Christoph Brachmann, Hashim Iqbal 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, Otthein Herzog

Automatic Movie Trailer Generation Based on Semantic Video Patterns

Abstract

Automatic video summarization has become an important field of research with the advances in digital audio and video analysis and much effort has been put into movie abstracting for large media databases. However, in the movie industry content summarization for advertising trailers has been perfected to a form of art. In this paper we introduce the approach of automatically generating entertaining Hollywood-like trailers based on a trailer grammar. The extraction of features from movies using state-of-the-art image and audio processing techniques builds the foundation for the selection of meaningful and usable material which is re-assembled according to defined grammar rules. We present a system for generating trailers for contemporary Hollywood action movies. User testing of our automatically produced trailers for this movie genre shows promising results that suggest further research in this field.

1 Introduction

Beside the original intention of a trailer or a teaser – advertising a particular movie – the short preview of a movie has become an attractive movie genre in itself (Kernan 2004), especially since many trailers are available on the Internet. With the development of current digital technology the question arises if and to what extent it is feasible to automate the process of trailer production based solely on extracted movie features. Such a system could provide improvements in different movie-related fields. For example, it could (a) suggest innovative ways of video browsing in digital movie databases, (b) help developing and testing experiments to formalize existing movie editing methods (film theory), and (c) simplify or even extend the work of editors.

In this paper, we present our approach of an automatic trailer generation system that we implemented and tested for action movies, and which can also be extended for use with other genres. In section 2, a short overview of previous work related to automatic trailer generation is given. Section 3 describes the ontology-based formalism we developed and now use as a basis for our system. Section 4 illustrates our system, which is capable of analyzing a movie and generating an action movie trailer for it. In section 5 we discuss experimental results of our system. Finally, section 6 draws a conclusion and addresses possible aspects of future work.

2 Related Work

The specific field of generating movie trailers automatically has only little related work so far. The more general task of summarizing video content has been explored in detail. Works that come close to our aim are Chen et al (2004) and Lienhart et al (1997), where the possibility of generating a movie trailer is mentioned explicitly. Both claim to do the composition of footage according to rules derived from film theory and present ways to retrieve crucial information for trailer generation. They do not focus on how to compose trailers.

Other works within the field of video summarization rather focus on the task of pure summarizing in order to provide means to handle the increasing amount of video data. Three basic approaches have evolved. The first one is video skimming as in Christel et al (1999) and Smith/Kanade (1998), where video material is analyzed and condensed to important scenes. Typically the linearity of the input video is preserved here. The second one is summarizing contents in a pictorial way (Uchihashi et al 1999, Yeung/Yeo 1997). In Uchihashi et al (1999), salient single frames of video sequences are captured, sized according to their importance, and arranged in a linear comic-strip-like way of telling a story. The third video-browsing approach is closely related to the pictorial summarization but focuses on a hierarchical, not necessarily linear way of presenting the video content (Ponceleon/Dieberger 2001, Zhang et al 1993). The degree of automation varies. There are completely automatic works (Lienhart et al 1997, Smith/Kanade 1998, Uchihashi et al 1999) and semi-automatic works (Zhu et al 2003). Typically automatic summaries highly depend on low level analysis of image and audio, while the semi-automatic summary tools provide some manual annotation framework enabling high-level analysis to conclude what is happening in a scene. This approach even uses a hierarchy for video summarization quite similar to that defined by our trailer grammar; however, it does not discuss video summarization for the movie trailer format.

Another interesting work is Ma et al (2002), focusing on the question of how a video is perceived by a user.

3 Trailer Grammar

In order to successfully re-assemble movie footage in a short video which can be called a *trailer*, first of all the meaning of this label must be understood. According to Arijon (2000), films are created based on an underlying *film grammar* to successfully communicate with the audience. Kernan (2004) argues that a trailer is not only a video of a defined length consisting of a random assembly of shots and scenes, but also a movie genre in its own right. Therefore we assume that trailers, as a special kind of film, can be described by syntactic elements and semantic rules which constitute a *trailer grammar*. After giving a definition of the term trailer we will examine how this grammar can be modeled via syntactic and semantic elements.

