

**Multimedia Content Analysis, Indexing and Summarization:  
A Perspective on Real-Life Use Cases**



**Multimedia Content Analysis, Indexing and Summarization:  
A Perspective on Real-Life Use Cases**

**Proefschrift**

ter verkrijging van de graad van doctor  
aan de Technische Universiteit Delft,  
op gezag van de Rector Magnificus prof. ir. K.C.A.M. Luyben,  
voorzitter van het College voor Promoties,  
in het openbaar te verdedigen op maandag 11 januari 2010 om 12:30 uur  
door

Suphi Umut NACI

Master of Science in Electrical and Electronic Engineering, Boğaziçi University, Turkije  
geboren te Istanbul, Turkije.

Dit proefschrift is goedgekeurd door de promotor:  
Prof. dr. ir. J. Biemond

Copromotor:  
Dr. A. Hanjalić

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof. dr. ir. Jan Biemond,	Technische Universiteit Delft, promotor
Dr. Alan Hanjalić,	Technische Universiteit Delft, copromotor
Prof. dr. Ingrid Heynderickx,	Technische Universiteit Delft
Prof. dr. ir. Geert-Jan Houben,	Technische Universiteit Delft
Prof. dr. Bülent Sankur,	Boğaziçi Üniversitesi, Turkey
Prof. dr. Jenny Benois-Pineau,	Université Bordeaux 1, France
Dr. Mauro Barbieri,	Philips Research Europe

The research for this thesis was conducted in the scope of the BSIK MultimediaN research program.

Multimedia Content Analysis, Indexing and Summarization: A Perspective on Real-Life Use Cases  
Naci, Suphi Umut  
Thesis, Delft University of Technology – With Ref. – With Summary in Dutch  
Published by TU Delft Mediamatica  
ISBN 978-90-813811-7-8

Copyright © 2009 by S.U. Naci

All rights reserved. No part of this thesis may be reproduced or transmitted in any form or by any means, electronic, mechanical, photocopying, any information storage and retrieval system, or otherwise, without written permission from the copyright owner.

*To my dear son, Arda...*

*Biricik ođlum Arda'ya...*



## Summary

The problem of finding images, video clips and music, given time, place, interest and mood has kept an immense number of scientists and technology developers busy in the past twenty years. However, straightforward attempts to apply text-based search to non-textual data still seem to be the only viable solution. In spite of the numerous ideas proposed so far in the MIR (Multimedia Information Retrieval) research field, it is remarkable that hardly any significant success story, and in particular a commercially relevant one, has been reported.

This thesis addresses the reasons that have prevented broad practical deployment of theories and algorithms for searching and retrieving content in multimedia data collections and proposes novel, generic and robust solutions. In particular, the thesis focuses on the problems that typically emerge when dealing with realistic use cases built around real-life systems, noisy data and highly unstructured and diverse content. A number of MIR aspects are selected that cover different MIR challenges, namely shot boundary detection, indexing videos of live music content and video summarization. All the algorithms proposed handle complex content and provide generic applicability and real-time operability. In shot boundary detection, spatiotemporal properties of the signal are used. Video indexing provides a new feature set based on crossing rate properties. Finally, in video summarization the audio modality is largely employed, emphasizing the importance of selecting the right contributions from different modalities.



## Samenvatting

De uitdaging om beelden, video clips en muziek op basis van tijd, plaats, belangstelling en stemming te vinden (te bepalen) heeft de afgelopen twintig jaar een groot aantal wetenschappers en technologieontwikkelaars beziggehouden. Niettemin, het rechttoe rechtaan toepassen van tekstgebaseerde zoekmethoden op niet-tekstgebaseerde data lijkt nog steeds de enige begaanbare weg. Ondanks talrijke ideeën die tot nu toe werden voorgesteld op het gebied van MIR (Multimedia Information Retrieval), is het opmerkelijk dat daar bijna geen noemenswaardig succesvol resultaat, vooral een met een commercieel belang, uit is voortgekomen.

In dit proefschrift worden de redenen behandeld die het praktisch gebruik van theorieën en algoritmen om naar informatie binnen grote multimedia databasen te zoeken hebben belemmerd en worden nieuwe, generieke, en robuuste oplossingen voorgesteld. In het bijzonder is aandacht besteed aan de problemen die zich voordoen bij het behandelen van realistische scenario's rond 'real-life' systemen, bij ruisachtige data en zeer ongestructureerde en diverse 'content'. Er zijn verscheidene MIR aspecten gekozen die leiden tot diverse uitdagingen op het gebied van de Multimedia Information Retrieval, namelijk, de detectie van video shots, het construeren van video-indexen van live muziekconcerten en de samenvatting van videodata. Alle voorgestelde algoritmen bewerken complexe content, zijn algemeen toepasbaar en werken in real-time. Bij 'shot boundary'-detectie worden de spatio-temporale eigenschappen van het signaal gebruikt. Video indexing biedt een nieuwe set van op nuldoorgangen gebaseerde kenmerken. Tenslotte wordt bij het samenvatten van video vooral de audio modaliteit gebruikt, daarbij de nadruk leggend op het belang van de keuze van de juiste bijdragen uit de verschillende modaliteiten.



