Christian Petersohn

Temporal Video Segmentation

Beiträge aus der Informationstechnik

Christian Petersohn

Temporal Video Segmentation

VOGT

Dresden 2010

Bibliografische Information der Deutschen Bibliothek
Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.ddb.de abrufbar.

Bibliographic Information published by Die Deutsche Bibliothek
Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibiograpic data is available in the internet at http://dnb.ddb.de.

Zugl.: Berlin, Techn. Univ., Diss., 2010

Die vorliegende Arbeit stimmt mit dem Original der Dissertation „Temporal Video Segmentation“ von Christian Petersohn überein.

Gesetzt vom Autor

ISBN 978-3-938860-39-7

Jörg Vogt Verlag
Niederwaldstr. 36
01277 Dresden
Germany

Phone: +49-(0)351-31403921
Telefax: +49-(0)351-31403918
e-mail: info@vogtverlag.de
Internet : www.vogtverlag.de

Temporal Video Segmentation

vorgelegt von
Dipl.-Ing. Christian Petersohn
aus Berlin

Von der Fakultät IV - Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften
Dr.-Ing.

genehmigte Dissertation

	Promotionsausschuss
Vorsitzender	Prof. Dr.-Ing. Clemens Gühmann Technische Universität Berlin
Gutachter:	Prof. Dr.-Ing. Thomas Sikora Technische Universität Berlin
	Prof. Dr. rer. nat. Ebroul Izquierdo Queen Mary, University of London

Tag der wissenschaftlichen Aussprache: 1. März 2010

Berlin 2010
D83

Contents

Abstract **xi**

Deutsche Zusammenfassung **xiii**

Preface **xvii**

Acknowledgments **xix**

1 Introduction **1**
1.1 General 1
1.2 Aims 2
1.3 Significance and Contribution 3
1.4 Application scenarios 6

2 Pattern recognition **7**
2.1 Design of a pattern recognition system 7
2.2 Learning concepts 8
2.3 Features 9
2.4 Classification 9
2.4.1 Bayes classification 9
2.4.2 Support Vector Machines 16
2.4.3 Feature cascade classifier/decision trees 23
2.4.4 Multi-class classification 25
2.5 Evaluation measures 27
2.6 Training and testing strategies 30
2.6.1 Overfitting and generalization 30
2.6.2 Cross-validation and parameter selection 32

3 Related work **35**
3.1 Hierarchy of temporal video segments 35
3.2 Shot boundary detection 38
3.2.1 Shot definition and shot transition types 38
3.2.2 Metadata-based shot boundary detection approach 39

3.2.3 Content-based shot boundary detection approaches 39
3.2.4 Drawbacks and limitations of current approaches 47
3.3 Representative images and segments inside of shots 51
3.3.1 Video abstraction types 51
3.3.2 Static storyboards 51
3.3.3 Segments inside of shots 59
3.3.4 Drawbacks and limitations of current approaches 61
3.4 Logical unit detection 63
3.4.1 Terms and definitions 63
3.4.2 Existing approaches to logical unit detection 65
3.4.3 Drawbacks and limitations of current approaches 80

4 Shot boundary detection 81
4.1 Optimal statistical shot boundary detection 82
4.2 Training and test data, performance evaluation 85
4.3 Overview of shot boundary detection system 89
4.4 Hard cut detection 92
4.4.1 Model of a hard cut 92
4.4.2 Features 93
4.4.3 Adaptive threshold 97
4.4.4 Feature fusion 100
4.4.5 Flash detector, short-term outlier frames 102
4.4.6 Training and experimental results 105
4.5 Fade detection 106
4.5.1 Model of a fade 106
4.5.2 Features and detection 107
4.5.3 Experimental results 108
4.6 Dissolve detection 110
4.6.1 Model of a dissolve 110
4.6.2 Detection overview 112
4.6.3 Candidate selection 112
4.6.4 Position determination 115
4.6.5 Feature extraction for candidate verification 117
4.6.6 Bayes, SVM and feature cascade for classification 125
4.6.7 Short dissolves 132
4.6.8 Experimental results 136
4.7 Wipe detection 139
4.7.1 Model of a wipe 139
4.7.2 Difference images 140
4.7.3 Evenness factor 142
4.7.4 Position determination 143
4.7.5 Double Hough transform 143