3.1 The Definition of a Trailer

In order to define a trailer with respect to an automatic generation, the term *trailer* refers to the fact that these short movies were originally shown at the end of a film program in movie theaters (Kernan 2004). During the 20th century, trailers evolved from pure advertisement to a movie genre with its own unique conventions, based on the demand to combine an artistic form with the highly commercial need of attracting the biggest possible audience. Since movies and trailers exist in many different forms in different cultural environments, we focus our automatic approach on the Western culture's most dominant trailer and movie industry: Hollywood blockbuster cinema. Trailers from this domain have developed a general formula that pays as little attention to genre or specific target groups as possible (Kernan 2004). Our aim is to produce short videos that resemble rather conventional *theatrical trailers* by having a length of more than one minute and featuring footage from the original movie. These are opposed to so-called *teaser trailers* which are typically produced before primary shooting is finished. Teasers consist mostly of texts, voice-overs, graphic elements, and which have a maximum running length of one minute. In the following the term *trailer* therefore refers to a theatrical trailer for a contemporary Hollywood movie.

3.2 The Syntactic Elements of a Trailer

Shots and *transitions* are usually the basic elements of any edited movie. Within these elements we presume that certain types can be identified by a shot-by-shot analysis of original movie trailers. In order to determine these types, an appropriate set of descriptions, i.e. an appropriate vocabulary, has to be defined. These descriptions inevitably involve a trade-off regarding level of detail. We developed the following guidelines to clarify this issue. Shot types must: (a) be able to cover all shots of a trailer, (b) be clearly distinguishable from each other (no redundancy), (c) have a well-defined meaning, (d) apply to as many existing trailers as possible, and (e) be defined based on the movie features that can be extracted by our analysis tools (technical feasibility). In order to distinguish between the original movie/trailer shots and the shots we produce for our trailers we refer to the latter as *clips*. In order to fulfill the requirements listed above we define the types of *clips* by the following *properties:*

- a category (reflecting the shot's formal features),
- the playback speed (to model effects like slow-motion or acceleration),
- the volume of the original footage sound (so that clips can be muted or amplified),
- and location, corresponding to the footage location in the source movie.

3.3 The Semantic Elements

In order to assemble these syntactic elements in a trailer-like way, semantic rules are needed. We propose to represent these rules as a hierarchy of super- and sub-patterns as shown in Figure 1. Each super-pattern consists of a number of sub-patterns either in a specific or random order. The highest level of patterns is the *trailer pattern*. Since there is not only one universal pattern which can describe all trailers at once, this pattern can be used to distinguish between different types of trailers. A trailer pattern consists of at least one *phase pattern*. The phase patterns again are composed of *sequence patterns*, which in turn consist of a number of *clip/transition pairs*. These pairs are the lowest level of the hierarchy.

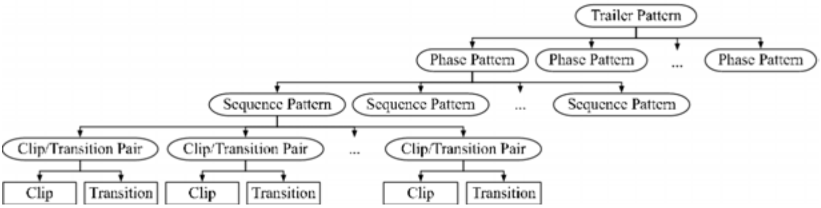

Figure 1. A branch of the hierarchical view of a generic trailer structure.

4 Trailer Generation System

Our system consists of two major components. The first one is a collection of various image and audio processing modules that provides a set of features extracted from the movie. The second component provides an implementation of the proposed trailer grammar. This component is able to categorize the annotated information of the first component and to use that data to automatically assemble a full trailer.

