optionally
title	title	2.64s	or vice versa
Spectacular	spectacular		optionally
credits	credits	4.8s	

Table 5.1 Clip Arrangement in the Abstracting Model of Action Trailers

Content	Category	Duration	Comment
Action Phase A			
Protagonist Close-Up, Silent	character1closeupsilent	0.8-1.2s	
Fast Action	fastAction	0.2-0.6s	
Gunshot	gunshot	0.4-1.0s	
Slow Action, Silent	slowAction	0.6-1.0s	
Fast Action	fastAction	0.2-0.6s	repeated two times
Actor Name	actor	1.32s	
Protagonist, Silent	character1silent	0.8-1.2s	
Slow Action, Silent	slowAction	0.6-1.0s	
Actor Name	actor	1.32s	
Person Close-Up, Silent	personcloseupsilent	0.8-1.2s	
Gunshot	gunshot	0.4-1.0s	optionally two times
Person Close-Up, Silent	personcloseupsilent	0.8-1.2s	
Actor Name	actor	1.32s	
Person Close-Up, Silent	personcloseupsilent	0.8-1.2s	
Gunshot	gunshot	0.4-1.0s	optionally two times
Person Close-Up, Silent	personcloseupsilent	0.8-1.2s	
Gunshot	gunshot	0.4-1.0s	
Fast Action	fastAction	0.2-0.6s	repeated four times
Scream, Silent	scream	0.4-1.0s	
Gunshot	gunshot	0.4-1.0s	
Slow Action, Silent	slowAction	0.6-1.0s	repeated three times
Gunshot	gunshot	0.4-1.0s	
Explosion	explosion	0.4-0.8s	
Slow Action, Silent	slowAction	0.6-1.0s	
Person Close-Up, Silent	personcloseupsilent	0.8-1.2s	
Slow Action, Silent	slowAction	0.6-1.0s	
Gunshot	gunshot	0.4-1.0s	

Table 5.2 Action Trailer Model, Action Phase Variant A

Content	Category	Duration	Comment
Action Phase B1			
Person Silent	personsilent	0.8-1.2s	
Explosion	explosion	0.4-0.8s	repeated two times
Fast Action	fastAction	0.2-0.6s	
Person Silent	personsilent	0.8-1.2s	
Fast Action	fastAction	0.2-0.6s	
Person Close-Up, Silent	personcloseupsilent	0.8-1.2s	
Protagonist, Silent	character1silent	0.8-1.2s	
Fast Action	fastAction	0.2-0.6s	
Explosion	explosion	0.4-0.8s	
Fast Action	fastAction	0.2-0.6s	
Person Silent	personsilent	0.8-1.2s	
Action Phase B2			
Slow Action, Silent	slowAction	0.6-1.0s	repeated three times
Fast Action	fastAction	0.2-0.6s	repeated two times
Explosion	explosion	0.4-0.8s	
Fast Action	fastAction	0.2-0.6s	
Explosion	explosion	0.4-0.8s	
Fast Action	fastAction	0.2-0.6s	
Slow Action, Silent	slowAction	0.6-1.0s	
Person Close-Up, Silent	personcloseupsilent	0.8-1.2s	
Person Silent	personsilent	0.8-1.2s	
Gunshot	gunshot	0.4-1.0s	
Fast Action	fastAction	0.2-0.6s	

Table 5.3 Action Trailer Model, Action Phase Variant B

Content	Category	Duration	Comment
Action Phase C			
Protagonist Close-Up, Silent	character1closeupsilent	0.8-1.2s	
Fast Action	fastAction	0.2-0.6s	
Gunshot	gunshot	0.4-1.0s	
Slow Action, Silent	slowAction	0.6-1.0s	
Fast Action	fastAction	0.2-0.6s	repeated two times
Protagonist, Silent	character1silent	0.8-1.2s	
Slow Action, Silent	slowAction	0.6-1.0s	
Person Close-Up, Silent	personcloseupsilent	0.8-1.2s	
Gunshot	gunshot	0.4-1.0s	optionally two times
Person Close-Up, Silent	personcloseupsilent	0.8-1.2s	
Action Phase B1 or B2			see table 5.3
Spectacular	spectacular		

Table 5.4 Action Trailer Model, Action Phase Variant C

5.5.2 Semantic Footage Categories

The existing categories from the SVP project were defined according to the needs of action trailers. However, additional categories were required for the newly chosen genres. The previously described structure of comedy trailers for example shows that dialogues and other speech-based sequences are crucial for these trailers. This required corresponding categories to be defined and parameters to be chosen.

The categories and the features used to segment the footage for horror and comedy trailers are shown in table 5.5. Compared to the existing categories from the SVP project, the temporal footage location is also taken into account.

ID	Comment
speechpart	speech-break-speech pattern, no disturbing music, between 3 and 10 seconds duration
speechpart_introduction	see category speechpart, footage location between 2%-30% of the movie
speechpart_story	see category speechpart, footage location between 30%-50% of the movie
speechpart_plotpoint2	see category speechpart, footage location between 60%-80% of the movie
character1closeupsilent	main actor, face probability 50-100%, face size between 0.2-0.8, no speech, duration 20-30 frames
character1silent	main actor, face probability 50-100%, face size between 0-0.2, no speech, duration 20-30 frames
character1speaking	main actor, face probability 50-100%, face size between 0-0.2, speech between 0.6-1.0, duration 25-100 frames
fastAction	motion magnitude between 0.4 and 1.0, no text and duration between 10 and 50 frames
personcloseupsilent	face probability 0.7-1.0%, face size between 0.2-0.8, no speech, duration 20-30 frames
personsilent	face probability 0.7-1.0%, face size between 0.0-0.2, no speech, duration 20-30 frames
quote_end	quote probability between 0.7-1.0, speech between 0.1-1.0, duration between 40-120 frames, footage location between 50%-100% of the movie
quote_beginning	quote probability between 0.7-1.0, speech between 0.1-1.0, duration between 40-120 frames, footage location between 0%-50% of the movie
quoteLong	quote probability between 0.7-1.0, speech between 0.1-1.0, duration between 100-175 frames
scream	scream probability 60%-100%, no text, duration between 10-50 frames, footage location between 1%-99% of the movie
scream_long	scream probability 60%-100%, no text, duration between 10-75 frames, footage location between 1%-99% of the movie
setting	sound volume between 0%-20%, no faces, no actors, no speech, no text, movement magnitude between 0-0.3, duration 25-100 frames, footage location between 20%-30% of the movie
slowAction	movement magnitude between 0.1-0.4, no text, duration between 25-50 frames
spectacular	sudden volume increase between 0.15-0.5

Table 5.5 Clip categories used for abstract generation

The clips contained in these categories resemble the basic footage units for the generation of the trailers.

ID	Comment
greenscreen	the approval notice often found preceding a trailer
company	the production company's name and -animation
blackframes	blank screen, black frames
tagline	the tagline of the movie
awards	awards won by the movie
directorproducer	the director or producer of the movie
actor	the main actors of the movie, up to 5
title	the movies title
credits	the trailer's credits

Table 5.6 Animation categories used for abstract generation

Additionally, table 5.6 shows a list of supported animation clips to supplement the generated video abstracts.

5.5.3 Creating Models

The previous sections described a generic structure of horror and comedy trailers and the categories of available video clips. In order to automatically generate trailers using the application developed in this thesis (see chapters 4.2 and 5.2 - 5.4), abstracting models need to be defined. These models describe *how* a trailer is structured and it can be generated employing the categorized video clips (see chapter 4.1.3) according to the generic trailer structures (see chapter 5.5.2).

Horror Trailers In table 5.7, the final model for the generation of horror trailers is shown. Two different story phases have been defined which reflect the arrangement style found in the two manually analyzed trailers. The sound archive for horror trailers consists of several different music excerpts to be used in the four trailer phases. The excerpts have been selected to reflect the properties of the soundtrack of the original trailer which were found during the analysis.

Content	Category	Duration	Comment
Intro Phase			
Company	company	≈2.5s	repeated up to 3 times
Setting/Character Introduction	setting, personsilent, speechpart_introduction	≈2.8-7.7s	up to 2 times
Story Phase (a)			
Tagline	tagline	≈4s	tagline of movie
Protagonist Speaking	character1speaking	≈11s	from 20%-50% of movie timeline
Quote	quote_beginning	≈11s	from 20%-50% of movie timeline
Story Phase (b)			
Slow Action	slowAction	≈0.5-0.7s	repeated up to 14 times
Black Frames	blackframes	≈0.3s	
Quote	quote_beginning	≈2s	
Protagonist Silent	character1closeupsilent	≈2s	
Quote	quote_beginning	≈2.5-3s	
Action Phase			
Actor Name	actor	1s	repeated 5 times
Fast Action	fastAction	≈0.35s	
Director Name	directorproducer	≈1.3s	
Spectacular	spectacular	≈0.4-1.7	
Person Closeup Silent	personcloseupsilent	≈0.6s	
Person Screaming	scream	≈0.5-0.7s	
Outro Phase			
Speech Part	speechpart_plotpoint2	≈2.4s	from climax of movie timeline
Quote	quote_end	≈5-11s	
Slow Action	slowAction	≈1s	
Movie title	title	≈2.7-3.2s	
Spectacular	spectacular	≈1s	
Credits	credits	≈5s	

Table 5.7 Clip Arrangement in the Abstracting Model of Horror Trailers

Comedy Trailers The trailer model for comedy trailers is shown in table 5.8. It consists mainly of speech part clips surrounded by transitional shots. These transitional shots fill the gaps between the speech part clips and belong to categories like *setting*, *slowAction* or *fastAction*. Shorter clips showing people are used as well. Additionally, typical trailer clips like the opening green screen, company logos and actor and director names are framing the structure.

Content	Category	Duration	Comment
Intro Phase			
Green Screen	greenscreen	2.33-4.21s	
Black Frames	blackframes	≈1s	
Company Logo	company	1.63-3.08s	
Protagonist Silent	character1silent	≈2.8-7.7s	up to two times
Speech Part	speechpart_introduction	≈2-5s	
Setting/Fast Action	setting, fastAction	≈0.5-1.0s	repeated 4 times
Humorous Scene	speechpart_introduction	≈2.7-3.3s	
Story Phase			
Protagonist Silent	character1silent	≈1.38s	
Humorous Scene	speechpart_plotpoint2		
Transitional Shot	setting	≈2.75s	
Tagline	tagline	≈2.75s	replacing a voice-over narrator
Action Phase			
Transitional Shot	slowAction	≈0.5-1.0s	repeated ≈8 times
Humorous Scene	speechpart	≈2.7-3.3s	
Actor Name	actor	≈1s	
Protagonist Silent	character1silent	≈0.5-1.0s	
Director Name	directorproducer	≈1.3s	
Outro Phase			
Movie title	title	≈2,7-3,2s	
Humorous Scene	speechpart	≈3,7-19s	
Credits	credits	≈5s	

Table 5.8 Clip Arrangement in the Abstracting Model of Comedy Trailers

The sound clips chosen for the background music of comedy trailers are more dominant and contain more pop music than the other genres. This corresponds to the findings from the analysis in chapter 4.1.2.

The trailer models were built and edited with the video abstracting application developed in this thesis. In figure 5.14 an excerpt of the actual abstracting model for horror trailers is shown. On top, the phasePattern level with the node *Outro_Horror* can be seen. This phase contains one sequencePatternList, called *Outro_Horror* as well. This sequencePatternList is shown right below the phasePattern in the same node with a light blue background. The child node below represents the sequencePattern *Outro_Horror* and the only child of type clipTransitionPairList named *Outro_Horror_fadeblack*. This sequencePatternList finally consists of the clips and transitions constituting a part of the finale trailer. In the example in figure 5.14, clips from the categories *quote_end*, *slowAction*, *title* and *spectacular* are concatenated via *fade_black* transitions.

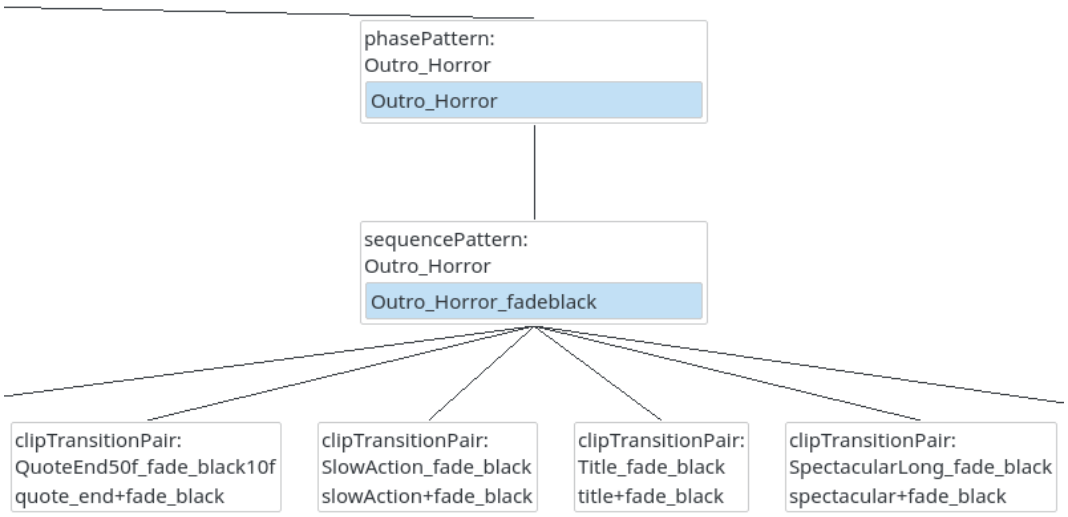

Figure 5.14 Abstracting Model Excerpt showing parts of the outro phase of the horror trailer model.