4.7.6 Detection overview and feature fusion . . . 148
4.7.7 Limitation of the wipe detection approach . . . 149
4.7.8 Experimental results . . . 149
4.8 Fusion of segmentation results . . . 151
4.9 Experimental results . . . 151
4.9.1 Transition type specific results . . . 151
4.9.2 TRECVID participation . . . 152
4.9.3 Comparison with TRECVID shot detection approaches . . . 155
4.9.4 Error analysis . . . 157
4.10 Summary and conclusion . . . 160

5 Sub-shot detection 163
5.1 Motivation and problem definition . . . 163
5.2 Algorithms for sub-shot detection . . . 165
5.2.1 Simple thresholding . . . 166
5.2.2 Clustering . . . 167
5.2.3 Motion analysis . . . 172
5.3 Evaluation and test set . . . 176
5.4 Test results . . . 181
5.5 Missed shot boundaries . . . 185
5.6 Selection of key-frames, classification . . . 187
5.7 Summary and conclusion . . . 188

6 Scene detection 189
6.1 Composition of scenes, challenges in detection . . . 190
6.1.1 Semantics . . . 190
6.1.2 Film grammar in the composition of scenes . . . 190
6.1.3 Definitions and guidelines to limit subjectivity . . . 193
6.1.4 Training and test set . . . 196
6.2 Shot similarity . . . 197
6.2.1 Choosing key-frames for similarity analysis . . . 197
6.2.2 Low-level features . . . 199
6.3 Gradual shot transitions as film grammar cues . . . 204
6.3.1 Gradual shot transitions in the data set . . . 204
6.3.2 Using gradual transitions to split scenes . . . 206
6.3.3 Using gradual transitions to suppress scene boundaries . . . 209
6.3.4 Reliability of gradual shot transitions as film grammar cues . 213
6.3.5 Experimental results . . . 215
6.4 Segmentation Methods . . . 219
6.4.1 Scene detection performance . . . 219
6.4.2 Remaining errors . . . 225
6.5 Multi-modal scene detection framework . . . 226

6.6 Summary and conclusion 231

7 Summary, conclusion, outlook **233**

A Histogram comparison **239**
A.1 State-of-the-art 239
A.2 Low-complexity proximity-based histogram dissimilarity measure . . 244

B Example of visual table-of-contents **251**

Bibliography **255**

Notation conventions **267**

Abstract

The presence and availability of video and multimedia data has steadily grown over the past years. Enabled by advances in storage and transmission capabilities, the huge amount of media content now triggers the need for technologies for video and multimedia content management. Simple and effective access to the content is needed. The objective of this thesis is to present steps toward simple and effective video access and browsing, to work towards technologies that can simplify annotation, automatic analysis, or video editing. This is done by developing methods and algorithms for the extraction of structural units in video on different hierarchical levels, cf. figure 1.

The first problem examined is the extraction of video shots. This is a fundamental task because shots are an important structural unit in video and most algorithms and techniques for further structuring, analysis, search and retrieval build upon the knowledge of shot boundaries. Extracting video shots is equivalent to detecting the shot boundaries in a video. The characteristics of the four different types of shot transitions are investigated, i.e. cut, fade, dissolve, and wipe, and a system of novel algorithms is presented with each algorithm specifically tailored to detect one of the shot transition types. The algorithms are designed to offer high detection rates with low computational complexity. They proved their performance in the TRECVID shot boundary detection task. The current version had the best overall detection results of the 18 shot detection systems that were evaluated on the official test set. The system is approximately twenty times faster than real time. It was one of the fastest in the contest.

While visually simple shots with little variance in content may well be regarded as basic units of video, there also exist visually complex shots with significant object or camera motion and a large variance in visual content. Such visually complex shots cannot sufficiently be represented by a single key-frame. A richer and adaptive representation is needed. This second problem is investigated and as a result a new level in the hierarchy of temporal video segments, named sub-shots, is proposed. Sub-shots are parts of shots. They are limited to small variations in semantic and visual content and are therefore suited as basic units for search and retrieval and for key-frame extraction. Three different algorithms for the automatic extraction of sub-shots are presented and evaluated. They are based on