Most trailers try to summarize the plot and setting of the announced movie and to introduce the relations between the main characters. Presently, the automatic extraction as well as the generation of a narrative, or at least some kind of dramatic arc, seems hardly feasible. Therefore, our approach focuses on a genre which relies significantly more on visual sensation, speed and effects than on narrative: the action movie. We performed a shot-by-shot analysis of various action movie trailers from the last 15 years. Thereby we identified specific grammar elements and selected appropriate image and audio processing modules that are able to provide the corresponding footage for generating an action movie trailer.

4.1 Extracting and Annotating Movie Features

In order to extract features of a given action movie we not only use methods of image and audio analysis on different levels of abstraction, but also derive data from Internet resources. By combining the output of several modules with each other we enhance the value and reliability of the annotated data. In order to have a basis for evaluating the output we manually annotated the action movie *The Transporter (2002)*. In our system features from a movie are extracted by the following modules:

- shot boundary detection based on gray-level histogram changes
- motion-based segmentation which divides a movie into frame ranges with homogeneous optical flow intensities
- face detection along with a k-means clustering of the detected faces based on Principal Component Analysis
- text detection which detects frame ranges with disturbing overlaid text, e.g. credits, subtitles etc.
- a tool which extracts movie data such as title, director, actors, genre, famous quotes, awards won and production company from the Internet Movie Database (www.imdb.com)
- audio-based segmentation which divides a movie into frame ranges with homogeneous sound volume intensities
- detection of sudden volume change in a movie
- detection of frame ranges comprising speech using a phoneme-based speech recognition
- speech recognition which detects frame ranges containing famous quotes of the movie (given by the IMDb) by using each quote as one entity in the language model
- detection of frame ranges with music along with a level of disturbance, based on image-based spectral analysis
- detection of sound events (explosion, crash, gunshot, scream) based on spectral feature extraction and Support Vector Machine classification

4.2 A Framework for Generating Trailers of an Annotated Movie

The second component of our system is provided with the annotation from the automatic feature extraction. This annotation is used in combination with our semantic patterns in order to generate a trailer of the particular movie. In addition to original movie footage we include automatically produced animations and add music and sound effects from a separate audio archive. The process of generating trailers from annotated movies is split into the following sub-components (see also Figure 2):

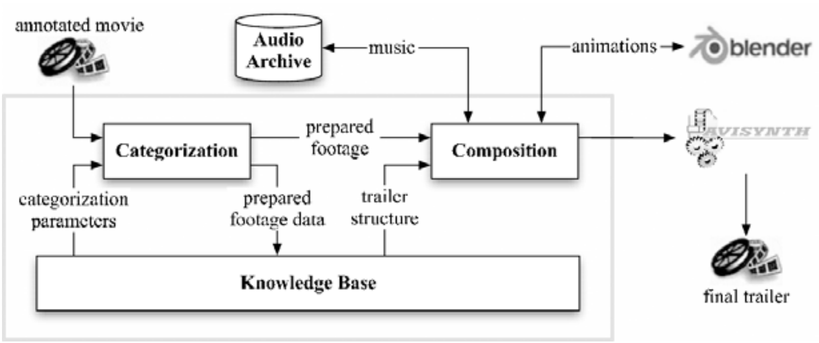

Figure 2: Diagram of the trailer generation process.

(1) In order to build a trailer we define a knowledge base that contains models for trailer structure elements and defines parameters for categories of video footage frame ranges. The movie annotation is filtered into the syntactic elements *clips* which then are classified into *categories* using these parameters. Next, the trailer model is created based on rules in the knowledge base and influenced by the availability of footage. In this way, the system generates a unique trailer model that is built to fit the available footage.