## Contents

<b>Summary</b>	<b>i</b>
<b>Samenvatting</b>	<b>iii</b>
<b>Introduction</b>	<b>1</b>
1.1 <i>Information Retrieval</i> .....	3
1.2 <i>On Content-Based Access to Audiovisual Data</i> .....	4
1.2.1 Multimedia Search .....	5
1.2.2 Multimedia Representation .....	8
1.2.3 Multimedia Content Analysis .....	9
1.2.4 Complexity Aspects .....	10
1.3 <i>The Mission, Scope and Organization of the Thesis</i> .....	11
1.3.1 Shot Boundary Detection .....	13
1.3.2 Video Indexing .....	13
1.3.3 Video Summarization .....	13
<b>Shot Boundary Detection based on Spatiotemporal Video Data Blocks</b>	<b>15</b>
2.1 <i>Introduction</i> .....	15
2.2 <i>The Concept of Spatiotemporal Video Data Blocks</i> .....	18
2.3 <i>Feature Extraction</i> .....	20
2.4 <i>Detecting Shot Boundaries</i> .....	25
2.4.1 Cut Detection .....	25
2.4.2 Fade/Dissolve Detection .....	27
2.4.3 Wipe Detection .....	28
2.5 <i>Evaluation</i> .....	31
2.5.1 Evaluation on own Dataset .....	31
2.5.2 Evaluation on the TRECVID Dataset .....	33
2.6 <i>Discussion</i> .....	35

---

<b>Shot Boundary Detection: Problem Solved?</b>	<b>37</b>
3.1 <i>Introduction</i> .....	37
3.2 <i>Qualitative Analysis of Gradual Shot Boundary Detectors</i> .....	38
3.2.1 A Selection of Representative Methods .....	40
3.2.2 Analysis of the Applicability Scope .....	44
3.2.3 Complexity Evaluation .....	47
3.3 <i>Quantitative Evaluation</i> .....	48
3.3.1 Literature-based Performance Evaluation .....	48
3.3.2 TRECVID-based Evaluation .....	51
3.4 <i>Shot Boundary Detection: Problem Solved?</i> .....	54
<b>Content-based Indexing of Live Concert Registrations</b>	<b>57</b>
4.1 <i>Introduction</i> .....	57
4.2 <i>Audio-based Indexing</i> .....	59
4.2.1 Feature Extraction: A Common Approach .....	59
4.2.2 Crossing Rate Features .....	62
4.3 <i>Video-assisted Audio Content Analysis</i> .....	67
4.3.1 Algorithm Design Choices .....	68
4.3.2 Visual Feature Extraction .....	70
4.3.3 Creating Visual Content Clusters .....	72
4.3.4 Updating Audio Content Classes Using Visual Clusters	74
4.4 <i>Experimental Evaluation</i> .....	77
4.4.1 Audio-based Indexing; Evaluation Setup.....	77
4.4.2 Instrument Solo / Applause Detection.....	79
4.4.3 Instrument Identification.....	80
4.4.4 Visually-assisted Indexing .....	82
4.5 <i>Conclusions</i> .....	84
<b>TRECVID BBC Rushes Summarization</b>	<b>87</b>
5.1 <i>Introduction</i> .....	87
5.2 <i>BBC Rushes: Data description and Summarization Requirements</i>	91
5.3 <i>Proposed Rushes Summarization Framework</i> .....	91

---

5.3.1 Data reduction .....	93
5.3.2 Redundancy Handling .....	95
5.3.3 Content Selection .....	97
5.3.4 Information Fusion and Summary Production.....	98
5.4 <i>TRECVID Evaluation</i> .....	100
5.5 <i>Discussion</i> .....	104
<b>Applications</b>	<b>107</b>
<i>A.1 Shot Boundary Detection</i> .....	108
<i>A.2 The MultimediaN Concert Video Browser System</i> .....	109
<i>A.3 Rushes Summarization</i> .....	111
<b>Bibliography</b>	<b>113</b>
<b>Acknowledgements</b>	<b>129</b>
<b>Curriculum Vitae</b>	<b>131</b>