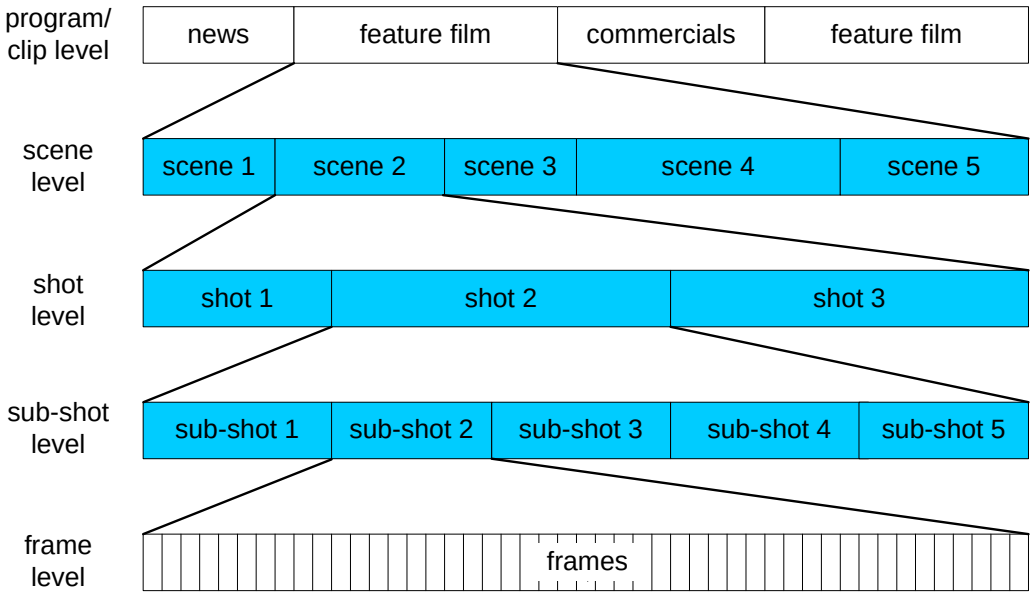

Figure 1: Proposed hierarchy of temporal video segments. This work is about segmenting video into temporal units on the scene, shot, and newly proposed sub-shot level.

visual content and motion analysis. The approaches to sub-shot detection based on clustering and motion analysis outperformed the simple thresholding algorithm.

The third problem addressed is the extraction of scenes. While shots and sub-shots are low-level units of video, humans will perceive the story or narrative of a video in terms of higher semantic units. Talking about the content of a video is usually based on entire scenes, not on single shots. Various types of known approaches to scene detection are investigated. A set of low-level visual features is evaluated based on their suitability for scene detection. Additionally, important aspects of film grammar regarding the composition of scenes are detailed. One important point regarding film grammar is that the types of shot transitions used by film editors in video are not randomly chosen. Cuts, fades, dissolves, and wipes are devices used to structure video and to provide local hints for the global structuring. An approach is presented suggesting improvements to known scene detection algorithms in two ways: First, to appropriately choose representative frames used for scene detection based on the position of detected gradual shot transitions; and second, to interpret gradual shot transitions as film grammar cues that have a separating or merging effect upon shots in their temporal proximity. A discussion is presented indicating how different thresholding mechanisms influence scene detection quality and experimental results are presented comparing different segmentation algorithms. As a last point approaches to multi-modal scene detection are discussed and a framework is presented that could be used for further research on this topic.

Deutsche Zusammenfassung

Zeitliche Videosegmentierung

Die Verfügbarkeit von Video und multimedialen Daten ist in den vergangenen Jahren immer weiter gestiegen. Durch Fortschritte bei der Entwicklung von Speicher- und Übertragungsmöglichkeiten existiert eine solche Menge an Mediendaten, dass auch die Technologien zu ihrer Verwaltung immer wichtiger werden. Ein einfacher und effektiver Zugriff auf die Inhalte ist notwendig. Das Ziel dieser Dissertation ist es, Schritte in Richtung eines einfachen und effektiven Zugriffs auf Videoinhalte aufzuzeigen. Außerdem werden Technologien entwickelt, die Vereinfachungen bei Annotation, automatischer Analyse oder auch beim Videoschnitt ermöglichen. Dies geschieht durch die Entwicklung von Methoden und Algorithmen zur automatischen Extraktion von zeitlichen Einheiten auf unterschiedlichen hierarchischen Ebenen in einem Video, siehe Figur 2.