(2) The composition framework translates the established trailer structure into specific trailer elements: Apart from video footage we incorporate runtime-created text animations for movie title, credits etc., as well as pre-produced music and sound effects content from our own audio archive. Footage, text animations and audio are finally composed to a unique, fully automatically generated trailer based on trailer semantics.

Knowledge Base Functionality

In order to incorporate trailer semantics, we implement a knowledge base that is designed to hold the knowledge for trailer construction using the public domain software CLIPS (www.ghg.net/clips/CLIPS.html). The trailer grammar concepts are represented in an ontology. We cast specific knowledge about action trailer syntax and semantics. As instances of these classes, using has-a relations to model our hierarchical view of the trailer structure. The properties of clips (category, speed, volume, location) are implemented as slots and among these the category is implemented as a class of several slots again, specifying a list of video analysis attributes for its classification. The combination of several annotation attributes to a category leads to semantic higher-level knowledge about the footage.

Categorization

Given a set of category definitions the categorization module processes the annotation data in order to build clips for each category. We build the clips based on frame ranges as opposed to a shot-based approach, allowing the categorization process to be independent from scene/shot information. Let A be a set of video frames and A_{movie} the set containing all original movie frames, then the frame set of an analyzed movie feature $A_{feature x}$ (e.g. all frames showing a face) is a subset of A_{movie}. The first step of the categorization process is to filter out the desired frames by corresponding thresholds (e.g. only get big faces indicating a close-up shot). This results in a new set which we refer to as an *attribute frame set* $A_{attribute x}$ with $A_{attribute x} \subseteq A_{feature x} \subseteq A_{movie}$. We process these attribute frame sets as tracks and perform an intersection as illustrated in Figure 3. Furthermore, for each clip we calculate a probability value based on weighting factors assigned to the attributes.

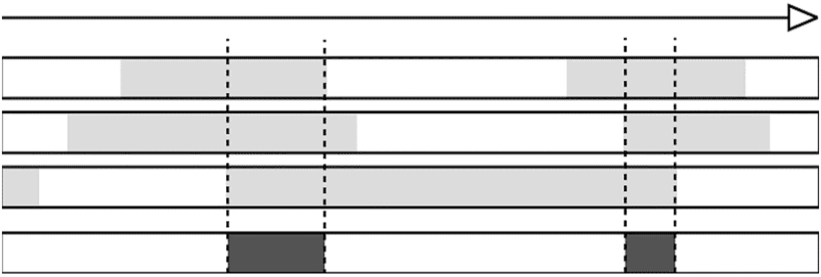

Figure 3: The intersection approach of the categorizer for a sample footage.

Trailer Structure

Once the movie footage has been segmented and categorized, information about the amount of clips within each category is stored in the knowledge base. The system then builds the trailer structure on an abstract level. In order to introduce variety into the trailer models, each semantic element in our hierarchy has a selection choice of lower level elements assigned to it. While offering multiple choices at each node in the trailer structure tree, the sequence of patterns can still be controlled to ensure consistency with the given trailer grammar. This approach grants easy and fast altering of the structure by linking more sub-patterns to a super-pattern or by deleting links. To avoid a purely random selection of a linked sub-pattern and to emphasize certain patterns, a weighting system is attached to the selection logic. Based on the trailer ontology and availability of categories, the knowledge base reasons which parts of the trailer structure fulfill all requirements. In case of certain parts failing due to lack of footage, fallback structures are considered first. If no such fallback exists, clip attributes are loosened: clips can then be chosen from random categories rather than specific ones. The result is a finished model of a trailer structure giving detailed information about which transitions to use, which background music to play, which clips of which category to show, what position within the movie they should come from, at what speed they should be played, and how high the volume of the original footage should be.

Selection of Clips

The clips in the trailer model come with properties regarding clip category, footage volume, speed, and location for footage selection. Given these parameters for the designated clips of the original movie, our system has three methods for clip selection: (1) Preferred location selection, based purely on the requested location and clip location in the movie, so the clip chosen is the one closest to the requested location. (2) Best clip from preferred location, which is similar to the preferred location selection, but additionally taking into consideration the quality of the clip so the clip chosen is the best clip available starting from the requested location. (3) A random clip of a given category is selected.