5.6 Additional Functions

Besides the main video abstracting application a few other tools and modules had to be developed as well in order to support and ease the abstraction process.

5.6.1 Speech Part Detection

During the modeling of Comedy trailers, it was quite difficult to locate dialogue sequences. Such dialogues often contain jokes or funny parts. Based on the output of the character audio detection (see chapter 3.2.7.2), a corresponding detector module was added to the already existing analyzing modules of the SVP system (see chapter 3.2.7.2). The features provided by this tool are the total duration of speech as well as the duration of the break.

Although intended for comic sequences, this module is also quite usable to find dialogue sequences in general. Combined with a filter to specific regions of the film's timeline (such as the plot points or ranges of the acts described in chapter 2.1.1), sequences of higher semantical value can be found.

5.6.2 Analyzer Module Controller

In order to prepare movies for the abstracting application, the automatic analysis using also the detector modules from the SVP project (see chapter 3.2.7.2) need to be performed. The resulting individual XML files have to be merged into a single one for each movie as well. As the set of movies was quite large, a script to automate this process was developed.

It invokes the following steps for each movie:

1. Extract the movie sound track
2. Run all the detector modules in the correct order and provide the required data
3. Validate the results
4. Merge all the individual feature XML files into a single one

The script takes the video file as parameter and produces the corresponding feature XML file (shown in figure 3.1) as output.

5.6.3 Trailer Annotation Player

The manual shot-based annotation of trailers is a very time-consuming process, as the trailer has to be stopped after each shot and the timestamps need to be noted. Often, a shot has to be replayed to gather all its properties. To assist the manual analysis a special video player was developed. It is based on the same programming libraries as the main application. Besides a video file this video player uses an additional shot list created by the shot detection module (see chapter 3.2.7.2) as input. Utilizing this information, the player allows the user to play and replay each shot individually. This eases and accelerates the annotation process considerably.

5.6.4 SVP Data Converter

During the development, sample categories and abstracting models were required for testing and as basis for the newly implemented trailer generation models (see chapter 5.5). To convert the models and categories defined via the Protégé editor in the SVP project into the new XML-based data format a converter script was written. It takes the Protégé instances file (*.pins) as input and produces two XML files, one for the trailer model and one for the category definition.

Chapter 6

Evaluation

In this chapter the newly created video abstracting application and the automatically created video abstracts are evaluated and rated.

6.1 Video Abstracting Application

Among common evaluation techniques for software systems are user studies and expert interviews which compare the usability and performance of the software to competing products. However, the developed software performs highly specific tasks, so similar products are very rare and not easily available (see chapter 3). Furthermore, experts in using video abstracting software are quite rare as well. This lack of competitive systems makes it very difficult to measure performance. Thus a descriptive evaluation of the product was performed, in addition also a rating of its output, the trailers it produced.

The overall objective of the application is to improve the video abstracting process. In chapter 4.2.2, the following requirements of the new application were defined:

1. Single application
2. Possibility to interact
3. Easier editing of categories and abstracting models

These requirements have been achieved and are described in detail in chapter 5 by means of workflow and user interface description. The application performs the steps of video abstracting specified by Truong and Venkatesh (see chapter 3.1.2): *excerpt segmentation*, *excerpt selection*, *excerpt shortening*, *multimodal integration* and *excerpt assembly*. Excerpt segmentation is done in the categorizer module, while excerpt selection and -shortening are performed by the abstract builder module (see chapter 5.4). The final steps of multi-modal integration and excerpt assembly are controlled and executed by the abstract viewer module and the GStreamer interface.

6.2 Automatically Generated Trailers

A number of common evaluation methods for video abstracts has been developed in recent works that are summarized below. In the next part of this chapter the setup of the evaluation as well as its outcome is described.

6.2.1 Evaluation Methods for Video Abstracts

For the evaluation of video abstracts three major approaches were described in chapter 3.1.3: *Result description*, *objective metrics* and *user studies*.

The result description is very common and simple as no comparison to other systems is performed. However, such an evaluation can be quite subjective and the results may not reflect the general performance of a method.

The problem with objective metrics is that they depend on some sort of fidelity function in order to measure success. Such a function would most likely focus on certain viewpoints and would as such also be prone to subjectivity. Additionally, a ground truth is needed for this kind of evaluation. It is also difficult to match human judgment with such metrics.

According to Truong and Venkatesh [2007], a user study would be the most realistic and useful method for evaluating video abstracts. However, it is difficult to set up and also highly subjective in regard to the audience's individual perception. Truong and Venkatesh [2007] also mentioned the difficulty in rating a video abstract's quality for test subjects.

6.2.2 Scenario

To evaluate the automatically generated trailers, a result description method is used. As a trailer is a complex multi-modal construct, choosing important criteria is crucial for the evaluation. Those can be divided into two questions: How good is the trailer in fulfilling the purposes that it is meant to achieve and how good is its technical quality?

According to Hediger [2001] (see chapter 2.2) trailers are supposed to:

- make the audience aware of the upcoming movie
- show what the audience may expect by
- convince people to watch the movie
- build up suspense

Contemporary trailers fulfill these functions by showing a model of the first 2/3 of the film. The content of the movie is communicated by dialogue excerpts and narrative voice-overs. The video track of a trailer follows the music and additional video sequences are used to illustrate the sound track.

Furthermore, Hediger [2001] mentions the following formal parameters of contemporary movie trailers (see chapter 2.2.3 as well):

- Duration: 120-150 seconds
- Average shot length 1.4 seconds
- Suspense plot
- Anonymous narrator
- Key art
- Grid cut
- Text overlays
- Title and actor names towards the end

6.2.2.1 Criteria

Based on these requirements more detailed questions can be formulated regarding the semantic level:

- Does the trailer transfer the mood of the movie (like fear, anxiety and mystery in horror; humor, slapstick and romance in comedy)?
- To what extent does the trailer summarize the first 2/3 of the story?
- Does the trailer build up a suspense plot (containing setting and character introduction, does it show the main conflicts, problems or challenges of the protagonists)?

Furthermore, the following questions regarding the technical quality should be asked:

- Are desired video sequences chosen (content- and footage location wise, is the duration correct)?
- Does the soundtrack fit?
- Is speech clearly perceivable and not interrupted due to false cuts?
- Are texts readable and displayed for the correct duration?
- Does the trailer comply with the formal parameters listed in chapter 6.2.2?

Trailer ID	Movie	Genre	Source
T1	Dreamcatcher	horror	automatic
T2	Dawn of the Dead	horror	automatic
T3	Bruce Almighty	comedy	automatic
T4	10 Things I Hate About You	comedy	automatic
O1	Dreamcatcher	horror	Hollywood
O2	Dawn of the Dead	horror	Hollywood
O3	Bruce Almighty	comedy	Hollywood
O4	10 Things I Hate About You	comedy	Hollywood

Table 6.1 Trailer and Corresponding Movie Selection for the Evaluation

6.2.2.2 Trailer Selection

Besides the metrics used, the evaluation requires material to be rated. For each of the two genres horror and comedy, two automatically generated trailers are analyzed. One is made for a movie that was annotated manually (see chapter 4.1.2) and the other one for a new movie. This allows us to see how the automatic trailers perform compared to the respective Hollywood trailers and how they perform in general.

For the genre horror the movies *Dreamcatcher* (2003, by Lawrence Kasdan) and *Dawn of the Dead* (2004, by Zack Snyder), and for the genre comedy *Bruce Almighty* (2003, Tom Shadyac) and *10 Things I Hate About You* (1999, by Gil Junger) were selected. For each of them, a trailer using the video abstracting application was generated. For clarity reasons, the automatically generated trailers were named T1-T4 while the original Hollywood trailers are referred to as O1-O4 (see table 6.1).

The original Hollywood trailers O1 and O3 have been used in the implementation process of the abstracting models (see chapter 5.5). The trailers O2 and O4 have not been not used previously. T1 and T2 were generated using the abstraction model *horror* and the sound archive *horror*, while the abstraction model *comedy* and the sound archive *comedy* were selected for the generation of T3 and T4.

6.2.3 Semantics

Each of the four automatically generated trailers were examined according to the criteria described in chapter 6.2.2.1. Table 6.1 shows the four automatically generated trailers.

The structure consisting of the four phases intro, story, action and outro which can

be found in the Hollywood trailers (see chapter 4.1.2), is present in the automatically generated trailers as well.

In the intro phase, the main characters and the setting are presented. The story phase introduces the relations, conflicts or problems of the movie story. The action phase consists of shorter and faster clips rising the tension up to the climax and presenting the actors and the director. Finally, the outro phase shows the movie title and uses slower and calmer clips to bring down the tension and leave an impression. The credits conclude the outro phase.

In comparison to the Hollywood trailers, the automatically generated ones contain simpler arrangements of video and audio clips. Complex cut patterns are not employed, such as the separation of video and audio in grid-cut. Also, the length of the automatic trailers is shorter. T1 has a duration of 105 sec, T2 of 90 sec, T3 of 111 sec and T4 of 113 sec, compared to O1 with a duration of 139 sec, O2 and O3 with 149 sec and O4 with 146 sec.

The comparison was performed on a semantical level, because the automatic trailers are not meant to be a rather exact copy but to follow the implemented structure of the original trailers. All the trailers were broken down into semantical units of three different kinds, similar to the manual analysis in chapter 4.1.2. The first category of semantical units is called *Speech* and contains noticeable dialogues or statements which transfer story information and may be rather long in duration. The second category consists of shorter sequences, such as setting or character introducing clips, which fill the gaps between the other sequences and contain less story information. This category is called *Filler*. The final semantic category consists of animated texts, such as the movie title or actor names, and is named *Animations*. The comparison is done phase-by-phase and separated into horror and comedy trailers.

6.2.3.1 Horror Trailers

Intro Phase In figure 6.1 the intro phases of the two automatically generated horror trailers are visualized and compared to the intro phases of the corresponding theatrical Hollywood trailers. All trailers start with animations, some with the greenscreen (sequences E1, G1 and H1), others directly with one or more film company animations: While the Hollywood trailers O1 and O2 feature multiple animations and company names (sequences F1, F3 and H2), the automatic trailers T1 and T2 only show one (sequences E2 and FG2). This is because the meta data contain only one company entry per movie. T1 continues with a shot showing a protagonist (sequence E4) and a dialogue sequence of two main characters (sequence E5). They are talking about the character *Duddits*. This introduction of main characters is similar to the

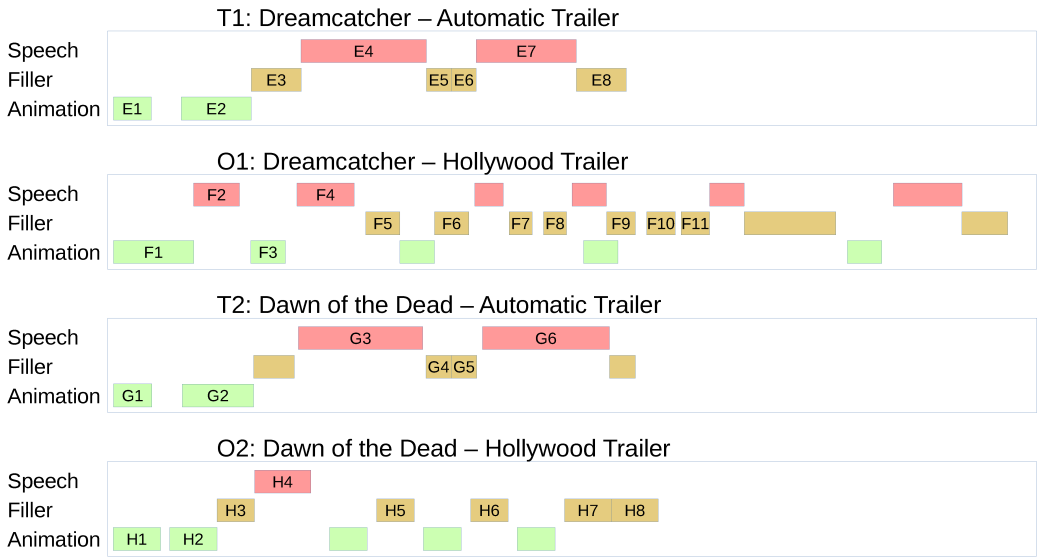

Figure 6.1 Comparison of Semantical Units in the Intro Phase of Horror Trailers

original trailer O1 (sequences F2 and F4), even the topic of the conversations is the same. After two shots of the snowy forest and other characters (sequence E5 and E6), the intro phase of T1 continues with another dialogue (sequence E7) between the two protagonists shown before. The end of this phase is marked by a setting shot showing the forest (sequence E8). This sequence is also similar to the footage in the original trailer O1 (sequences F5-F11).