Das erste in dieser Arbeit behandelte Themenfeld ist die Extraktion von Videoshots. Dies ist eine grundlegende und wichtige Aufgabe, da Shots die Basiseinheiten in einem Video sind und die meisten Algorithmen und Methoden zur Strukturerkennung, Analyse und Suche in Videos auf Shotinformationen aufbauen. Extraktion von Videoshots bedeutet, die Übergänge zwischen Shots zu finden. Die Charakteristika von vier verschiedenen Shotübergangstypen werden untersucht. Das sind harter Schnitt, Ein-/Ausblendung, Überblendung und Wischblende. Es wird ein System neuer Algorithmen präsentiert mit jeweils einem spezialisierten Algorithmus für jeden Shotübergangstyp. Die Algorithmen sind auf hohe Erkennungsqualität bei gleichzeitig niedriger Rechenkomplexität ausgelegt. Sie haben ihre Leistungfähigkeit beim internationalen TRECVID-Wettbewerb für Shoterkennungssysteme bewiesen. Die aktuelle Version erreichte, bezogen auf alle Shotübergänge, die beste Erkennungsleistung auf dem offiziellen Testset. Gleichzeitig ist das Verfahren etwa zwanzigmal schneller als Echtzeit und damit eines der schnellsten im Feld.

Während visuell einfache Shots, also Shots mit nur kleinen Änderungen des Bildinhalts, Basiseinheiten in einem Video sind, gibt es auch visuell komplexe Shots mit umfangreicher Objekt- oder Kamerabewegung und starker Änderung des Bildinhalts. Diese visuell komplexen Shots können nicht hinreichend durch ein einzelnes Keyframe repräsentiert werden. Eine umfassendere und adaptive Repräsentation wird benötigt. Dies ist das zweite behandelte Themenfeld. Als Lösung

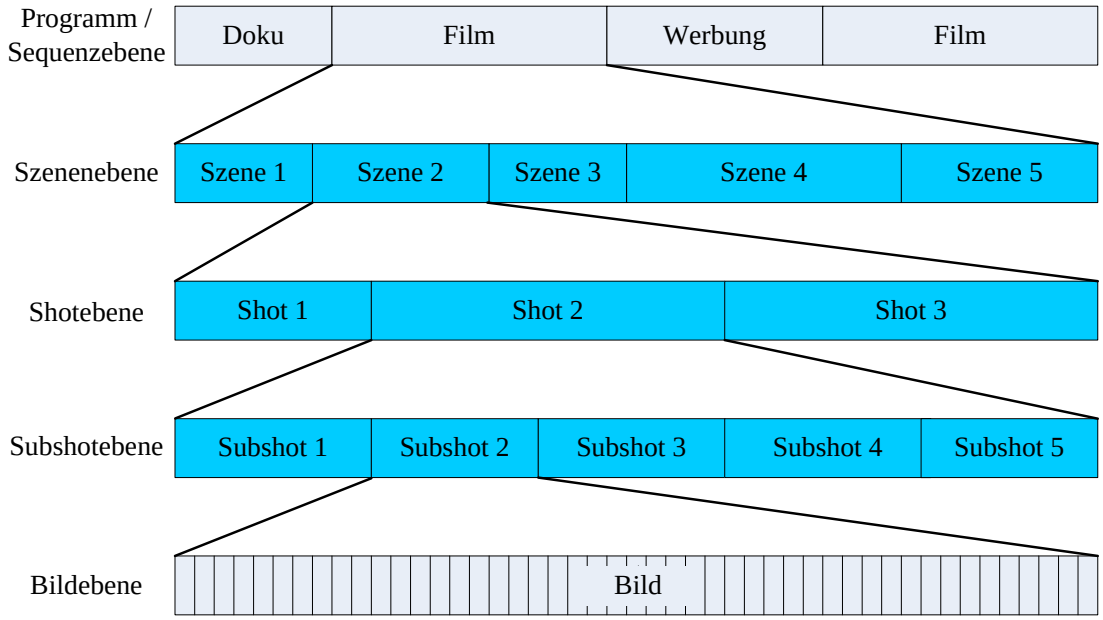

Abbildung 2: Vorgeschlagene Hierarchie zeitlicher Videosegmente. In dieser Arbeit wird die Segmentierung von Video in die zeitlichen Einheiten auf der Szenen-, Shot- und neu vorgeschlagenen Subshotebene behandelt.