3D Text Animations and Audio

Text animations displaying information on movie title, release date, actor names, movie company as well as legal disclaimers are one distinctive feature of movie trailers and an essential component of a trailer structure. The composition system dynamically creates a script from which the 3D software Blender (blender.org) produces one digital video file per animation ready to be used for final composition, using predefined animation templates. For additional music soundtrack and sound effects, we provide the possibility to incorporate preproduced sound files.

Final Composition

Selected footage, animation clips and audio soundtrack are composed into a final video using Avisynth (www.avisynth.org) scripts. True to the methods used in standard movie trailers, sound effects are added to text animations to make them more effective. Changes in trailer soundtrack are masked by special transition sound effects. Fade/flash shot transitions (as determined by the trailer structure model) are implemented. The result is the finished trailer modeled according to a trailer structure created using our trailer ontology.

4.3 Automatically Generating Action Movie Trailers

We define 26 clip categories, such as *CharacterSpeaking*, *PersonSilent*, *Quote*, *FastAction*, *Explosion*, *Setting*, *Gunshot*, *Scream* etc., and 3 transition categories, namely *HardCut*, *FadeBlack*, *FlashWhite*. The definition of each category includes a set of attributes along with specific value ranges for the annotated features and weighting factors. An extension of our set of categories is possible and would be necessary to model more complex trailers or trailers of other movie genres.

As the basic structure of our action movie trailer pattern we identified five different phase patterns representing typical stages of action trailers. These phase patterns are again made up of sequence patterns corresponding to typical action trailer shot sequences. The phases are:

- *Intro* (slow and moody shots of locations and people together with speech establishing a conflict or introducing the main characters). Sequence patterns usually involve clips such as *Setting*, *CharacterSilent* or *CompanyName*.
- *Story* (medium fast shots of action and people together with dialogue to wrap up the task the main characters have to face). Sequence patterns use Intro clips but also add action with clips such as *SlowAction*, *Shout*, or *Fire*.
- *Break* (a long and very significant or dramatic comment by one of the main characters – typically without background music). Clips used in sequence patterns are *Quote* and *QuoteLong*.
- *Action* (a fast montage with loud sound of the fastest action scenes together with close-ups of the main characters). Sequence patterns use *ActorName*, *FastAction*, *Explosion*, *Gunshot* and clips of other similar categories.
- *Outro* (typically very calm or without any music and shows – sometimes mixed with close-ups or a short shot of one of the main characters uttering an extremely comic or tough comment – the title and credits of the movie together with a release date). Sequence patterns use clips such as *Quote*, *Title*, *Credits* or *Spectacular*.

With the defined elements (clips, transitions and patterns), as well as their relations to each other we can describe an action movie trailer in a formal way. This description can be used by our system to generate an action trailer from any movie (as long as the automatic analysis provides enough footage for the different categories).

In order to include text animations we provide four animation templates which all have a different artistic style, each matched to the action movie genre. Our audio archive is a collection of pre-produced sound files and consists currently of 37 music files and 22 sound effect files. Currently we use four categories of music files according to the mood of our trailer phases (Intro, Story, Action, Outro) and three sound effect types that are mostly used in professional trailers ("boom", "woosh" and "wooshbang").

5 Experimental Results

For several action movies, we let our system generate five test trailers per movie, each being unique due to the random choices built into the system architecture.[1] Many of the generated trailers give a quite good impression of the story and its characters. Although the framework makes no attempt to understand the storyline, important elements of the story are often contained in the selected shots. A good selection of quotes from the main characters seems to create an implicit storytelling, so the quotes have probably the greatest overall impact on the perceived quality of our trailers. We also tested a comedy, *Groundhog Day* (1993), and an old action movie, *James Bond: From Russia With Love* (1963). As expected, the results were unsatisfactory – for different reasons. The detection modules mostly failed for the James Bond movie due to the very different image and audio qualities, whereas for the comedy the action trailer pattern was clearly not fitting.