# Chapter 1

## Introduction

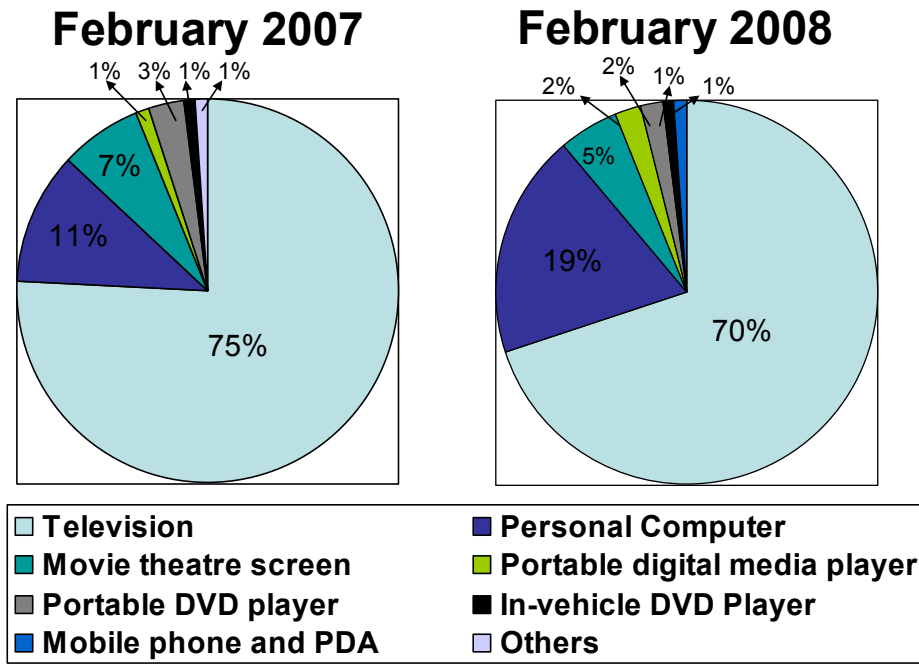
The advances in digital recording technology and the enormous expansion of possibilities to acquire, exchange, store and share non-textual or *audiovisual* information, such as music, images and video, have lead to a rapidly growing share of the data carrying these types of information in private and commercial digital data collections and on the Web in general. Also in absolute terms, the amount of this type of data an average user gets exposed to has increased tremendously over the past years. According to a recent report from the International Data Corporation (IDC) [Gantz08], the amount of digital data created, captured and replicated worldwide will reach 1800 exabytes\* by 2011. Considering that this number was 281 exabytes in 2007, the trend of rapid growth will inevitably keep the discussions regarding physical handling of so much data alive, both concerning their storage and their transfer.

With the increasing amount of data a user gets exposed to, the importance of effectively and efficiently accessing any part of this data in a given context becomes of utmost importance. In this thesis we focus on a particularly important and challenging type of data access, namely *content-based data access* that targets those data clusters carrying meaningful, or *semantic*, content for the user. This content can be characterized in terms of the *semantic concepts* it represents. These concepts can be defined at various (abstraction) levels ranging from *static semantic concepts* (e.g. “car”, “person”, “house”, “beach”, “Amsterdam”, “mountain”, “Eiffel Tower”), via *dy-*

---

\* 1 exabyte equals to 1 billion gigabytes

*namic semantic concepts or events* (e.g. “moving car”, “fight”, “explosion”, “instrument solo”, “vocal section”) to *affective semantic concepts or moods* (“romantic”, “exciting”, “sad”).



**Figure 1.1** Increase in share the Internet takes as a means to consume and share audio-visual data. Internet persistently replaces the traditional sequential-access media, which accentuates the problems of content management, organization and content-based access<sup>†</sup>.

In fact, the increasing need for content-based data access has emerged from the trend towards *non-linear* and *on-demand* data access. In combination with huge sizes of data collections that cannot be searched in a classical (linear) fashion any more, this trend has been supported by the recent developments of the Internet, new Internet-based concepts for interactive information management and sharing, such as Semantic Web [Shadbolt07], Web 2.0 [Magedanz08], social networks [Backstrom06] and Peer-to-Peer (P2P) networks [Park08], and the appearance of Internet-connected, powerful and highly interactive portable devices capable of capturing and displaying audiovisual content. As an illustration, Figure 1.1 indicates an increase of video consumption via personal computers and mobile devices as

<sup>†</sup> Source: Ipsos MediaCT©, 2008.

opposed to the classical TV over the past two years. Internet is the main factor that accelerated the efforts in on-demand data access. It is only after the emergence of the systems and concepts mentioned above that the problems related to data management, organization and content-based access have really turned into practical issues, for which robust and reliable solutions need to be found.