wird in dieser Arbeit die Segmentierung von Shots auf einer neuen Hierarchiestufe der zeitlichen Videosegmente, in sogenannte Subshots, vorgeschlagen. Subshots sind Teile von Shots. Sie sind begrenzt auf nur kleine Änderungen im semantischen und visuellen Inhalt und sind daher geeignet, als Basiseinheiten bei der Vidoesuche oder für die Keyframeextraktion verwendet zu werden. Drei verschiedene Algorithmen für die automatische Extraktion von Subshots werden präsentiert und evaluiert. Sie basieren auf der Analyse von visuellem Inhalt bzw. Bewegung. Der Clusteralgorithmus und der Bewegungsanalysealgorithmus zur Subshoterkennung liefern dabei bessere Ergebnisse als der Schwellwertalgorithmus.

Das dritte behandelte Themenfeld ist die Erkennung von Szenen in einem Video. Während Shots und Subshots kleine einfache Videoeinheiten sind, basiert die menschliche Wahrnehmung des Videoinhalts eher auf größeren semantischen Einheiten. Wird beispielsweise über einen Film erzählt, so erfolgt das normalerweise mit Hilfe von Szenen. Verschiedene Typen bekannter Szenenerkennungsverfahren werden analysiert. Mehrere Videomerkmale werden bezüglich ihrer Tauglichkeit für die Szenenerkennung evaluiert. Zusätzlich werden wichtige Aspekte der Filmgrammatik bei der Komposition von Szenen beschrieben. Ein wichtiger Punkt, bezogen auf Filmgrammatik, ist, dass der Typ eines Shotübergangs beim Filmschnitt nicht willkürlich gewählt wird. Harte Schnitte, Ein-, Aus-, Über- und Wischblenden sind Bausteine für die Strukturierung eines Videos und liefern lokal Hinweise auf die globale Szenenstruktur. Es wird ein Verfahren entwickelt und präsentiert, das Verbesserungen zu bekannten Szenenerkennungsverfahren auf zweierlei Weise erreicht:

Erstens werden geeignete Bilder für die Analyse in Szenenerkennungsverfahren unter Beachtung der Lage der graduellen Shotübergänge ausgewählt. Zweitens werden graduelle Shotübergänge als filmgrammatische Hinweise interpretiert, die sowohl trennende als auch vereinende Wirkung auf die zeitlich benachbarten Shots haben können. Es wird untersucht, wie verschiedene Schwellwertverfahren die Qualität der Szenenerkennung beeinflussen. Messergebnisse für den Vergleich mehrerer Segmentierungsverfahren werden präsentiert. Schließlich werden noch Ansätze zur multimodalen Szenenerkennung diskutiert und ein entsprechendes Rahmenwerk vorgestellt, das für weitere Untersuchungen in diesem Themenfeld genutzt werden kann.

Preface

The original work presented in this thesis has been published before. The work dealing with the extraction of shots is published in:

1. Christian Petersohn. Fraunhofer HHI at TRECVID 2004: Shot boundary detection system. In TREC Video Retrieval Evaluation Online Proceedings, TRECVID, 2004.

2. Christian Petersohn. Dissolve shot boundary determination. In Proc. IEE European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology, EWIMT, pages 87-94, London, UK, 2004.

3. Christian Petersohn. Wipe shot boundary determination. In Proc. IS&T/SPIE Electronic Imaging 2005, Storage and Retrieval Methods and Applications for Multimedia, pages 337-346, San Jose, CA, USA, 2005.

4. Christian Petersohn. Verfahren und Vorrichtung zum automatisierten Vergleichen zweier Sätze von Messwerten. Patent granted. DE 102007051612.8-52 2007.

The work on the extraction of sub-shots is published in:

5. Christian Petersohn. Sub-shots - basic units of video. In Proc. IEEE International Conference on Systems, Signals and Image Processing and EURASIP Conference Focused on Speech and Image Processing, Multimedia Communications and Services, Maribor, Slovenia, 2007.