Compared to the original trailer O2, the automatic trailer T2 has a different pattern in the intro phase. Both start with the greenscreen (sequences G1 and H1) and animations (sequences G2 and H2) of the movie company (although O2 has two company animations). The original trailer O2 then uses a longer scene showing American daily suburban life of a family (sequences H3-H8). Instead T2 starts in medias res (see chapter 2.1.6) showing a woman in a blood-covered shirt and a police officer pointing a gun at her (sequence G3). Separated by black frames, the two following shots introduce further characters (sequences G4 and G5). The end of the intro phase in T2 also presents a main threat of the story of the film by including a scene of the same police officer stating “maybe they are coming for us” (sequence G6). Although the methods of introducing the zombie apocalypse differ in O2 and T2, both trailers fulfill this narrative purpose.

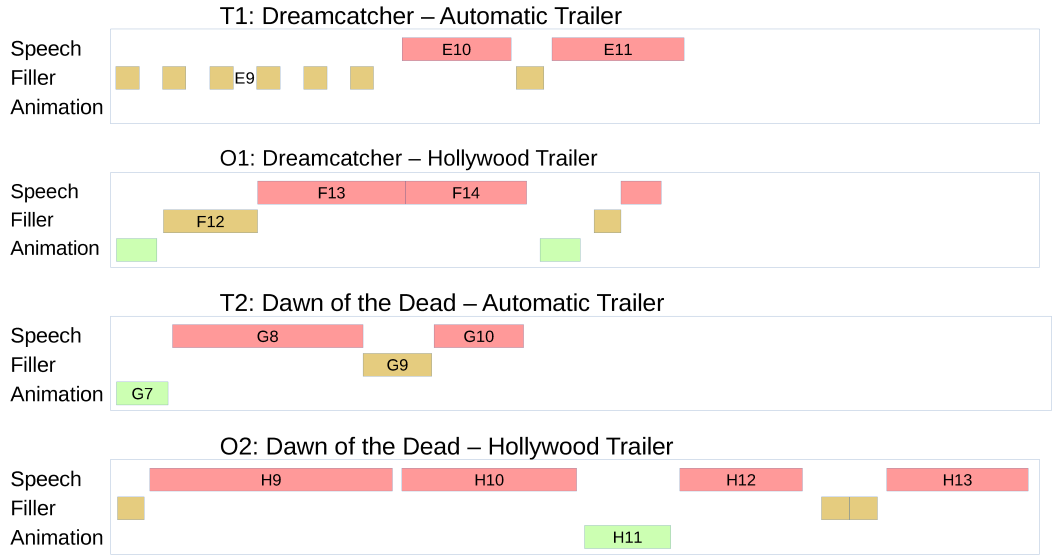

Figure 6.2 Comparison of Semantical Units in the Story Phase of Horror Trailers

In the intro phases of both automatically generated trailers, the main characters are shown and introduced. T1 also features a spoken reference to the character Duddits. In both trailers the music and cut pattern create the desired mood and begin to build up suspense which fulfills the requirements on this phase.

Story Phase The story phases of the horror trailers are illustrated in figure 6.2. The automatic trailers employ two different patterns in this phase. T1 builds up tension with a pattern of alternating black frames and filler clips (sequence E9), showing shots of different characters. The story phase of T1 ends with two speech scenes, one showing two characters in a tense situation (sequence E10) and the other one showing a monologue of a man walking through the snowy forest (sequence E11).

The corresponding original trailer O1 has a similar structure, starting with several filler shots accompanied by a voice-over narrator explaining background information (sequence F12). Later on, more information is given by diegetic voice-overs from the antagonist (sequence F13 and F14). Such story-summarizing narration is missing in the automatic trailer.

The story phase of the second automatic trailer T2 uses a semantical pattern more similar to the corresponding Hollywood trailer O2. It begins with a textual

animation of the tagline of the movie (sequence G7) which is also used in the original trailer O2 (although not in the beginning but in sequence H11). The remainder of the story phase is combined out of three scenes showing a dispute between characters (sequence G8), how they try to survive (sequence G9) and how they fight against zombies in an abandoned mall (sequence G10).

In comparison, the Hollywood trailer O2 starts with the surprising appearance of the daughter who has turned into a zombie (sequence H9). The narration then follows the female character from the intro phase while she realizes the apocalypse (sequence H10). Radio broadcasts are used to transfer story information via sound bridges (sequence H11). In subsequent shots, the journey of the female character into the same mall shown in T2 is told (sequence H12). Finally, the gathering of survivors at the mall is shown in O2 (sequence H13).

The story phase surfaces one of the problems of the automatically generated trailers, their semantical structure is less continuous than in the original ones. Also T1 and T2 suffer from the missing narration, either non-diegetic (like the narrator in T1) or non-diegetic (like the radio broadcast in T2). Although the story phases of the two trailers intensify the mood and continue to build suspense, they have shortfalls in regard to storytelling which is not surprising as it is not contained in the abstracting models.

Action Phase The action phases of the horror trailers are compared in figure 6.3. The automatic trailer for T1 uses a pattern of alternating actor names (sequence E12) and different action shots from the snowy forest (sequence E13), combined with mysterious music. Also, the director's name appears after five iterations (sequence E14). Towards the end, a spectacular shot is showing one of the main characters being attacked by a monster (sequence E15). Two more character and action shots follow (sequence E16). Because the action phase of the trailer model is based on the structure of O1, the action phase of T1 corresponds to the one of O1.

In the trailer T2, a similar concept is chosen. The pattern of actor name animations and action sequences is used as well (sequence G11 and G12), followed by the director's name (sequence G13). Three short action sequences form the end of the action phase (sequence G14). The original trailer O2 has a more complex arrangement. It starts with alternating blank screen and action shots (sequence H14). Diegetic voice-over narration is illustrated by setting and action sequences which show the arrival of zombies at the mall (sequence H15). In the following sequences, the characters talk about the possibility of zombies getting in (sequence H16). The second half of the action phase of O2 consists of a montage of action shots com-

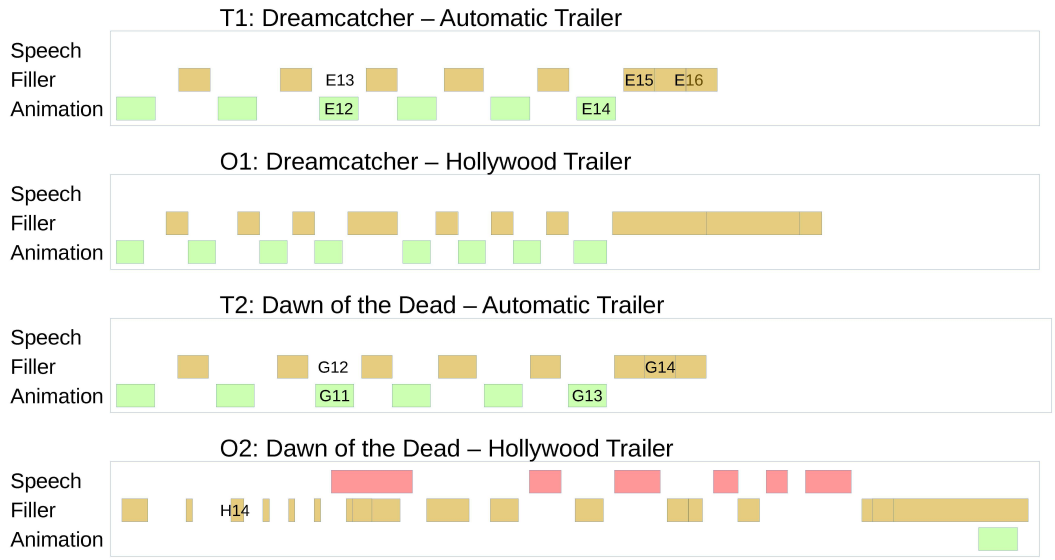

Figure 6.3 Comparison of Semantical Units in the Action Phase of Horror Trailers

bined with sound effects (sequence H17). Towards the end, the image seems to be projected from a film strip and the strip starts to melt (sequence H18).

The model for the action phase is the same in T1 and T2, so a similar arrangement can be expected. It is also found in O1. However, the original Hollywood trailer O2 has a different structure and has a focus on telling more story information than O1. The fast-paced pattern of names, black frames and short filler shots of the two automatic trailers intensifies the suspense until it reaches its climax, also the music reaches a peak. Thus the action phases correspond well to the desired requirements.

Outro Phase Figure 6.4 visualizes the outro phases of the four trailers. The outro phases of T1 and T2 are based on the same model as well and thus share a common structure. They open with two speech-focused sequences with less music which slows down the tempo and show further semantically interesting footage (sequences E17, E18, G14 and G15). A short filler action shot (sequence E19) marks the transition towards the movie title animation (sequence E20) and the credits (sequence E22). Between those two text animations, a final spectacular footage scene (sequence E21) is inserted to act as button (see chapter 2.2.3). This structure can also be found in O1, while the outro phase of O2 simply consists of the textual animations of

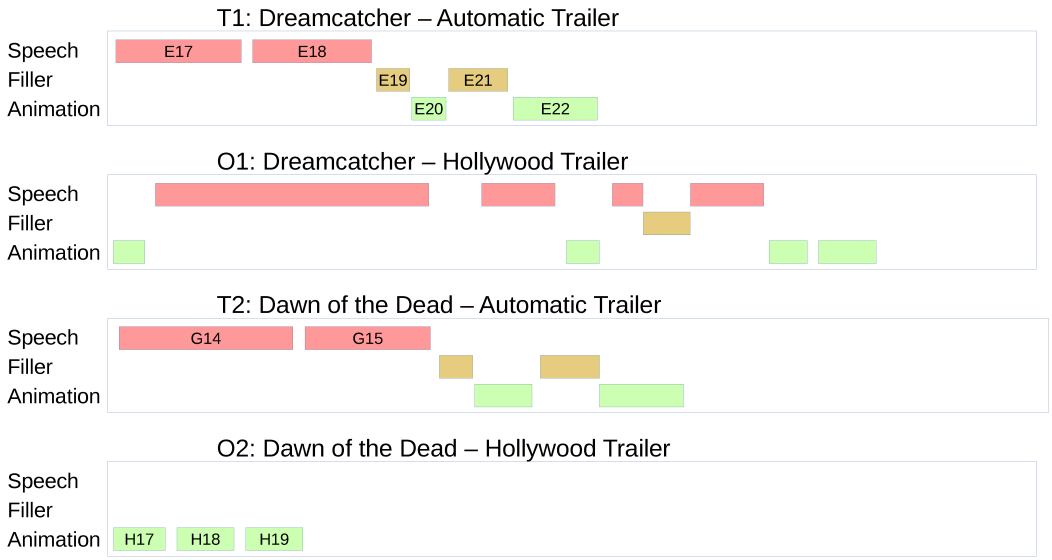

Figure 6.4 Comparison of Semantical Units in the Outro Phase of Horror Trailers

the movie title (sequence H17), the release date (sequence H18) and the credits (sequence H19).

Both outro phases of the automatic trailers serve the intended purposes of reducing the pace and calming down the trailer after the climax by showing longer and calmer speech-based sequences. The music slows down and is calmer as well. The movie title is shown as expected and a final spectacular button scene is included in both trailers. Both automatic trailers are well ended by the credits.

6.2.3.2 Comedy Trailers

In comparison to action and horror trailers, the structure of comedy trailers is similar but less distinctive. The trailers are also dividable into the four phases but their appearance is not as striking. In the analyzed comedy trailers (see also chapter 4.1.2), the phases are composed of longer speech-focused sequences, separated by character or setting filler shots.