## 1.1 Information Retrieval

The development of theoretical and algorithmic solutions to the problem of content-based access to audiovisual data could be approached by first investigating which solutions there already exist to enable content-based access to textual information, and whether and how these solutions could be expanded to non-textual data categories. The problem of content-based access to text collections has already been investigated for many years and can be defined as the problem of finding the text documents that are meaningful for the user in a given context.

Content-based access to text collections has been studied in the field of computer science referred to as *Information Retrieval* (IR). The term “information retrieval” was coined by Calvin Northrup Mooers [Mooers59] already in 1948 and has been used more and more frequently with the developments in computer and networking technologies and with the data explosion we mentioned earlier in this chapter. IR addresses the problem of searching and finding the most relevant (fragments of) text documents in large text collections. The IR field has reached a high level of maturity over the past years. Critical contribution to this positive development came from the Text Retrieval Conference (TREC)<sup>‡</sup> that has been sponsored by the United States National Institute of Standards and Technology (US NIST) and organized since 1992 with the purpose of providing a discussion forum and a platform for large-scale evaluation of IR theories, algorithms and systems.

An IR process consists of three main steps: *query*, *search* and *presentation*. When the user wishes to access a specific set of text documents in a collection, the search request is first formulated as a query in a format that is appropriate as input into the further IR steps. This is typically a word, a combination of words, word options (i.e. either one word or another one) or a phrase. Based on the query, a specific search algorithm extracts a group of documents that is most relevant to the query. The trivial way of measuring the relevance is by investigating whether there is a match between the words from the query and the words in the document. However, measuring the relevance to the query can also be done in a much more sophisticated fashion. Examples of techniques that can be used for this purpose are those based on word stemming [Paice96], TF-IDF models [Salton88, Robertson97], word collocations [Ferret02] and Latent Semantic Analysis [Kontostathis06], depending on whether word generalization, the occurrence statistics

---

<sup>‡</sup> <http://trec.nist.gov>

and relevance of individual words, or explicit or implicit semantic relations among the words, respectively, are exploited. Furthermore, IR techniques have also been developed for dedicated search domains, such as Internet, where the specificities of the domain are exploited to further improve the document relevance measurement. For instance, Google's PageRank algorithm [Langville06] measures the importance or relevance of a web page to a search query not only based on the text content in that page, but also by investigating the hyperlinks leading to that page. Finally, in the last step, the documents are presented to the user by a list ranked according to the document relevance.

## 1.2 On Content-Based Access to Audiovisual Data

Since late 1980s, a growing community of scientists has been pursuing the challenge of expanding the philosophy underlying the concepts from the IR field to apply it to audiovisual data. These efforts permitted the creation of a new research field often referred to as *Multimedia Information Retrieval* (MIR). Here, the term *multimedia* serves primarily to indicate a broader scope of data types to be dealt with, compared to the IR field, and encompasses the entire scope of different *modalities*, like visual, audio and text, either taken individually or co-existing together in compound multimedia documents. A good example of such a compound document is a video that we refer to in this thesis in general as an image sequence with an accompanying soundtrack.

As a clear indication of the growing importance of multimedia information retrieval in the IR community, a separate *video retrieval track* has been added to the scope of TREC in 2001. Later in 2003, this track became an independent evaluation platform and conference, named TRECVID<sup>§</sup>. The main reason behind the separation of TRECVID from the TREC initiative was the rapid development of the MIR field and a growing scientific community requiring a strong, dedicated evaluation benchmark involving common datasets and protocols for large-scale evaluation of MIR theories and algorithms.

Three main research challenges have emerged in the MIR field in the past years. These challenges are listed in the items below and addressed in more detail in the remainder of this section:

- **Multimedia Search** - how to search for a required piece of multimedia information in a given data collection,
- **Multimedia Representation** - how to present the multimedia content to the user in a compact and intuitive manner, and
- **Multimedia Content Analysis** - how to preprocess multimedia data to identify data chunks with coherent content.

---

<sup>§</sup> <http://www-nlpir.nist.gov/projects/trecvid/>

### 1.2.1 Multimedia Search

The multimedia search challenge lies in enabling query formulation and document relevance computation for an image, audio or video collection with the same robustness and reliability as in the IR systems operating on text document collections. Three main research directions have emerged towards a realization of such an expansion:

- **Query-by-Example (QBE)**, where the pieces of audiovisual data are used as an alternative to a text query to search in audiovisual data collections,
- **Multimedia Indexing**, where the chunks of multimedia data are first assigned textual labels corresponding to the semantic concepts represented by the data, so a multimedia collection can be searched using text queries, and
- **Query and Refinement**, where the multimedia data collection is first searched using a query from one modality, and then the search results are updated based on a refinement of the initial search results by using the information from alternative sources, either through user interaction (relevance feedback) or from a different modality (re-ranking).