6. Christian Petersohn. Verfahren zur zeitlichen Segmentierung eines Videos in Videobildfolgen und zur Auswahl von Keyframes für das Auffinden von Bildinhalten unter Einbeziehung einer Subshot-Detektion. Patent filed. 2007.

The work on the extraction of scenes is published in:

7. Christian Petersohn. Logical unit and scene detection: a comparative survey. Invited paper. In Proc. IS&T/SPIE Electronic Imaging 2008, Multimedia Content Access: Algorithms and Systems II, San Jose, CA, USA, 2008.

8. Christian Petersohn. Improving scene detection by using gradual shot transitions as cues from film grammar. In Proc. IS&T/SPIE Electronic Imaging 2008, Multimedia Content Access: Algorithms and Systems II, San Jose, CA, USA, 2008.

9. Christian Petersohn. Automatisiertes Verfahren zur zeitlichen Segmentierung eines Videos in Szenen unter Berücksichtigung verschiedener Typen von filmgrammatikbasierten Übergängen zwischen Bildfolgen. Patent filed. 2007.

The shot boundary detector has been used to automatically segment the TRECVID video corpus since 2005. This master shot boundary reference, together with the key frames, is publicly available for the TRECVID 2005 data.

10. Georges Quenot, Christian Petersohn, Kevin Walker. TRECVID 2005 Keyframes & Transcripts. Linguistic Data Consortium, Philadelphia, 2007.

A short summary of the temporal video segmentation work presented in this thesis has been published in

11. Christian Petersohn. Temporal video structuring for preservation and annotation of video content. In Proc. IEEE International Conference on Image Processing ICIP 2009, Cairo, Egypt, 2009.

Acknowledgments

Conducting this thesis is like walking on a long and winding road. This challenging but richly rewarding journey is made possible by the help, directly and indirectly, from people "around" me.

First, I would like to thank my supervisor, Professor Thomas Sikora, for his guidance and encouragement. In the discussions he initiated important improvements to this work.

I would also like to thank Prof. Ebroul Izquierdo for being willing to be my co-supervisor.

Professor Hans-Joachim Grallert, Dr. Siegmund Pastoor and Dr. Ulrich Leiner supported me at the Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute and gave me the possibility to pursue my research.

Special thanks goes to my college Thomas Meiers who has been a constant source of encouragement and support. Especially mathematical discussions have been very fruitful with him.

I would like to thank my colleges Marco Kunze and Christian Hentschel for their support and humor. Sharing the room with them has been very enjoyable.

Sina Scheuplein and Jens Binder supported me towards the end of this work. Having lunch together has been very pleasant.

Lebop, thank you for all the discussions and comments. Being able to join your family was such a marvelous and lovely experience. You have a huge heart. I have learned so much from you.

Sally and Julia, thank you for being photo models. It is so much fun to have you. I am glad you are around.

Doris, it is not always easy to live with someone who has PhD thesis hanging over his back. You have always supported me and kept me focused on the important things in life. Thanks for your love, you are my precious gift.

Most importantly, I would like to thank my parents and family for their enormous support. You have always believed in me. Thank you for your love, encouragement and backup.

Chapter 1

Introduction

1.1 General

Our society is increasingly dependent upon the ability to manage the information tide. The ever growing digital storage capacities, improved compression techniques and higher transmission rates afford us the luxury of rapidly storing and transmitting multiple types of data, such as text, audio, images, and video to and from huge archives. But storing and transmitting data makes sense only if useful information from these archives can be browsed, searched and retrieved.

For text this is comparatively simple. Text is organized in a natural hierarchy of characters, words, sentences, paragraphs and chapters, etc. Books have a table of contents that is used by the reader to quickly identify sections of interest. Text can be searched using algorithms like full text search or approximate string search. Simple indices like inverted files exist based on the alphabetical order of letters. The alphabet is also useful in the navigation of a lexicon. Dictionaries and thesauruses can assist in broadening user queries.

Compared to text, video is inherently more difficult to manage. Viewers are not provided with a structured table of video contents, per se. Historically, navigation in video has been very limited and often restricted to simple backward and forward operations. Video does not have such an intrinsic order for frames as the alphabet provides for letters. Many different descriptions, like color and edge histograms or face descriptors focusing on various features on different semantic levels and even different modalities (video, audio, text), have been proposed. Video indexing and search is much more difficult since low-level descriptors usually lack a one dimensional ordering and the harder-to-build high-level concept detectors only cover the aspects for which they were built.