The evaluation of comedy trailers was performed in the same way as for the horror trailers. Two trailers (T3 and T4) were built automatically and compared to the corresponding original Hollywood trailers (O3 and O4). T4 and O4 are made for a completely new movie, while the source for T3 and O3 has served as basis for the

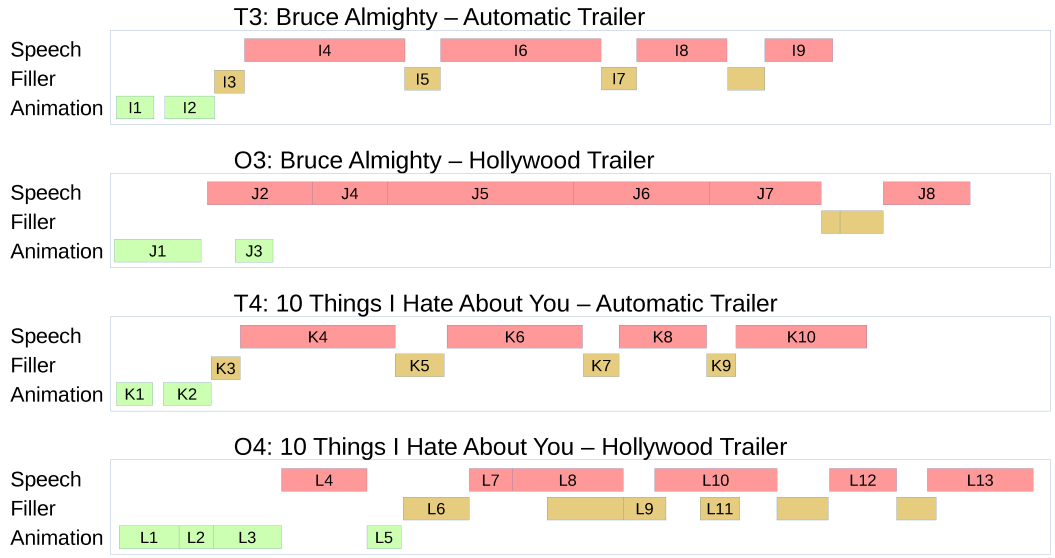

Figure 6.5 Comparison of Semantical Units in the Intro Phase of Comedy Trailers

creation of the comedy trailer models.

Intro Phase Figure 6.5 illustrates the semantic structure of the intro phases in the comedy trailers. Both the automatic trailers (T3 and T4) and the original Hollywood trailers (O3 and O4) start with a green screen (sequences I1, J1, K1, L1) and logo(s) of the producing film companies (sequences I2, J1, J3, K2, L2). The automatic trailer for T3 continues with a pattern of alternating scenes and filler shots. It starts with a filler shot introducing the main character Bruce (sequence I3). The next scene shows the work routine of Bruce in a TV station and how his boss sends him to the Niagara falls to film a documentary (sequence I4). After a humorous filler shot (sequence I5), this scene continues with two sequences containing footage from the Niagara falls (sequences I6 and I8). In between, another filler shot (sequence I7) introduces Grace, Bruce's girlfriend. Towards the end of the intro phase, a scene (sequence I9) shows that Bruce is not happy in the current situation of his life.

Compared to the original trailer O3, T3 has a less continuous plot. The intro phase of O3 is organized according to his daily routine. It starts with him waking up (sequence J2), showing him walking his dog (sequence J4) and having difficult times at work (sequence J5). Another scene shows how he is being hunted by a mob

(sequence J6) and finally he complains towards god why he is hating him (sequences J7 and J8).

After the opening text animations (sequences K1 and K2), the intro phase of the automatic trailer T4 continues with a filler scene introducing the setting of the film by showing a party (sequence K3). In the next speech scene (sequence K4) two students named Cameron and Michael are talking about a girl on a school yard. The girl is Bianca, the sister of the female main character Kat. The next two scenes introduce the father of Bianca and Kat (sequence K5) and provide background information through a conversation between Kat and the father (sequence K6). A shot showing a party (sequence K7) forms the transition to a humorous scene (sequence K8). The final part of the intro phase opens with a filler shot showing Kat (sequence K9) and a scene showing her shouting towards a biker (sequence K10).

The intro phase of the Hollywood trailer O4 in contrast features a prominent voice-over non-diegetic narrator, who is filling the informational gaps between the speech-based scenes. It starts with a scene (sequence L4) of a conversation between Bianca and another girl. After a text animation (sequence L5), the beginning of a relationship between Bianca and Cameron is shown (sequences L6 and L7), similar to T4 (sequence K4). This scene is followed by a scene of the father talking about his view on dating (sequence L8). By showing two filler scenes (sequence L9) of Kat, her character is introduced. A dialogue scene between the director of the school and Kat continues the introduction (sequence L10). Similar to T4 (sequence K10) a filler sequence shows her driving a car (sequence L11). Towards the end of the introduction phase, the two sisters have a discussion with their father (sequence L12) during which he points out that Bianca is only allowed to date if Kat does as well (sequence L13).

The automatic trailers begin with the desired structure by showing a greenscreen and company logo. Furthermore, the intro phases fulfill their purpose of introducing the protagonists and their situation as well as the settings of the stories. They also establish a comic atmosphere by showing humorous scenes and thus provide a good start for the automatic trailers.

Story Phase The story phase of the comedy trailers is depicted in figure 6.6. In T3 the story phase opens with a filler shot of Bruce (sequence I10). This shot is followed by a scene showing an argument between Bruce and his girlfriend (sequence I11). A humorous filler shot of Bruce in a coffee shop (sequence I12) leads over to the animated tagline of the movie (sequence I13). The story phase of O3 is composed

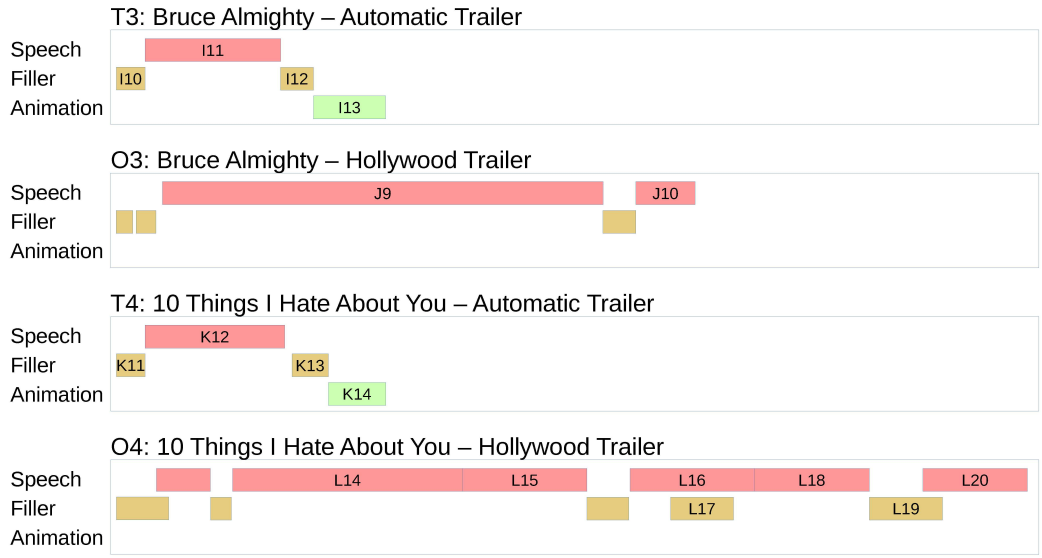

Figure 6.6 Comparison of Semantical Units in the Story Phase of Comedy Trailers

of a larger scene showing a dialogue between Bruce and God (sequence J9) which introduces the premise of the story: God is giving Bruce his powers to show that he can do the job better. The next two shots show Bruce using his powers for the first time in a coffee shop (sequence J10).

In T4 the story phase starts with a filler shot of Bianca (sequence K11) and a sequence showing a car passing by at night (sequence K12). After a setting scene showing a classroom (sequence K13), the story phase ends with the animation of the movie tagline (sequence K14).

The corresponding original trailer O4 has a much longer story phase, composed out of seven main scenes. In the first scenes, Cameron and Michael try to find a boy to date Kat (sequence L14) and choose the character Patrick (sequence L15). The scene also introduces Patrick and shows how the two friends try to convince him to date Kat (sequence L16). This scene is illustrated by a filler shot showing Kat tackling another girl during school sports (sequence L17). The second part of the story phase of O4 deepens the character Kat by showing a dialogue between Kat and the director (sequence L18). Footage from both scenes is also included in the automatic trailer T4. After a transitional filler shot of a party (sequence L19), the

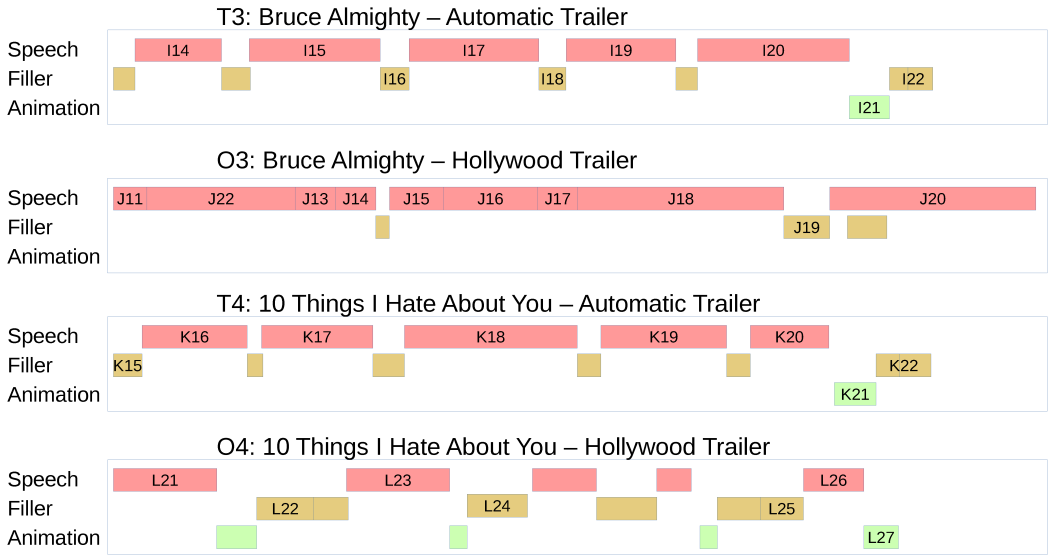

Figure 6.7 Comparison of Semantical Units in the Action Phase of Comedy Trailers

final scene of the story phase shows Patrick flirting with Kat at the party (sequence L20).

Similar to the automatically generated horror trailers the story phase in automatically generated comedy trailers is rather problematic. Although composed of speech-based sequences and action sequences, the storyline of the movie is not clearly visible. A good scene is shown in T3 containing a conflict between Bruce and his girlfriend but T4 includes only general sequences. Especially the lack of a narration is disadvantageous and only partly covered by the animated tagline.

Action Phase The action phases share a more similar semantical structure which is shown in figure 6.7. In most trailers speech-based scenes alternate with shorter filler sequences.

In the automatic trailer T3 the action phase starts with the introduction of his girlfriend's background as a kindergarten teacher (sequence I14). In the next scene, God is making jokes by telling Bruce that he is dead (sequence I15). The following short filler shot shows Bruce in a coffee shop (sequence I16). Both scenes are also used in the corresponding Hollywood trailer O3 (sequences J20 and J10). The

following scenes show Bruce getting beaten up by an angry mob (sequence I17), the same incident is shown in O3 as well (sequence J6). The following shot (sequence I18) shows Bruce being mad at God and is also part of the original trailer O3 (sequence J8). In the next scene, Bruce is talking to his girlfriend (sequence I19). The final longer speech-based sequence shows a humorous scene of Bruce at work in an argument with a colleague (sequence I20). After a textual animation of the actor's name (sequence I21), the action phase closes with two shorter filler sequences (sequence I22) showing the Bruce and his girlfriend. The trailer O3 has a rather dominant narrative arc, compared to the automatic (T3) one. It starts with Bruce trying out his new powers by letting water burst out of hydrants (sequence J11) and by increasing the breasts of his girlfriend (sequence J12). In between God asks if he is "having fun?" (sequence J13). Several more scenes show Bruce walking on the sea (sequence J14), pulling the moon closer for a romantic evening (sequence J15) and answering e-mails sent to God (sequence J16). He is also using it to make his dog use the toilet (sequence J17), to get revenge at work (sequence J18) and finally to split the coffee in his mug and cars on a street to make room for his sports car (sequence J19). In a final dialogue scene above the clouds in the sky, a cliffhanger is shown in which God saying to Bruce he might be dead (sequence J20). As mentioned before, this scene is also part of T3.

The action phase of T4 starts with a short sequence of Kat and Patrick (sequence K15). It is followed by a conversation between Bianca and her father about her dress for the prom (sequence K16). The next sequences show the development of the relationship between Kat and Patrick (sequence K17). A humorous scene shows Cameron and Micheal looking for a guy to date Kat (sequence K18), a scene also found in the original trailer O4 (sequence L4). The remainder of the action scene consists mainly of two scenes focusing on the relationship between Kat and Patrick, first on a party (sequence K19) and later in a car (sequence K20). After the textual animation of the main actor (sequence K21), this phase ends with two more filler shots showing Kat and Patrick (sequence K22).