#### *Query-by-Example*

Inspired by the achievements in text-based search and driven by the increasing maturity of the computer vision field, the scientific community considered already in late 1980s to specify search queries in a non-textual form [Yoshitaka99] and to extrapolate the text-based search concept onto non-textual data collections. First attempts in this direction addressed image databases, for which the difficulty in expressing the content of an image using a simple text query (“an image is worth more than a thousand words”) led to the need for an effective alternative solution. This solution envisioned the use of an example image as a query and the formulation of the search task as to find all other images in the collection that are similar to the query in terms of its content. The similarity between the query and other images in the data collection is then computed using the signal properties, or *features*, that are extracted from the images and used to represent the image content. Typical examples of image features are color, texture and shape, which can be quantified using a number of different models like, for instance, the histograms representing color distributions, autocorrelation function or fractal dimension describing a texture, or edge maps revealing the shapes of the objects found in an image [DelBimbo99].

Initial attempts to realize the QBE concept in the visual domain assumed a rather simple relation between the image features and the content aspects of an image of interest. For instance, instead of targeting the retrieval of the images that

represent the concept “Paris”, the early systems like [Kato91], [Gong94], [Ashley95], or [DelBimbo96] were only able to retrieve the images showing a sufficient similarity in terms of a selected feature and the applied feature model (e.g. “Find me all images with the same color distribution like in the query”). While more sophistication in the search process has been introduced by the approaches that tried to consider the structure of the visual scene by looking at the coexistence and the spatial relations among the objects [Haarslev97], a real breakthrough could be achieved only by coming much closer to the actual (semantic) content of images when choosing the type and level of feature-based image representation. Examples of such representations are the *Scale-Invariant Feature Transform* (SIFT) [Lowe04] that enabled various advanced applications, like object-based image/video search (e.g. [Grauman06], [Sivic08]).

Successes in expanding the QBE principle from the text domain have also been achieved in the audio domain. There, various music retrieval solutions have been proposed, like Query-by-Humming [Ghias95] and cover song retrieval [Ellis2007], [Sera2008], which aim at finding all songs that are similar to the hummed melody, or are different versions of the query song, respectively.

### *Multimedia Indexing*

A straightforward expansion of the IR retrieval concept to audiovisual data would be through adding textual annotations to images and (pieces of) audio or video data streams. While this process was done manually in the past, there has been an increasing effort over the last years towards developing automated multimedia indexing mechanisms.

Automatically assigning text labels to pieces of audiovisual content requires solutions that are capable of bridging the so-called *semantic gap*, which was defined in [Smeulders00] as the “lack of coincidence between the information that one can extract from the digital data and the interpretation that the same data has for a user in a given situation”. In other words, theory and algorithms are required that are capable of revealing the presence of a semantic concept in an image or a video clip based on the measurements performed at the feature level and providing the information about e.g. color distribution or edge statistics in an image, sound energy and pitch of the audio track, or motion intensity and motion field entropy measured between consecutive frames of a video. The development of multimedia indexing algorithms has typically been approached through the deployment of the techniques from machine learning and pattern recognition to train semantic concept detectors based on features extracted from data. Enormous complexity and challenge in searching for solutions for the multimedia indexing task has inspired and involved a steadily growing research community since the end of 1990s.

While numerous indexing solutions have been proposed so far (e.g. [Hauptmann08]), recent TRECVID reports [Over08b] have shown that their performance has still not reached the level at which their practical deployment would

become possible. An explanation for this is not difficult to find: enabling unconstrained text-based multimedia search by relying on semantic concept labels would require (a) that the entire searchable semantic space of the data collection is represented by a limited set of labels, and (b) that all pieces of audiovisual data corresponding to a given semantic concept can be retrieved from the collection using the corresponding label. Fulfilling these requirements in a general case is a rather difficult task, mainly due to the enormous richness of the content in a typical collection and its possible interpretations by the users. Not surprisingly, first analyses have already been published questioning the current principles underlying the development of automated multimedia indexing solutions and searching for theoretical and practical limits of such solutions [Hauptmann07].

More recently, the research effort towards developing automated multimedia indexing solutions expanded to the affective level. Approaches have been proposed that aim at assigning labels to audiovisual content describing the affective state this content may elicit at users. Examples of affective labels are “sad”, “happy” and “romantic”. In addition to the knowledge base exploited by the traditional indexing methods mentioned above, the affective indexing methods (e.g. [Chan05]) also build on the knowledge from the fields of affective computing [Picard00], psychology and psychophysiology [Hanjalic05].