An important prerequisite for video visualization, abstraction, navigation, search and retrieval is the automatic discovery of temporal video structure. Only after temporal video segmentation is performed can the resulting units be marked and used in indexing, search and retrieval. Through hierarchical structuring it is

possible for users to browse video similar to the way they browse text in books or magazines.

The first step in temporal video segmentation generally is the determination of video shot boundaries. Video shots are defined as a sequence of consecutive frames taken by a single camera act. Shots are often considered the primitives for higher-level content analysis, indexing, and classification [Hanjalic, 2002].

Temporally adjacent shots are then grouped to form logical units which are also called scenes, logical story units (LSUs) or simply story units [Kender and Yeo, 1998, Hanjalic et al., 1999, Vendrig and Worring, 2002]. Scenes are defined as events that take place in one setting in a continuous time period [Ascher and Pincus, 1999]. They convey the semantic meaning of the video to the viewers [Rui et al., 1999].

Shots are physical units in video that can be detected using low-level features and transition models. Scenes, on the other hand, are defined on a higher, semantic level. Sub-shots are extracted for shots that contain complex visual material.

In the design of video analysis and indexing applications three main questions have to be answered [Snoek and Worring, 2005]: 1) What (temporal) units are to be indexed, e.g. the entire video document or single frames? 2) How are they to be indexed, i.e. what type of analysis is to be performed and on which modalities? And 3) which index is to be used with which labels, e.g. the names of the players in a soccer match, their time dependent position, or both?

This work is about video segmentation on a *temporal* axis which is substantially different from video segmentation on a *spatial* axis. *Spatial* video segmentation tries to find spatial areas within frames that belong together, e.g. to find and extract objects within a video frame, whereas *temporal* video segmentation tries to find the temporal units in video.

1.2 Aims

The main aim of this thesis is to identify and extract the temporal units in video on various hierarchical levels. Extraction is done on three different levels.

1. *Scenes* - Scenes are high-level semantic structural units in video that allow an abstract representation. In particular, this work focuses on different scene detection algorithms, features and shot representations employed, and how film grammar can be used to improve the performance.

2. *Shots* - Shots are regarded as the basic units in video. Different transition types - hard cuts, fades, dissolves, and wipes - have to be detected.

3. *Sub-Shots* - Visually complex shots cannot be comprehensively represented by a single key-frame and may themselves contain a temporal structure above

the frame level. Sub-shots are identified as structural units below the shot level. Different algorithms for sub-shot detection are proposed, compared and evaluated.

After the temporal units in video have been extracted, they can be represented by key-frames and a visual table of contents can be constructed, cf. figure 1.2, which is a much richer representation than the state-of-the-art in current video portals, cf. fig. 1.1.

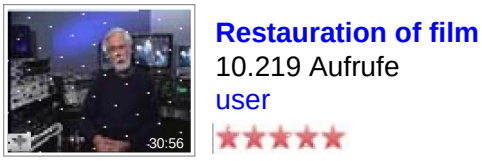

Figure 1.1: State-of-the-art example of the presentation of a video in a video portal, e.g. YouTube. One key-frame and a little additional text, user rating, etc., is used for the entire video sequence.

1.3 Significance and Contribution

Shots

The first task in recovering the temporal structure of video is to segment it into its shots. This is equivalent to finding all shot transitions in a video. In this work novel approaches to hard cut, fade, dissolve, and wipe detection are proposed. Separate detectors are developed for the different types of shot transitions. All detectors are developed to have low computational complexity. The shot detector also works for black-and-white video as only luminance information is needed. Because specific detectors are used for different shot transition types, the system does not only detect the positions of shot boundaries but also the type and duration of the transitions. The system was evaluated in the TRECVID contest. It outperformed the algorithms of all other participating teams in the hard cut detection task and was among the best for gradual transition detection (fades, dissolves, wipes).