The corresponding original trailer starts with a scene of Bianca and her father (sequence L21). After a text insert, a humorous scene of Cameron and Michael riding a bike down a hill follows (sequence L22). The trailer then returns the focus onto the developing relationship of Kat and Patrick by employing several speech-based scenes (sequence L23) and transitional filler shots (sequence L24). These sequences are supported by a voice-over narrator. Towards the end, the relationship between Cameron and Bianca is picked up as well by a scene showing them kissing in a car (sequence L25) and by a conversation between Cameron and Patrick talking about it (sequence L26). Finally, another textual animation (sequence L27) concludes the

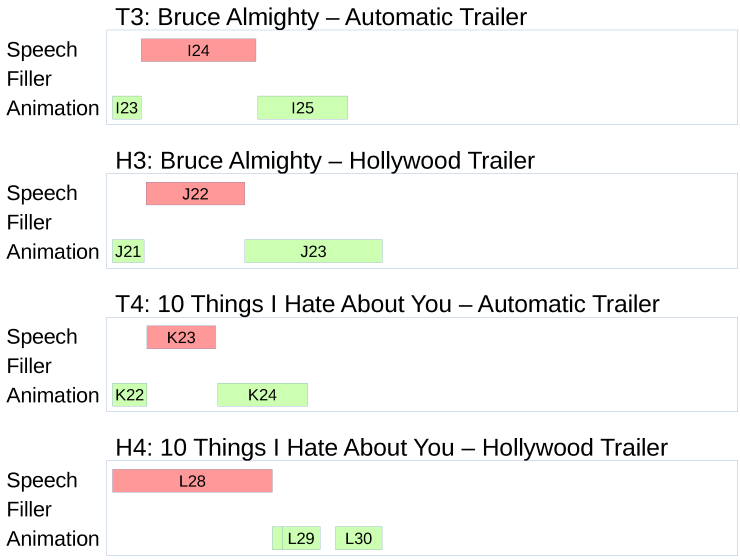

Figure 6.8 Comparison of Semantical Units in the Outro Phase of Comedy Trailers

action phase.

The action phases of the two automatic trailers both show several humorous and speech-based scenes which present further insights into the movie. The trailer T3 includes the climax scene of O3 and T4 includes parts of the storyline of the movie by showing the development of the relationship between Kat and Patrick which is advantageous. The name of the main actor is shown as well. Since the pattern of the automatic trailers in this phase is similar to the original Hollywood trailers the results are satisfying as well.

Outro Phase The outro phases of the comedy trailers are all very similar, as illustrated in figure 6.8. Except for the Hollywood trailer O4, all trailers contain the movie title (sequences I23, J21, K22, L29) in form of a graphical animation, a final humorous scene forming the button (sequences I24, J22, K23, L28) and the credits (sequences I25, J23, K24) or release information (sequence L30).

T3 shows Bruce reporting live from an asteroid impact site (sequence I24), whereas in O3 the scene shows his girlfriend being surprised by the dog sitting on the toilet and reading the newspaper (sequence J22). The final humorous scene of T4 shows

Patrick singing a song to impress Kat (sequence K23). In O4, the final humorous scene is located before the title and shows how another guy is trying to impress Bianca (sequence L28). This trailer also did not feature credits but a "coming soon"-text insert instead (sequence L30).

The end of T3 and T4 contains all desired elements: the movie title, a final spectacular scene acting as button and the credits and thus fulfills all requirements.

6.2.4 Technical Quality

So far, the evaluation of the automatic trailers was done on a semantical level. However, the technical — or syntactical — quality of the trailers is an important criterion as well. Shortfalls in this domain can be noticed quite easily and may disturb the viewing experience.

In general, the quality is quite good. However, some automatic trailers still have problems with speech-based sequences. A few sequences do not contain the begin or end of a spoken sentence and thus are not very well understandable or sound irritating. In some cases the footage does contain unwanted music which may interfere with either spoken text or the supplemented trailer music. A similar problem exists with movie footage containing undesired text, such as the movie title, names of actors or the credits which the text detection module (see chapter 3.2.7.2) failed to identify.

These problems occur to the quality of the input data of the application. Although the detection modules should allow the application to avoid specific footage material, false detections result in such shortfalls in the automatic trailers.

Table 6.2 shows a comparison of the formal parameters for contemporary Hollywood trailers described by Hediger [2001] (see chapter 2.2.4) to the four automatically generated trailers. The average duration of the automatic trailers is shorter because of their simpler structure. The average shot length is comparable for horror trailers (1.3-1.8 sec) and longer (2.2-2.9 sec) for comedy trailers because of the focus on longer dialogue sequences and the missing grid cut. Key art depends on the individual movie, so a generic realization is hardly possible. Grid cut is not used as the application does not support it. Text overlays in form of the tagline, the movie title and the actor's name(s) are included in every automatically generated trailer. Also the button is included in each of them.

Parameter	Hollywood	Automatic
Duration	120-150 sec	90-113 sec
Average shot length	1.4 sec	1.27-2.94 sec
Suspense plot:	yes	partly
Anonymous narrator:	82.3%	no
Key art:	63.6%	no
Grid cut:	60.7%	no
Text overlays	53.4 %	yes
Title:	90.5%	yes
Actor name(s):	64.6%	yes
Button	22%	yes

Table 6.2 Comparison of Formal Parameters of Trailers

6.2.5 Summary

The automatically generated trailers share a less complex structure by means of narration and by means of multi-modal arrangement. They are based on the syntactical patterns found in the original Hollywood trailers and thus share these patterns. Hollywood trailers may use other forms of such patterns and the automatic ones can only use the patterns encoded in the abstracting models.

Compared to the Hollywood trailers, the automatic ones have less clearly visible story lines. Besides the difficulty in selecting proper footage, this is due to the lack of an accompanying narration, either via a non-diegetic voice over or via sound bridges and grid cut in the automatic trailers. Such narration helps in transferring knowledge and filling informational gaps. Such voice-overs can partly be substituted by text animations, such as the movie tag line. Grid cut is currently not implemented in the abstracting application and would require additional concepts for selecting suitable illustrative clips.

The focus on speech-based footage based on its temporal position in the movie, required due to the lack of other distinctive syntactical features, works quite well in the comedy trailers. The automatic horror trailers use this approach for interesting footage selection as well, but they also profit from the syntactical categories used in the previous generation of the action trailer. Especially in the action part of the horror trailers, fast-action and spectacular clips contribute to the overall impression. Simple black sequences also contribute a lot the atmosphere of horror trailers.

The soundtrack of the automatic trailers matches typical trailer music. The sound

parts arranged in horror trailers provide a mysterious and frightening atmosphere which supports the purpose of these trailers a lot. Similar, the casual sounds used in the automatic comedy trailers helps their positive and humorous ambiance.

Although the automatic trailers have some flaws regarding the cut points and audible arrangements they provide an acceptable preview and transfer the mood successfully.

Chapter 7

Conclusion and Future Work

Intention In this thesis the already existing approach for the automatic generation of action trailers was extended towards other genres bearing more semantical information.

After the introduction, an insight into the domain of movies, genres and trailers was given by means of excursuses in chapter 2. Typical steps of film production were described, important terms and perspectives of genre theory introduced and the history and structure of trailers presented. The next chapter focused on the scientific domain of video abstracts and movie trailers in particular. A choice of video abstracting systems dealing with movies were described in chapter 3. Especially the SVP approach [Brachmann et al., 2006], which served as basis for this thesis, was presented in detail.

In chapter 4 and chapter 5, the approach of this work and its implementation was described. A new interactive application for the categorization of movie footage, for design and implementation of video abstracting models, and for the final assembly of video abstracts was developed. This application allows for the creation of two new trailer models for the genres horror and comedy.

The two genres horror and comedy were chosen because they reflect a more semantical level of storytelling. In a manual analysis of two horror and two comedy trailers, generic genre structures were identified and used as a basis for the implementation of two new trailer models. While the SVP action trailer model uses a rather large set of different categories and specialized detection modules, the trailer models developed in this thesis combined a smaller set of general features with temporal information (e.g. the speech part detection module). This combination of syntactical categories; for example speaking characters, with a temporal position constraint allows for finding interesting semantical parts of the movie for dialogue-based movie genres. Additionally, new music and sound effects have been selected and added.

To evaluate the performance of the new video abstracting application two horror

and two comedy trailers were generated using the previously implemented trailer models. These four trailers were evaluated and compared on a semantical level against the corresponding original Hollywood trailers for the same movies. In the evaluation in chapter 6 the automatic trailers showed satisfying results and are comparable to the original trailers by their syntactical and semantical structure to a certain degree. However, the automatic trailers have shortfalls due to the quality of the feature extraction. The face detection for example returns a large number of false negatives and may mark a frame range as not containing faces, although there are some. This makes it quite difficult to extract non-character sequences from a movie. Similar problems exist in the character audio module and the music detection. As a consequence, the system has sometimes difficulties in determining the right start and end point for sequences.

Future Work The new video abstracting application may easily be enhanced by employing improved or new detector modules. Furthermore, it has been shown that new genres can be added by simply adapting existing trailer models or by creating entirely new ones. The software may also be used in other domains than movie trailers by simply creating corresponding abstracting models.

By modifying the application, other use cases can be covered as well, such as content retrieval by using the categorizer part. The system could also be adapted to export a recommendation for a video abstract to other video editing systems.

As the analysis of movie trailers and corresponding literature pointed out, many Hollywood trailers make use of an external narrator. The inclusion of such a feature into the automatic video abstracting software would thus allow for more realistic trailers to be generated.

The generation of trailers is a form of art, which requires a human editor to understand the story of the underlying movie in order to select and form appropriate material. To reach the same quality level, a computer system also needs more understanding of semantics, as required in in many computer science research domains.

Bibliography

Lalitha Agnihotri, Nevenka Dimitrova, and John R. Kender. Design and evaluation of a music video summarization system. In *2004 IEEE International Conference on Multimedia and Expo*, volume 3, pages 1943–1946, June 2004. doi: 10.1109/ICME.2004.1394641.

Lalitha Agnihotri, John Kender, Nevenka Dimitrova, and John Zimmerman. Framework for personalized multimedia summarization. In *Proceedings of the 7th ACM SIGMM International Workshop on Multimedia Information Retrieval*, MIR '05, New York, NY, USA, 2005. ACM. doi: 10.1145/1101826.1101841.

K. Aizawa, K. Ishijima, and M. Shiina. Summarizing wearable video. In *Proceedings. 2001 International Conference on Image Processing*, volume 3, pages 398–401, 2001. doi: 10.1109/ICIP.2001.958135.

Rick Altman. *Film/Genre*. British Film Inst., London, 1999.

Shigemi Aoyagi, Ken'ich Kourai, Koji Sato, Toshihiro Takada, Toshiharu Sugawara, and Rikio Onai. Implementation of flexible-playtime video skimming. volume 5305, 2003. doi: 10.1117/12.538808.

Yasuo Ariki, Masahito Kumano, and Kiyoshi Tsukada. Highlight scene extraction in real time from baseball live video. In *Proceedings of the 5th ACM SIGMM International Workshop on Multimedia Information Retrieval*, MIR '03, New York, NY, USA, 2003. ACM. doi: 10.1145/973264.973297.

Alia Asaad, Serkan Bahceci, Vivien Chan, Sarah Felsmann, Alexander Gryanik, Jorge Hey, Bjoörn Hinze, Seyedeh Niloufar Hosseini, Boris Kamenov, Alexander Mann, Mohammad Reza Mossadegh, Franklin Okwor, Haoyue Qiu, André Rust, Paul Schwermer, Yuping Shi, Berit Steenbock, Philipp Steiner, Fuyi Sun, Nadir Sunar, Ayse Taneri, Till von Wenzlawowicz, Malte Wirkus, Daniel Wu, Arne Jacobs, Thorsten Hermes, and Otthein Herzog. ADDiCT - Automatic Dramaturgy Detection in order to Create Trailers - Projektbericht, 2008.

Jürgen Assfalg, Marco Bertini, Carlo Colombo, Alberto Del Bimbo, and Walter Nunziati. Semantic annotation of soccer videos: automatic highlights identification. *Computer Vision and Image Understanding*, 92(2?3), 2003. doi: http://dx.doi.org/10.1016/j.cviu.2003.06.004.

N. Babaguchi. Towards abstracting sports video by highlights. In *2000 IEEE International Conference on Multimedia and Expo*, volume 3, pages 1519–1522, 2000. doi: 10.1109/ICME.2000.871056.

N. Babaguchi, Y. Kawai, and T. Kitahashi. Generation of personalized abstract of sports video. In *IEEE International Conference on Multimedia and Expo 2001.*, pages 619–622, Aug 2001. doi: 10.1109/ICME.2001.1237796.

A. Bagga, Jianying Hu, Jialin Zhong, and G. Ramesh. Multi-source combined-media video tracking for summarization. In *Proceedings 16th International Conference on Pattern Recognition*, volume 2, pages 818–821, 2002. doi: 10.1109/ICPR.2002.1048428.

Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *Proceedings of the 9th European Conference on Computer Vision - Volume Part I*, ECCV'06, Berlin, Heidelberg, 2006. Springer-Verlag. doi: 10.1007/11744023_32.

D. Bordwell, J. Staiger, and K. Thompson. *The Classical Hollywood Cinema: Film Style and Mode of Production to 1960*. Taylor & Francis, 2003.

David Bordwell and Kristin Thompson. *Film art: an introduction*. McGraw-Hill, New York, 9. ed. edition, 2010.