In addition to our own evaluation, we performed a test viewing with a group audience of 59 people. Since producing movie trailers is a creative process, a purely statistical analysis, like a shot-by-shot comparison of generated and official trailers, would not help much. Official trailers merely present the trailer artist's choice out of millions of other possible approaches. Using them as a ground truth would unnecessarily restrict the scope of possible best solutions.

We showed two trailers generated by our software, for *Transporter 2* and *Terminator 2*. As a reference, we showed two professionally produced trailers with different aesthetic appearances, for *War Of The Worlds* and *Miami Vice*, and three pseudo-trailers that each used a different level of randomness for the selection and ordering of shots or frame ranges, combined with music. One of these was generated by the commercial software Muvee Autoproducer

[1] Trailers produced by our system can be downloaded from www.tzi.de/svp.

(www.muvee.com). The test viewers were asked to rate the same six aspects for each trailer. The detailed scores of all trailers are shown in Table 1.

As expected, the overall rating of the random trailers is significantly lower than any of the others, while *War Of The Worlds* performed best. Our automatically generated trailers received high ratings for good composition and "cuts & effects", and lower ratings for "narrative aspects". The *Terminator 2* trailer received its highest rating (7.66) in the category "cuts & effects". This shows that our system succeeds in timing the cuts and adding animation screens with emphasizing audio effects to enhance the genre-typical powerful appearance. The categories "character introduction" and "plot introduction", which depend highly on sophisticated high-level analysis, received the lowest scores (6.36, 6.88). The results are still surprisingly good considering the fact that the system makes no attempt to analyze the movie's storyline, but instead tries to imitate that behavior by showing faces and inserting pieces of dialogues. As a consequence, the quotes in the generated *Terminator 2* trailer reveal some important clues about the characters and the plot (e.g. the Terminator's role as a protector from the future), but other important aspects, especially of the background story, are missing.

The results show that our attempts of automatic categorization and composition appear to be generally successful. They suggest that our automatic trailers are a clear improvement over random shot selection methods. Furthermore, it seems that wrongly chosen shots, resulting from inaccuracies of the analysis modules, typically do not disturb the flow of the trailer. Also, our results show no noticeable difference between the judgment of people who had seen the movies and of those who had not. For further evaluation, it may be interesting to perform viewer tests on trailer variations from only one movie to make the results more comparable between the different types of test trailers.

	scene selection	composition	cuts & effects	character introduction	plot introduction	advertisement value	total score
War of the Worlds, 2005 (official trailer)	8.41	7.91	7.79	7.47	7.40	8.16	7.86
Miami Vice, 2006 (official teaser)	4.97	6.27	6.27	3.27	3.59	4.95	4.89
The Transporter, 2002 (Muvee Autoprod.)	5.05	4.03	4.22	2.97	3.59	3.95	3.97
Bad Boys, 1995 (random frame ranges)	4.64	3.67	3.41	3.19	3.22	3.52	3.61
Blade, 1998 (random order of selected clips)	4.16	3.24	3.24	4.07	4.07	3.43	3.70
Transporter 2, 2005 (generated)	7.47	7.54	7.90	6.80	6.80	7.37	7.31
Terminator 2, 2001 (generated)	7.58	7.63	7.66	6.36	6.88	7.46	7.26

Table 1. Detailed average ratings from the user testing (max. score: 10).