The hard cut detector uses a novel combination of pixel, histogram, and edge differences with an adaptive thresholding scheme. Flash detection additionally improves detection results. Dissolve and fade detection is done with edge energy statistics, average frame luminance, pixel and histogram differences, and novel linearity and evenness measures. Gradual transition candidates are checked by analyzing frame differences of frame pairs with various temporal distances to enable the detection of gradual transitions independent of their duration. A dedicated one-frame-dissolve detector is proposed that allows dissolve detection without lower bound on their duration. For wipe detection an evenness factor is defined and

Figure 1.2: Proposed visual table of contents with a key-frame for the entire video sequence (top left), key-frames for each scene (vertical left), shot key-frames for each scene (horizontal), and sub-shots (bottom line).

proposed to be used to exploit the observation that during a wipe spatial zones of change move through the image. For the remaining wipe candidates usage of a new approach for the detection of uniform movement of linear zones of change, using double Hough transform, is proposed.

Due to its low computational complexity and the excellent performance of the shot detection system, the algorithms developed in this thesis were used to provide the official TRECVID master shot reference for the different higher-level tasks in TRECVID 2005, 2006, 2007, 2008 and 2009.

Sub-shots

Shots are commonly regarded as the basic unit of video above the frame level. However, shots may contain very complex and diverse content. A key-frame representation with only one or two key-frames may not suffice to visually summarize the content of a shot. For search, retrieval, and video analysis, smaller consistent units are sometimes needed. To address these issues a new level in the hierarchy of temporal video segments, named sub-shots, is proposed. Three different algorithms for the automatic segmentation of shots into sub-shots, based on visual content and motion analysis, are proposed, compared and experimentally evaluated. Best results can be achieved using a frame clustering approach or the motion analysis algorithm.

Scenes

Scenes are semantic structural units in video. Manual segmentation of a video into scenes is a subjective task. Based on common editing patterns and conventions used in the film-making craft, rules and guidelines are presented to limit subjectivity and to provide a clearer understanding of the scene definition.

Focus is then directed toward the investigation of visual features for scene detection on one hand and on thresholding and segmentation methods on the other. Various visual low-level features may be employed for scene detection. Feature quality is investigated for a set of histogram and block-based features with different color spaces, number of color channels, and key-frame selection methods. The measurements offer a profound evaluation and comparison of the different features.

After detailing important aspects of film grammar regarding logical units and scenes, clearly the types of shot transitions used by film editors in video are not randomly chosen. Cuts, fades, dissolves and wipes are devices in film grammar used to structure video. This work illustrates how knowledge of film grammar can be used to improve scene boundary detection algorithms. Three improvements regarding the use of gradual shot transitions as features in the scene detection process are proposed. (1) The selection of key-frames for shot similarity measurement should take the position of gradual shot transitions into account. (2) Gradual shot transitions have a separating effect. Local cues can be used to improve the

global structuring into scenes. (3) Gradual shot transitions also have a merging effect upon shots in their temporal proximity. Coherence values and shot similarity values used during scene detection have to be modified to exploit this fact. The proposed improvements can be used together with a variety of scene detection approaches. Experimental results indicate that considerable improvements in terms of precision and recall are achieved.

Different thresholding and segmentation methods are investigated and experimentally evaluated. A new globally adaptive thresholding method with multiple coordinated thresholds and window sizes to detect unequally pronounced scene boundaries is proposed, which outperforms other well-known and frequently used scene detection approaches.

Finally, a discussion on multi-modal scene detection is presented and a framework is proposed that can combine features from different modalities into one similarity or coherence measure for scene detection. This framework can serve as a basis for further research.

1.4 Application scenarios

The goal of automatic temporal structure detection in video is to extract the units needed for further processing and subsequent applications. Possible application scenarios that benefit from or require the extraction of temporal video units include:

1. *Video Browsing* - A scalable and browsable visual representation should enable the user to efficiently judge the visual content without having to completely watch it.

2. *Navigation and Access* - Non-linear narrative based navigation and access to specific parts of the video should be possible. For example, a certain news story or part of a documentation should be easily accessible without having to rely on tedious fast forward or fast backward operations only.

3. *Video Editing* - Temporal units enable the user to rearrange video content or to easily identify and reuse or discard parts of a video, such as commercial breaks.

4. *Video Annotation, Analysis, and Search* - The extraction of temporal units is a prerequisite in order to have units other than an entire video to annotate, analyze, and use in search scenarios.