Jean-Yves Bouguet. *Pyramidal Implementation of the Lucas Kanade Feature Tracker*. Intel Corporation, Microprocessor Research Labs, 1999.

Christoph Brachmann, Hashim Chunpir, Silke Gennies, Benjamin Haller, Philipp Kehl, Astrid Paramita Mochtarram, Daniel Möhlmann, Christian Schrumpf, Christopher Schultz, Björn Stolper, Benjamin Walther-Franks, Arne Jacobs, Thorsten Hermes, and Otthein Herzog. Semantic Video Patterns - Project Report, May 2006.

William F Brewer and Edward H Lichtenstein. Event schemas, story schemas, and story grammars. *Center for the Study of Reading Technical Report; no. 197*, 1980.

Edward Buscombe. The idea of genre in the american cinema. *Film genre reader III*, pages 12–26, 2003.

Rui Cai, Lie Lu, Hong-Jiang Zhang, and Lian-Hong Cai. Highlight sound effects detection in audio stream. In *Proceedings of the 2003 International Conference on Multimedia and Expo - Volume 3*, ICME '03, Washington, DC, USA, 2003. IEEE Computer Society.

J. Calic and E. Izuierdo. Efficient key-frame extraction and video analysis. In *Proceedings. International Conference on Information Technology: Coding and Computing.*, pages 28–33, April 2002. doi: 10.1109/ITCC.2002.1000355.

Hyun Sung Chang, Sanghoon Sull, and Sang Uk Lee. Efficient video indexing scheme for content-based retrieval. *Circuits and Systems for Video Technology, IEEE Transactions on*, 9(8):1269–1279, Dec 1999. doi: 10.1109/76.809161.

Peng Chang, Mei Han, and Yihong Gong. Extract highlights from baseball game video with hidden markov models. In *2002 International Conference on Image Processing. Proceedings*, volume 1, pages I–609–I–612, 2002. doi: 10.1109/ICIP.2002.1038097.

S.B. Chatman. *Story and Discourse: Narrative Structure in Fiction and Film.* Cornell paperbacks. Cornell University Press, 1980.

Shu ching Chen, Mei ling Shyu, Min Chen, and Chengcui Zhang. A decision tree-based multimodal data mining framework for soccer goal detection. In *Proc. of IEEE International Conference on Multimedia and Expo*, 2004.

Michael G. Christel, Michael A. Smith, C. Roy Taylor, and David B. Winkler. Evolving video skims into useful multimedia abstractions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '98, New York, NY, USA, 1998. ACM Press/Addison-Wesley Publishing Co. doi: 10.1145/274644.274670.

F. Coldefy and P. Bouthemy. Unsupervised soccer video abstraction based on pitch, dominant color and camera motion analysis. In *Proceedings of the 12th Annual ACM International Conference on Multimedia*, MULTIMEDIA '04, New York, NY, USA, 2004. ACM. doi: 10.1145/1027527.1027588.

F. Coldefy, P. Bouthemy, M. Betser, and G. Gravier. Tennis video abstraction from audio and visual cues. In *2004 IEEE 6th Workshop on Multimedia Signal Processing*, pages 163–166, Sept 2004. doi: 10.1109/MMSP.2004.1436457.

Jim Collins. Genericity in the nineties: Eclectic irony and the new sincerity. *Film theory goes to the movies*, pages 242–263, 1993.

M. Cooper and J. Foote. Summarizing video using non-negative similarity matrix factorization. In *2002 IEEE Workshop on Multimedia Signal Processing*, pages 25–28, Dec 2002. doi: 10.1109/MMSP.2002.1203239.

M. Cooper and J. Foote. Discriminative techniques for keyframe selection. In *2005 IEEE International Conference on Multimedia and Expo*, July 2005. doi: 10.1109/ICME.2005.1521470.

Serhan Dagtas and Mohamed Abdel-Mottaleb. Multimodal detection of highlights for multimedia content. *Multimedia Syst.*, 9(6), June 2004. doi: 10.1007/s00530-003-0130-3.

Gamhewage C. de Silva, Toshihiko Yamasaki, and Kiyoharu Aizawa. Evaluation of video summarization for a large number of cameras in ubiquitous home. In *Proceedings of the 13th Annual ACM International Conference on Multimedia*, MULTIMEDIA '05, New York, NY, USA, 2005. ACM. doi: 10.1145/1101149.1101329.

Daniel DeMenthon, Vikrant Kobla, and David Doermann. Video summarization by curve simplification. In *Proceedings of the Sixth ACM International Conference on Multimedia*, MULTIMEDIA '98, New York, NY, USA, 1998. ACM. doi: 10.1145/290747.290773.

Ryan Alexander Diduck. Ideology and rhetoric in the classical hollywood movie trailer. 2008.

A. Divakaran, R. Radhakrishnan, and K.A. Peker. Motion activity-based extraction of key-frames from video shots. In *Proceedings. International Conference on Image Processing 2002*, volume 1, pages I–932–I–935, 2002. doi: 10.1109/ICIP.2002.1038180.

Ajay Divakaran, Kadir A. Peker, Regunathan Radhakrishnan, Ziyou Xiong, and Romain Cabasson. Video summarization using mpeg-7 motion activity and audio descriptors. In Azriel Rosenfeld, David Doermann, and Daniel DeMenthon, editors, *Video Mining*, volume 6 of *The Springer International Series in Video Computing*, pages 91–121. Springer US, 2003. doi: 10.1007/978-1-4757-6928-9_4.

Anastasios D. Doulamis, Nikolaos D. Doulamis, and Stefanos D. Kollias. A fuzzy video content representation for video summarization and content-based retrieval. *Signal Processing*, 80(6), 2000. doi: http://dx.doi.org/10.1016/S0165-1684(00)00019-0.

N. Doulamis, A. Doulamis, Y.S. Avrithis, and S.D. Kollias. Video content representation using optimal extraction of frames and scenes. In *Proceedings. 1998 International Conference on Image Processing*, volume 1, pages 875–879, Oct 1998. doi: 10.1109/ICIP.1998.723660.

Mark S. Drew and James Au. Video keyframe production by efficient clustering of compressed chromaticity signatures (poster session). In *Proceedings of the Eighth ACM International Conference on Multimedia*, MULTIMEDIA '00, New York, NY, USA, 2000. ACM. doi: 10.1145/354384.354534.

F. Dufaux. Keyframe selection to represent a video, March 23 2004. US Patent 6,711,587.

Ahmet Ekin and A. Murat Tekalp. Framework for tracking and analysis of soccer video. volume 4671, 2002. doi: 10.1117/12.453120.

B. Erol, D.-S. Lee, and J. Hull. Multimodal summarization of meeting recordings. In *Proceedings. 2003 International Conference on Multimedia and Expo*, volume 3, pages III–25–8, July 2003. doi: 10.1109/ICME.2003.1221239.

Brigitte Fauvet, Patrick Bouthemy, Patrick Gros, and Fabien Spindler. A geometrical key-frame selection method exploiting dominant motion estimation in video. In Peter Enser, Yiannis Kompatsiaris, Noel E. O'Connor, Alan F. Smeaton, and Arnold W.M. Smeulders, editors, *Image and Video Retrieval*, volume 3115 of *Lecture Notes in Computer Science*, pages 419–427. Springer Berlin Heidelberg, 2004. doi: 10.1007/978-3-540-27814-6_50.

A.M. Ferman and A.M. Tekalp. Two-stage hierarchical video summary extraction to match low-level user browsing preferences. *IEEE Transactions on Multimedia*, 5(2):244–256, June 2003. doi: 10.1109/TMM.2003.811617.

Syd Field. *The screenwriter's workbook*. A Dell Trade paperback. Dell, New York, 1984.

Bill Fisher. *addttp4 - Text-to-phone software*. NIST, May 2000. URL `https://www.nist.gov/itl/iad/mig/tools`. last visited: April 30, 2017.

Edward Morgan Forster. *Aspects of the Novel*. RosettaBooks, 2010.

G. Genette. *Narrative Discourse: An Essay in Method*. Cornell paperbacks. Cornell University Press, 1983.

D. Gibson, N. Campbell, and B. Thomas. Visual abstraction of wildlife footage using gaussian mixture models and the minimum description length criterion. In *Proceedings. 16th International Conference on Pattern Recognition*, volume 2, pages 814–817, 2002. doi: 10.1109/ICPR.2002.1048427.

A. Girgensohn and J. Boreczky. Time-constrained keyframe selection technique. In *1999. IEEE International Conference on Multimedia Computing and Systems*, volume 1, pages 756–761, Jul 1999. doi: 10.1109/MMCS.1999.779294.

Fred Goldberg. *Motion picture marketing and distribution: getting movies into a theatre near you*. Focal Press, Boston [u.a.], 1991.

Yihong Gong. Summarizing audiovisual contents of a video program. *EURASIP Journal on Advances in Signal Processing*, 2003(2), 2003. doi: 10.1155/S1110865703211082.

Yihong Gong and Xin Liu. Summarizing video by minimizing visual content redundancies. In *IEEE International Conference on Multimedia and Expo*, pages 607–610, Aug 2001. doi: 10.1109/ICME.2001.1237793.

Yihong Gong and Xin Liu. Video summarization and retrieval using singular value decomposition. *Multimedia Syst.*, 9(2), August 2003. doi: 10.1007/s00530-003-0086-3.

Barry Keith Grant. *Film genre: from iconography to ideology*. Short cuts. Wallflower, London [u.a.], 2007.

Lifang Gu, Don Bone, and Graham Reynolds. Replay detection in sports video sequences. In Nuno Correia, Teresa Chambel, and Glorianna Davenport, editors, *Multimedia '99*, Eurographics, pages 3–12. Springer Vienna, 2000. doi: 10.1007/978-3-7091-6771-7_2.

Riad Hammoud and Roger Mohr. A probabilistic framework of selecting effective key-frames for video browsing and indexing, 2000.

Mei Han, Wei Hua, Wei Xu, and Yihong Gong. An integrated baseball digest system using maximum entropy method. In *Proceedings of the Tenth ACM International*

Conference on Multimedia, MULTIMEDIA '02, New York, NY, USA, 2002. ACM. doi: 10.1145/641007.641081.

Seung-Hoon Han and In-So Kweon. Scalable temporal interest points for abstraction and classification of video events. In *IEEE International Conference on Multimedia and Expo*, July 2005. doi: 10.1109/ICME.2005.1521512.

A. Hanjalic. Generic approach to highlights extraction from a sport video. In *Proceedings. 2003 International Conference on Image Processing*, volume 1, pages I–1–4, Sept 2003. doi: 10.1109/ICIP.2003.1246883.

A. Hanjalic, R. L. Lagendijk, and J. Biemond. Automated high-level movie segmentation for advanced video-retrieval systems. *IEEE Trans. Cir. and Sys. for Video Technol.*, 9(4), June 1999. doi: 10.1109/76.767124.

Alan Hanjalic, R.L. Lagendijk, and Jan Biemond. A new method for key frame based video content representation. *Image Databases and Multi-Media Search*, 1998.

M. Hawley. *Structure out of sound*. PhD thesis, MIT, Massachussetts, 1993.

Liwei He, Elizabeth Sanocki, Anoop Gupta, and Jonathan Grudin. Auto-summarization of audio-video presentations. In *Proceedings of the Seventh ACM International Conference on Multimedia (Part 1)*, MULTIMEDIA '99, New York, NY, USA, 1999. ACM. doi: 10.1145/319463.319691.

Vinzenz Hediger. *Verführung zum Film: der amerikanische Kinotrailer seit 1912*. PhD thesis, Marburg, 2001.

Derek Hoiem, Yan Ke, and Rahul Sukthankar. Solar: Sound object localization and retrieval in complex audio environments. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, March 2005.

Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon. *Spoken Language Processing - A Guide to Theory, Algorithm, and System Development*. Prentice Hall PTR, New Jersey, 2001.

Bogdan Ionescu, Patrick Lambert, Didier Coquin, Laurent Ott, and Vasile Buzuloiu. Animation movies trailer computation. In *Proceedings of the 14th Annual ACM International Conference on Multimedia*, MULTIMEDIA '06, New York, NY, USA, 2006. ACM. doi: 10.1145/1180639.1180770.

Go Irie, Takashi Satou, Akira Kojima, Toshihiko Yamasaki, and Kiyoharu Aizawa. Automatic trailer generation. In *Proceedings of the International Conference on Multimedia*, MM '10, New York, NY, USA, 2010. ACM. doi: 10.1145/1873951.1874092.

A. Jaimes, T. Echigo, M. Teraguchi, and F. Satoh. Learning personalized video highlights from detailed mpeg-7 metadata. In *2002 International Conference on Image Processing. Proceedings*, volume 1, pages I–133–I–136, 2002. doi: 10.1109/ICIP.2002.1037977.

A. Joshi, S. Auephanwiriyakul, and R. Krishnapuram. On fuzzy clustering and content based access to networked video databases. In *Eighth International Workshop on Research Issues In Data Engineering, 1998. 'Continuous-Media Databases and Applications'. Proceedings*, pages 42–49, Feb 1998. doi: 10.1109/RIDE.1998.658277.