### *Query and Refinement*

In view of the considerable problems in bridging the semantic gap for solving the multimedia indexing problem in a general use case, iterative search concept has been introduced as a viable alternative. There, additional information, other than that used in the initial search step, is deployed to refine the results obtained in this step and steer the search process towards the relevant set of multimedia documents. After a first set of retrieved documents is generated using a classical QBE approach, the additional information is typically deployed to provide a relevance-based weighting of the elements in the retrieved set with respect to the query. Then, in the next iteration, this weighting is used to modify the search criteria towards an improved set of retrieved documents. Ideally, after a limited number of iterations, all retrieved documents will be relevant. Compared to the multimedia indexing task, in which the relevance of the multimedia documents is specified a priori, by defining the training data sets used to build semantic concept detectors, the document relevance in the relevance feedback approach is learned iteratively, at query time.

The iterative approach described above has originally been introduced for the task of image retrieval [Rui98], and has relied on a human (user) to provide relevance weighting after each iteration. More recent examples of this interactive *relevance feedback* approach can be found in [Zheng08] and [Tao09]. Variations to this concept have also been proposed, such as the one considering active learning. There, mechanisms are deployed that steer the process of collecting the input from

the user in order to make this input as useful for the search process as possible [Huang08]. Putting the user centrally in the search loop is not only beneficial for bridging the semantic gap, but also for refining the set of relevant multimedia documents towards the personal preferences of the user and for the given use context.

More recently, another class of iterative multimedia search approaches emerged, in particular for the video retrieval task. These approaches can be referred to as *video search re-ranking*, and -as opposed to the interactive ones- do not rely on the involvement of the user in the iterative search process. Instead, information from a different modality is used to refine the initial search result that is typically obtained in the text domain and using a text-based query. Recent approaches in this direction (e.g. [Hsu06], [Tian08]) rely on visual modality in the refinement phase and deploy different optimization mechanisms to infer the best possible re-ranking of the search results after using the visual information and by preserving the basic relevance information contained in the initial text-based ranked list.

### 1.2.2 Multimedia Representation

The second MIR research challenge, i.e. representation of multimedia data, concerns representing large multimedia documents in a compact form that communicates enough information to the user to be able to grasp the content and/or the organization of the content throughout a multimedia document in an efficient, natural and intuitive manner. The main difference compared to multimedia search is that here the retrieval system does not expect any query from the users indicating the type of multimedia documents that are to be retrieved. Instead, the retrieval scenario in this case envisions an interaction with the multimedia collection through a browsing interface, where the collection is offered to the user by means of the most relevant parts of the multimedia content that are detected and put together into *summaries* or *abstracts* automatically by the system. Based on looking at such a compact representation of a given multimedia document, the user may decide whether to retrieve the document or not.

The most prominent example of long multimedia documents for which summaries or abstracts would be helpful is video. *Automatic video summarization* is the common name for the class of theoretical and algorithmic solutions that aim at extracting a condensed version of a video document providing full information about the content of that document. As opposed to a *video abstract*, where only the most relevant parts (e.g. highlights [Hanjalic05]) should be included, the information contained in a video summary should be sufficient to obtain an overview of the entire video content and its temporal development (e.g. a story line).

Creating a video summary can be approached in various ways. A content representation using keyframes is the simplest form of a summary. A *keyframe* is generally defined as a representative video frame which provides critical information

about the video content in a given time interval either in terms of the topic being treated (e.g. in TV news), or regarding the people, objects, scene structure or dramatic events (e.g. in movies) [Hanjalic05]. In a keyframe-based video summarization approach, the extracted keyframes are typically displayed either on a static storyboard, as a dynamic slideshow [Amir03], or in the form of a video poster [Yeung97].

More recently, *dynamic video summaries* have been proposed as well. For instance, in the approach targeting the generation of video skims [Sundaram02], first a measure is defined to evaluate the visual content complexity of a video. Then, the result of this measurement is used to map the video content onto the minimum length frame sequence that is considered to be sufficient to comprehend the content of the video. The summarizing frame sequence is generated by taking into account the informativeness and coherence of its visual content, the constraints of audiovisual synchronization and the visual and audio syntax.