6 Conclusion and Future Work

This paper presents a novel approach of intensively using a trailer grammar in combination with data automatically extracted out of a movie by different image and audio analysis techniques for generating a Hollywood-like action movie trailer. First, a trailer grammar was defined that can be applied to various movie genres. Second, a system was implemented, which provides means for using extracted features to build a trailer according to any defined trailer pattern based on our trailer grammar. One such trailer pattern was created by manually analyzing several action movie trailers. Using our system we generated trailers for some action movies according to this pattern, and these have shown that automatic trailer generation is not only possible, but can even achieve good results. Still, our trailers lack some elements a manually edited trailer comprises, e.g. telling a coherent story or voice-over narration.

Testing our system with further action movies should form the basis for refining current analysis modules, categorization and composition components, and our action trailer model. Also, many expansions are conceivable. The classification of movie footage into semantic categories could be expanded by adding more categories (e.g. *Kissing*, *Fight)* based on more sophisticated image and audio processing techniques. Concerning the composition framework, more animation styles as templates and a way of matching styles to movie content could be added. We also believe that the effect of a generated trailer could be vastly improved by adding pre-produced generic voice-overs to the soundtrack.

To enable our system to handle movie genres besides the action genre some significant expansions are necessary. First, further analysis modules need to be added and existing ones need to be improved to provide a richer annotation usable across movie genres. Second, trailer patterns for comedy, romance, drama, horror etc. need to be developed and incorporated into the generation knowledge base. Finally, the composition system would have to include animation and audio templates to render these models. A possible extension and a slightly complementary approach of our system is described by Kehl (2007). The author uses 14 dramatic trailer units and 33 narrative units in order to generate a trailer. Based on these units, the narrative content may be covered by this approach.

While in this work we have provided a basis for automatic trailer generation, a major aspect of future work remains the trade-off between technical feasibility and the semantic and aesthetic requirements the movie trailer must live up to.

Bibliography

Arijon, D. *Grammatik der Filmsprache*. Frankfurt a.M.: Zweitausendeins, 2000.

Christel, M. G., A. G. Hauptmann, A. Warmack, and S. A. Crosby. "Adjustable filmstrips and skims as abstractions for a digital video library." *ADL* (1999): 98-104.

Chen, H.-W., J.-H. Kuo, W.-T. Chu, and J.-L. Wu. "Action movies segmentation and summarization based on tempo analysis." *MIR '04: Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval.* New York, NY: ACM, 2004: 251–258.

Kehl, P. *Structures of Narrative: A Formal Description of Movie Trailers*. Master Thesis, University of Bremen, 2007.

Kernan, L. *Coming Attractions – Reading American Movie Trailers*. Austin, TX: University of Texas Press, 2004.

Lienhart, R., S. Pfeiffer, and W. Effelsberg. "Video abstracting." *Communications of the ACM*. 40.12 (1997): 54-62.

Ma, Y., L. Lu, H. Zhang, and M. Li. "A user attention model for video summarization." *Proceedings of ACM Multimedia* 2002: 533-542.

Ponceleon, D. and A. Dieberger. "Hierarchical brushing in a collection of video data." *Proceedings of the 34th Hawaii International Conference on System Sciences*. 2001: 1654-1661.

Smith, M. A. and T. Kanade. "Video skimming and characterization through the combination of image and language understanding." *IEEE International Workshop on Content-Based Access of Image and Video Database*. 1998: 61-70.

Uchihashi, S., J. Foote, A. Girhensohn, and J. Boreczky. "Video manga: Generating semantically meaningful video summaries." *Proceedings of ACM Multimedia*. 1999: 383-392.

Yeung, M. and B. Yeo. "Video visualization for compact presentation and fast browsing of pictorial content." *IEEE Trans CSVT*. 7 (August 1997): 771-785.

Zhu, X., J. Fan, A. K. Elmagarmid, and X. Wu. "Hierarchical video content description and summarization using unified semantic and visual similarity." *Multimedia Systems*, 9.1 (2003): 31-53.

Zhang, H. J., A. Kankanhalli, and S. W. Smoliar. "Automatic partitioning of full-motion video." *Multimedia Systems*, 1.1 (1993): 10-28.