S.M. Kaminsky. *American film genres: approaches to a critical theory of popular film.* Pflaum Pub., 1974.

Michael A. Smith Takeo Kanade. Video skimming and characterization through the combination of image and language understanding. In *Proceedings of the 1998 International Workshop on Content-Based Access of Image and Video Databases (CAIVD '98)*, CAIVD '98, Washington, DC, USA, 1998. IEEE Computer Society.

Eung Kwan Kang, Sung Joo Kim, and Joon Soo Choi. Video retrieval based on scene change detection in compressed streams. *IEEE Transactions on Consumer Electronics*, 45(3):932–936, Aug 1999. doi: 10.1109/30.793648.

Hong-Wen Kang and Xian-Sheng Hua. To learn representativeness of video frames. In *Proceedings of the 13th Annual ACM International Conference on Multimedia*, MULTIMEDIA '05, New York, NY, USA, 2005. ACM. doi: 10.1145/1101149.1101242.

Yoshihiko Kawai, Hideki Sumiyoshi, and Nobuyuki Yagi. Automated production of tv program trailer using electronic program guide. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*, CIVR '07, New York, NY, USA, 2007. ACM. doi: 10.1145/1282280.1282287.

Changick Kim and Jenq-Neng Hwang. Object-based video abstraction for video surveillance systems. *IEEE Transactions on Circuits and Systems for Video Technology*, 12(12):1128–1138, Dec 2002. doi: 10.1109/TCSVT.2002.806813.

Jae-Gon Kim, Hyun Sung Chang, Kyeongok Kang, Munchurl Kim, Jinwoong Kim, and Hyung-Myung Kim. Summarization of news video and its description for content-based access. *International Journal of Imaging Systems and Technology*, 13(5), 2003. doi: 10.1002/ima.10067.

V. Kobla, D. DeMenthon, and D. Doermann. Detection of slow-motion replay sequences for identifying sports videos. In *1999 IEEE 3rd Workshop on Multimedia Signal Processing*, pages 135–140, 1999. doi: 10.1109/MMSP.1999.793810.

L.J. Latecki, D. de Wildt, and Jianying Hu. Extraction of key frames from videos by optimal color composition matching and polygon simplification. In *2001 IEEE Fourth Workshop on Multimedia Signal Processing*, pages 245–250, 2001. doi: 10.1109/MMSP.2001.962741.

Hun-Cheol Lee and Seong-Dae Kim. Rate-driven key frame selection using temporal variation of visual content. *Electronics Letters*, 38(5):217–218, Feb 2002. doi: 10.1049/el:20020112.

Hun-Cheol Lee and Seong-Dae Kim. Iterative key frame selection in the rate-constraint environment. *Signal Processing: Image Communication*, 18(1), 2003. doi: http://dx.doi.org/10.1016/S0923-5965(02)00089-9.

Hyowon Lee, Alan F. Smeaton, Catherine Berrut, Noel Murphy, Seán Marlow, and Noel E. O'Connor. Implementation and analysis of several keyframe-based browsing interfaces to digital video. In *Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL '00, London, UK, 2000. Springer-Verlag.

Sangkeun Lee and M.H. Hayes. An application for interactive video abstraction. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. Proceedings*, volume 5, pages V–905–8, May 2004. doi: 10.1109/ICASSP.2004.1327258.

Shih-Hung Lee, C.H. Yeh, and C.-C.J. Kuo. Video skimming based on story units via general tempo analysis. In *2004 IEEE International Conference on Multimedia and Expo*, volume 2, pages 1099–1102, June 2004. doi: 10.1109/ICME.2004.1394402.

T. Leitch. *Crime Films*. Genres in American Cinema. Cambridge University Press, 2002.

Ying Li, Shrikanth Narayanan, and C.-C.Jay Kuo. Movie content analysis, indexing and skimming via multimodal information. In Azriel Rosenfeld, David Doermann, and Daniel DeMenthon, editors, *Video Mining*, volume 6 of *The Springer International Series in Video Computing*, pages 123–154. Springer US, 2003. doi: 10.1007/978-1-4757-6928-9_5.

Wen-Nung Lie and Chun-Ming Lai. News video summarization based on spatial and motion feature analysis. In Kiyoharu Aizawa, Yuichi Nakamura, and Shin'ichi Satoh, editors, *Advances in Multimedia Information Processing - PCM 2004*, volume 3332 of *Lecture Notes in Computer Science*, pages 246–255. Springer Berlin Heidelberg, 2005. doi: 10.1007/978-3-540-30542-2_31.

Rainer Lienhart and Jochen Maydt. An extended set of haar-like features for rapid object detection. In *IEEE ICIP*, volume 1, September 2002.

Rainer Lienhart, Silvia Pfeiffer, and Wolfgang Effelsberg. Video abstracting. *Commun. ACM*, 40(12), December 1997. doi: 10.1145/265563.265572.

Rainer Lienhart, Silvia Pfeiffer, and Wolfgang Effelsberg. Automatic trailer production. *Handbook of Multimedia Computing*, 5:361, 1998.

Rainer W. Lienhart. Dynamic video summarization of home video. volume 3972, 1999. doi: 10.1117/12.373569.

Tianming Liu, Hong-Jiang Zhang, and Feihu Qi. A novel video key-frame-extraction algorithm based on perceived motion energy model. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(10):1006–1013, Oct 2003. doi: 10.1109/TCSVT.2003.816521.

Tiecheng Liu and John R. Kender. Optimization algorithms for the selection of key frame sequences of variable length. In Anders Heyden, Gunnar Sparr, Mads Nielsen, and Peter Johansen, editors, *Computer Vision - ECCV 2002*, volume 2353 of *Lecture Notes in Computer Science*, pages 403–417. Springer Berlin Heidelberg, 2002a. doi: 10.1007/3-540-47979-1_27.

Tiecheng Liu and J.R. Kender. An efficient error-minimizing algorithm for variable-rate temporal video sampling. In *2002 IEEE International Conference on Multimedia and Expo. Proceedings*, volume 1, pages 413–416, 2002b. doi: 10.1109/ICME.2002.1035806.

Tiecheng Liu and J.R. Kender. Rule-based semantic summarization of instructional videos. In *2002 International Conference on Image Processing. Proceedings*, volume 1, pages I–601–I–604, 2002c. doi: 10.1109/ICIP.2002.1038095.

Tieyan Liu, Xudong Zhang, Desheng Wang, Jian Feng, and Kwok-Tung Lo. Inertia-based cut detection technique: a step to the integration of video coding and content-based retrieval. In *5th International Conference on Signal Processing Proceedings*, volume 2, pages 1018–1025, 2000. doi: 10.1109/ICOSP.2000.891700.

Tieyan Liu, Xudong Zhang, Jian Feng, and Kwok-Tung Lo. Shot reconstruction degree: a novel criterion for key frame selection. *Pattern Recognition Letters*, 25 (12), 2004a. doi: http://dx.doi.org/10.1016/j.patrec.2004.05.020.

Tieyan Liu, Xudong Zhang, Jian Feng, and Kwok-Tung Lo. Shot reconstruction degree: a novel criterion for key frame selection. *Pattern recognition letters*, 25 (12):1451–1457, 2004b.

S Lu, I King, and MR Lyu. Video summarization using greedy method in a constraint satisfaction framework. In *Proceedings of 9th International Conference on Distributed Multimedia Systems*, pages 456–461, 2003.

Shi Lu, I. King, and M.R. Lyu. Video summarization by video structure analysis and graph optimization. In *2004 IEEE International Conference on Multimedia and Expo*, volume 3, pages 1959–1962, June 2004a. doi: 10.1109/ICME.2004.1394645.

Shi Lu, M.R. Lyu, and I. King. Video summarization by spatial-temporal graph optimization. In *Proceedings of the 2004 International Symposium on Circuits and Systems*, volume 2, pages II–197–200, May 2004b. doi: 10.1109/ISCAS.2004.1329242.

Shi Lu, M.R. Lyu, and I. King. Semantic video summarization using mutual reinforcement principle and shot arrangement patterns. In *Proceedings of the 11th International Multimedia Modelling Conference*, pages 60–67, Jan 2005. doi: 10.1109/MMMC.2005.64.

Yu-Fei Ma, Lie Lu, Hong-Jiang Zhang, and Mingjing Li. A user attention model for video summarization. In *Proceedings of the Tenth ACM International Conference on Multimedia*, MULTIMEDIA '02, New York, NY, USA, 2002. ACM. doi: 10.1145/641007.641116.

Sen Marlow, David A. Sadlier, Noel O'Connor, and Noel Murphy. Audio processing for automatic tv sports program highlights detection, 2002.

K. Masumitsu and T. Echigo. Video summarization using reinforcement learning in eigenspace. In *Proceedings. 2000 International Conference on Image Processing*, volume 2, pages 267–270, Sept 2000. doi: 10.1109/ICIP.2000.899351.

A. Miene, A. Dammeyer, Th. Hermes, and O. Herzog. Advanced and adapted shot boundary detection. In D. W. Fellner, N. Fuhr, and I. Witten, editors, *Proc. of ECDL WS Generalized Documents*, 2001.

Kenichi Minami, Akihito Akutsu, Hiroshi Hamada, and Yoshinobu Tonomura. Video handling with music and speech detection. *IEEE MultiMedia*, 5(3), 1998. doi: http://dx.doi.org/10.1109/93.713301.

Koichi Miura, Reiko Hamada, Ichiro Ide, Shuichi Sakai, and Hidehiko Tanaka. Motion based automatic abstraction of cooking videos. In *Proc. ACM Multimedia 2002 Workshop on Multimedia Information Retrieval*, pages 29–32, 2002.

Shingo Miyauchi, Noboru Babaguchi, and Tadahiro Kitahashi. Highlight detection and indexing in broadcast sports video by collaborative processing of text, audio, and image. *Systems and Computers in Japan*, 34(12), 2003. doi: 10.1002/scj.10493.

Stephen Neale. *Genre*. British Film Inst., London, 1980.

Haung Wei Ng, Y. Sawahata, and K. Aizawa. Summarization of wearable videos using support vector machine. In *2002 IEEE International Conference on Multimedia and Expo. Proceedings*, volume 1, pages 325–328, 2002. doi: 10.1109/ICME.2002.1035784.

Chong-Wah Ngo, Yu-Fei Ma, and Hong-Jiang Zhang. Automatic video summarization by graph modeling. In *Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2*, ICCV '03, Washington, DC, USA, 2003. IEEE Computer Society.

Nosa Omoigui, Liwei He, Anoop Gupta, Jonathan Grudin, and Elizabeth Sanocki. Time-compression: Systems concerns, usage, and benefits. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '99, New York, NY, USA, 1999. ACM. doi: 10.1145/302979.303017.

H. Pan, P. van Beek, and M.I. Sezan. Detection of slow-motion replay segments in sports video for highlights generation. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings*, volume 3, pages 1649–1652, 2001. doi: 10.1109/ICASSP.2001.941253.

K.A. Peker and A. Divakaran. Adaptive fast playback-based video skimming using a compressed-domain visual complexity measure. In *2004 IEEE International Conference on Multimedia and Expo (ICME)*, volume 3, pages 2055–2058, June 2004. doi: 10.1109/ICME.2004.1394669.

K.A. Peker, A. Divakaran, and Huifang Sun. Constant pace skimming and temporal sub-sampling of video using motion activity. In *Proceedings 2001 International Conference on Image Processing*, volume 3, pages 414–417, 2001. doi: 10.1109/ICIP.2001.958139.

M. Petkovic, V. Mihajlovic, W. Jonker, and S. Djordjevic-Kajan. Multi-modal extraction of highlights from tv formula 1 programs. In *2002 IEEE International Conference on Multimedia and Expo. Proceedings*, volume 1, pages 817–820, 2002. doi: 10.1109/ICME.2002.1035907.

N. Peyrard and P. Bouthemy. Motion-based selection of relevant video segments for video summarisation. In *Proceedings of the 2003 International Conference on Multimedia and Expo - Volume 1*, ICME '03, Washington, DC, USA, 2003. IEEE Computer Society.

Silvia Pfeiffer, Rainer Lienhart, Stephan Fischer, and Wolfgang Effelsberg. Abstracting digital movies automatically. *Journal of Visual Communication and Image Representation*, 7(4), 1996. doi: http://dx.doi.org/10.1006/jvci.1996.0030.

S.V. Porter, M. Mirmehdi, and B.T. Thomas. A shortest path representation for video summarisation. In *Proceedings. 12th International Conference on Image Analysis and Processing*, pages 460–465, Sept 2003. doi: 10.1109/ICIAP.2003.1234093.

V. Propp, L. Scott, and L.A. Wagner. *Morphology of the Folktale: Second Edition*. Publications of the American Folklore Society. University of Texas Press, 2010.