The most critical practical difference between the search and representation challenges lies in the evaluation step. While the results of the search step can be evaluated in terms of how many items are properly or falsely identified as relevant to the query (e.g. using the precision and recall measures) with respect to the available relevance ground truth, obtaining such ground truth for the video summarization task is far from trivial. Different summaries can equally well represent the content of a given video. For this reason, video summary evaluation and benchmarking has become an important research topic itself. In general, two different approaches have been established in this direction. The first one is referred to as *extrinsic evaluation*, which looks at how the created summary helps to perform another task, like for instance, the one of a video editor selecting the raw footage in the process of video montage. For example, [Taskiran06] used an extrinsic evaluation scheme in which the participants were asked to do a multiple-choice test after watching a video summary. The questions were defined to assess the capability of the participants to recognize the key aspects of the content of the video being summarized. The second approach is *intrinsic evaluation*, which only considers the representativeness of the summary in view of the original video. For instance, [Ferman03] used an intrinsic evaluation approach wherein a neutral observer, with the full knowledge of the original video, rated the summary videos for redundant and missing objects and the events covered.

### 1.2.3 Multimedia Content Analysis

Prior to performing search or representation steps on long (temporally extended) multimedia documents like, for instance, video, typically some preprocessing steps are needed either to identify interesting pieces of video to be considered as retrieval units, or to reveal as much of the structure regarding the temporal content flow as possible to facilitate feature extraction. A good example of the former is

segmentation of a video into *semantic segments* or *scenes* characterized by coherent content, and the latter can best be illustrated by video segmentations into *shots*.

A shot is a single continuous camera take. Due to a high video frame rate of 25-30 frames a second, the visual content over a series of consecutive video frames can be considered rather consistent. This visual content consistency, in combination with a typically short duration (e.g. not longer than a couple of seconds) make a shot ideal as the elementary video unit serving as input into further video processing, indexing and representation steps. A *shot boundary* is the transition from one shot to the next one. This transition might be an abrupt one (i.e. a cut), but may also be realized through applying an editing effect between two shots. Examples of such effects are dissolves, fades and wipes. The process of detecting shot boundaries and obtaining a shot-based video segmentation is referred to as *shot boundary detection*. Shot boundary detection is the oldest branch of the MIR research field. What started in late 1980s with a number of pioneering approaches exploiting the visual consistency of the frames within a shot and using simple visual features, like color histograms or edge statistics, to detect a transition between two shots, has grown into a large international effort, the results of which have been evaluated in the TRECVID context for many years.

Regarding the video segmentation into semantic segments, while in TV news or documentaries such segments correspond to news stories or topics, respectively, movies can be divided into scenes or episodes. Different approaches have been proposed for this *high-level video segmentation*. Representative works include time-constrained clustering [Yeung97], time-adaptive grouping [Rui98], content recall [Kender98] and fast-forward shot linking [Hanjalic99]. Although proposed by different researchers and independent of each other, all these approaches can be related to a single generic underlying principle, namely the *principle of content coherence* [Hanjalic04]. According to this principle, the presence of a semantic segment boundary at a given time stamp is evaluated based on the feature-level relations between shots taken from before and after this time stamp. While the concept is generic, domain-specific adaptation is required regarding the feature selection representing the shots. While visual features appear suitable for movies, text and speech features have proved most useful when parsing TV news programs.

#### 1.2.4 Complexity Aspects

In addition to the dimension characterizing different MIR challenges and corresponding theories and algorithms described in the previous sections, these theories and algorithms can also be analyzed along the complexity dimension. Here, we refer to the complexity not with respect to the computational resources required to perform MIR tasks, but rather in view of the abstraction level of the MIR tasks, which is known to determine the difficulty of developing the corresponding algorithms and evaluating them in a practical use context.

If one starts with analyzing the problem of shot boundary detection, one could realize that solving this problem does not require any content interpretation at the semantic level. The task consists primarily of detecting the breaks in the visual content flow, either the abrupt or gradual ones, which can be approached purely at the level of visual features. Compared to this, the step towards detecting semantic segments already requires basic assumptions about the content organization at the semantic level and about the way the feature-level relations among the shots map onto the scene-level content coherence. While no real understanding of content at the level of semantic concepts is required for such segmentation, content understanding becomes necessary in the multimedia indexing task. However, although the presence of the semantic gap imposes significant difficulty in devising robust indexing solutions, it is even more difficult to grasp the importance and relevance of particular pieces of video content when generating a summary. This is due to the fact that not only the binary decision regarding the presence of particular objects or persons in a scene may be relevant for content selection in a summary, but also an interpretation of a particular story context (e.g. episode, news topic) and its relation to other story contexts in a video.