Jian quan Ouyang, Jin-Tao Li, and Yong-Dong Zhang. Replay boundary detection in mpeg compressed video. In *2003 International Conference on Machine Learning and Cybernetics*, volume 5, pages 2800–2804, Nov 2003. doi: 10.1109/ICMLC.2003.1260028.

Regunathan Radhakrishnan, Ajay Divakaran, and Ziyou Xiong. A time series clustering based framework for multimedia mining and summarization. In *Proceedings of the 6 th ACM SIGMM International Workshop on Multimedia Information Retrieval*, 2004.

Z. Rasheed and M. Shah. Scene detection in hollywood movies and tv shows. In *Proceedings. 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages II–343–8, June 2003. doi: 10.1109/CVPR.2003.1211489.

Zeeshan Rasheed, Yaser Sheikh, and Mubarak Shah. Semantic film preview classification using low-level computable features. In *3rd International Workshop on Multimedia Data and Document Engineering*, 2003.

M.T. Roeder. *A History of the Concerto*. Amadeus Press, 1994.

Jiawei Rong, Wanjun Jin, and Lide Wu. Key frame extraction using inter-shot information. In *2004 IEEE International Conference on Multimedia and Expo*, volume 1, pages 571–574, June 2004. doi: 10.1109/ICME.2004.1394256.

Roni Rosenfeld and Philip Clarkson. Statistical language modeling using the cmu-cambridge toolkit. 1997.

Yong Rui, Anoop Gupta, and Alex Acero. Automatically extracting highlights for tv baseball programs. In *Proceedings of the Eighth ACM International Conference on Multimedia*, MULTIMEDIA '00, New York, NY, USA, 2000. ACM. doi: 10.1145/354384.354443.

T. Schatz. *Hollywood genres: formulas, filmmaking, and the studio system*. McGraw-Hill, 1981.

Huang-Chia Shih and Chung-Lin Huang. Detection of the highlights in baseball video program. In *2004 IEEE International Conference on Multimedia and Expo*, volume 1, pages 595–598, June 2004. doi: 10.1109/ICME.2004.1394262.

Alan F. Smeaton, Paul Over, and Wessel Kraaij. Trecvid: Evaluating the effectiveness of information retrieval tasks on digital video. In *Proceedings of the 12th Annual ACM International Conference on Multimedia*, MULTIMEDIA '04, New York, NY, USA, 2004. ACM. doi: 10.1145/1027527.1027678.

Alan F. Smeaton, Bart Lehane, Noel E. O'Connor, Conor Brady, and Gary Craig. Automatically selecting shots for action movie trailers. In *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, MIR '06, New York, NY, USA, 2006. ACM. doi: 10.1145/1178677.1178709.

M.A. Smith and T. Kanade. Video skimming and characterization through the combination of image and language understanding techniques. In *1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Proceedings*, pages 775–781, Jun 1997. doi: 10.1109/CVPR.1997.609414.

Vivian Carol Sobchack. *The limits of infinity*. Barnes, South Brunswick, 1980.

Jim Soto. WRITING A SCREENPLAY (The Syd Field Paradigm). URL `http://jimsoto.weebly.com/uploads/3/2/2/0/32208869/12._writing_a_screenplay_syd_field_-_outline.pdf`. last visited: April 30, 2017.

Janet Staiger. Hybrid or inbred: The purity hypothesis and hollywood genre history. *Film genre reader III*, pages 185–199, 2003.

A. Stevenson. *Oxford Dictionary of English*. Oxford Dictionary of English. OUP Oxford, 2010.

Masaru Sugano, Yasuyuki Nakajima, Hiromasa Yanagihara, and Akio Yoneyama. Generic summarization technology for consumer video. In *Proceedings of the 5th Pacific Rim Conference on Advances in Multimedia Information Processing - Volume Part II*, PCM'04, Berlin, Heidelberg, 2004. Springer-Verlag. doi: 10.1007/978-3-540-30542-2_1.

Xinding Sun and Mohan S. Kankanhalli. Video summarization using r-sequences. *Real-Time Imaging*, 6(6), December 2000. doi: 10.1006/rtim.1999.0197.

Hari Sundaram and Shih-Fu Chang. Condensing computable scenes using visual complexity and film syntax analysis. In *PROCEEDINGS OF ICME 2001*, 2001.

Hari Sundaram and Shih fu Chang. Video skims: Taxonomies and an optimal generation framework. In *In Proc. IEEE International Conference on Image Processing 2002*, 2002.

Y. Takahashi, N. Nitta, and N. Babaguchi. Video summarization for large sports video archives. In *2005. ICME 2005. IEEE International Conference on Multimedia and Expo*, pages 1170–1173, July 2005. doi: 10.1109/ICME.2005.1521635.

Yukinobu Taniguchi, Akihito Akutsu, and Yoshinobu Tonomura. Panoramaexcerpts: Extracting and packing panoramas for video browsing. In *Proceedings of the Fifth ACM International Conference on Multimedia*, MULTIMEDIA '97, New York, NY, USA, 1997. ACM. doi: 10.1145/266180.266396.

C.M. Taskiran, Z. Pizlo, A. Amir, D. Ponceleon, and E.J. Delp. Automated video program summarization using speech transcripts. *Multimedia, IEEE Transactions on*, 8(4):775–791, Aug 2006a. doi: 10.1109/TMM.2006.876282.

C.M. Taskiran, Z. Pizlo, A. Amir, D. Ponceleon, and E.J. Delp. Automated video program summarization using speech transcripts. *Multimedia, IEEE Transactions on*, 8(4):775–791, Aug 2006b. doi: 10.1109/TMM.2006.876282.

Laura Teodosio and Walter Bender. Salient video stills: Content and context preserved. In *Proceedings of the First ACM International Conference on Multimedia*, MULTIMEDIA '93, New York, NY, USA, 1993. ACM. doi: 10.1145/166266.166270.

Dian Tjondronegoro, Yi-Ping Phoebe Chen, and Binh Pham. Integrating highlights for more complete sports video summarization. *IEEE Multimedia*, 11(4), 2004a. doi: http://doi.ieeecomputersociety.org/10.1109/MMUL.2004.28.

D.W. Tjondronegoro, Yi-Ping Phoebe Chen, and Binh Pham. Classification of self-consumable highlights for soccer video summaries. In *2004 IEEE International Conference on Multimedia and Expo*, volume 1, pages 579–582, June 2004b. doi: 10.1109/ICME.2004.1394258.

Ba Tu Truong and Svetha Venkatesh. Video abstraction: A systematic review and classification. *ACM Trans. Multimedia Comput. Commun. Appl.*, 3, February 2007. doi: http://doi.acm.org/10.1145/1198302.1198305.

A. Tudor. *Theories of film*. Cinema one. Secker & Warburg [for] the British Film Institute, 1974.

Shingo Uchihashi, Jonathan Foote, Andreas Girgensohn, and John Boreczky. Video manga: Generating semantically meaningful video summaries. In *Proceedings of the Seventh ACM International Conference on Multimedia (Part 1)*, MULTIMEDIA '99, New York, NY, USA, 1999. ACM. doi: 10.1145/319463.319654.

J. Vermaak, P. Perez, A. Blake, and M. Gangnet. Rapid summarisation and browsing of video sequences. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 2002.

Paul Viola and Michael J. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE CVPR*, 2001.

Kongwah Wan and Changsheng Xu. Robust soccer highlight generation with a novel dominant-speech feature extractor. In *2004 IEEE International Conference on Multimedia and Expo*, volume 1, pages 591–594, June 2004a. doi: 10.1109/ICME.2004.1394261.

Kongwah Wan and Changsheng Xu. Efficient multimodal features for automatic soccer highlight generation. In *Proceedings of the 17th International Conference on Pattern Recognition*, volume 3, pages 973–976, Aug 2004b. doi: 10.1109/ICPR.2004.1334691.

Jinjun Wang, Changsheng Xu, Engsiong Chng, and Qi Tian. Sports highlight detection from keyword sequences using hmm. In *2004 IEEE International Conference on Multimedia and Expo*, volume 1, pages 599–602, June 2004. doi: 10.1109/ICME.2004.1394263.

Robert Warshow. *The immediate experience: movies, comics, theatre & other aspects of popular culture*. Harvard University Press, 1962.

N. Wilkens. Detektion von Videoframes mit Texteinblendungen in Echtzeit. Master's thesis, Universität Bremen, 2003.

Linda Williams. Film bodies: Gender, genre, and excess. *Film genre reader III*, pages 141–159, 2003.

W. Wolf. Key frame selection by motion analysis. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing. Conference Proceedings*, volume 2, pages 1228–1231, May 1996. doi: 10.1109/ICASSP.1996.543588.

J.K. Wu, M.S. Kankanhalli, J.H. Lim, and D. Hong. *Perspectives on Content-Based Multimedia Systems*. The Information Retrieval Series. Springer US, 2006.

Yu-Xiang Xie, Xi-Dao Luan, Song-Yang Lao, Ling-Da Wu, Peng Xiao, and Jun Wen. Edu: A model of video summarization. In Peter Enser, Yiannis Kompatsiaris, NoelE. O'Connor, AlanF. Smeaton, and ArnoldW.M. Smeulders, editors, *Image and Video Retrieval*, volume 3115 of *Lecture Notes in Computer Science*, pages 106–114. Springer Berlin Heidelberg, 2004. doi: 10.1007/978-3-540-27814-6_16.

Wei Xiong, John Chung-Mong Lee, and Rui-Hua Ma. Automatic video data structuring through shot partitioning and key-frame computing. *Mach. Vision Appl.*, 10(2), June 1997. doi: 10.1007/s001380050059.

Ziyou Xiong, R. Radhakrishnan, and A. Divakaran. Generation of sports highlights using motion activity in combination with a common audio feature extraction framework. In *Proceedings. 2003 International Conference on Image Processing*, volume 1, pages I–5–8, Sept 2003a. doi: 10.1109/ICIP.2003.1246884.

Ziyou Xiong, R. Radhakrishnan, A. Divakaran, and T.S. Huang. Audio events detection based highlights extraction from baseball, golf and soccer games in a unified framework. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings*, volume 5, pages V–632–5, April 2003b. doi: 10.1109/ICASSP.2003.1200049.

Ziyou Xiong, R. Radhakrishnan, A. Divakaran, and T.S. Huang. Effective and efficient sports highlights extraction using the minimum description length criterion in selecting gmm structures [audio classification]. In *2004 IEEE International Conference on Multimedia and Expo*, volume 3, pages 1947–1950, June 2004. doi: 10.1109/ICME.2004.1394642.

Zhe Xu and Ya Zhang. Automatic generated recommendation for movie trailers. In *2013 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, pages 1–6, June 2013. doi: 10.1109/BMSB.2013.6621738.

Itheri Yahiaoui, Bernard Merialdo, and Benoit Huet. Automatic video summarization. In *Proc. CBMIR Conf*, 2001.

M. M. Yeung and Boon-Lock Yeo. Video visualization for compact presentation and fast browsing of pictorial content. *IEEE Trans. Cir. and Sys. for Video Technol.*, 7(5), 1997. doi: 10.1109/76.633496.

M.M. Yeung and B. Liu. Efficient matching and clustering of video shots. In *International Conference on Image Processing. Proceedings*, volume 1, pages 338–341, Oct 1995. doi: 10.1109/ICIP.1995.529715.

M.M Yeung, C. Li, and R.W. Lienhart, editors. *Scalable hierarchical video summary and search*, volume 4315, 2001. doi: 10.1117/12.410967.

Bin Yu, Wei-Ying Ma, Klara Nahrstedt, and Hong-Jiang Zhang. Video summarization based on user log enhanced link analysis. In *Proceedings of the Eleventh ACM International Conference on Multimedia*, MULTIMEDIA '03, New York, NY, USA, 2003. ACM. doi: 10.1145/957013.957095.

Xiao-Dong Yu, Lei Wang, Qi Tian, and Ping Xue. Multilevel video representation with application to keyframe extraction. In *Proceedings. 10th International Multimedia Modelling Conference*, pages 117–123, Jan 2004. doi: 10.1109/MULMM.2004.1264975.

Hong Jiang Zhang, Jianhua Wu, Di Zhong, and Stephen W. Smoliar. An integrated system for content-based video retrieval and browsing. *Pattern Recognition*, 30 (4), 1997. doi: http://dx.doi.org/10.1016/S0031-3203(96)00109-4.

Xu-Dong Zhang, Tie-Yan Liu, Kwok-Tung Lo, and Jian Feng. Dynamic selection and effective compression of key frames for video abstraction. *Pattern Recogn. Lett.*, 24(9-10), June 2003. doi: 10.1016/S0167-8655(02)00391-4.

Ming Zhao, Jiajun Bu, and Chun Chen. Audio and video combined for home video abstraction. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings*, 2003.

Yueting Zhuang, Yong Rui, T.S. Huang, and S. Mehrotra. Adaptive key frame extraction using unsupervised clustering. In *Proceedings. 1998 International Conference on Image Processing*, volume 1, pages 866–870, Oct 1998. doi: 10.1109/ICIP.1998.723655.