### 1.3 The Mission, Scope and Organization of the Thesis

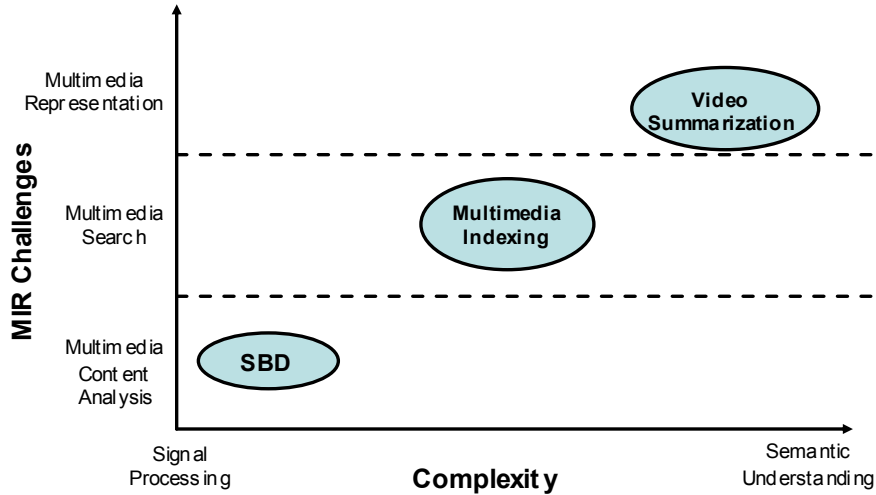
The problem of finding images, video clips and music, given time, place, interest and mood has kept an immense number of scientists and technology developers busy in the past twenty years. However, straightforward attempts to apply text-based search to non-textual data (e.g. Google image search) still seem to be the only viable solution. In spite of the numerous ideas proposed so far in the MIR research field, it is remarkable that hardly any significant success story, and in particular a commercially relevant one, has been reported.

This thesis addresses the reasons that have prevented broad practical deployment of theories and algorithms for searching and retrieving content in multimedia data collections and proposes novel solutions. In particular, the thesis focuses on the problems that typically emerge when dealing with realistic use cases built around real-life systems, noisy data and highly unstructured and diverse content. The details of these real use cases can be found in Appendix A.

Since the MIR field is too large to be considered in its entirety, a number of MIR aspects are selected that cover different MIR challenges and different complexity levels described in Section 1.2. The selected MIR aspects include:

- Shot boundary detection,
- Multimedia indexing, and
- Video summarization

These aspects are positioned in Figure 1.2 along the axes MIR Challenges and Complexity as identified in Section 1.2.



**Figure 1.2** MIR aspects to be discussed in the thesis according to their relation to MIR challenges and the expected level of complexity of related solutions.

The selection of the three MIR aspects listed above is largely motivated by the context in which the research for this thesis has been performed. This context is defined by the Dutch BSIK MultimediaN research program (2004-2009), in which a number of industrial stakeholders expressed interest in practically deployable MIR solutions in a number of real-life use cases. More specifically, a real-time shot boundary detector capable of analyzing general video collections with constant high performance was required (Appendix A1). Furthermore, a system was required for indexing live concert registrations in terms of semantic concepts such as “instrumental section” or “solo”, and for providing non-linear access to different sections (e.g. songs) of a concert (Appendix A2).

This motivation is further strengthened by the selection of tasks and use cases covered in the TRECVID evaluation benchmark, and this particularly related to the shot boundary detection and video summarization tasks. Shot boundary detection has been extensively evaluated over several years and the numerous results reported there and obtained on large and diverse video collections can be seen as an invaluable source of information when searching for an answer to the question whether the shot boundary detection problem can be considered solved. Regarding video summarization, TRECVID has provided the first opportunity for the research community to develop and evaluate video summarization techniques using a

large video collection and covering a real-life use case defined by the professional BBC editors (Appendix A3).

In view of the mission and scope defined above, the thesis is organized in the following thematic units.

### 1.3.1 Shot Boundary Detection

In view of the extensive previous work on the subject and our objective to evaluate practical usability of the available solutions under the real-time constraint and on a general video collection, we split the coverage of this subject over two chapters of this thesis. In Chapter 2, we first propose a novel shot boundary detection algorithm, its design choices, which explicitly address the issues of generic applicability and real-time operability. We will evaluate this algorithm first individually, using a TRECVID data set, and then also in a comparative study in Chapter 3 together with other solutions that were evaluated in the TRECVID context. Furthermore, Chapter 3 will bring a detailed qualitative and quantitative analysis of most representative algorithms proposed in the past years, and provide recommendations regarding the applicability of various individual and hybrid solutions in a real-life use case.

### 1.3.2 Video Indexing

Chapter 4 addressed the problem of indexing the videos of live music concert registrations. We propose a novel approach for automatically indexing such videos in terms of the events that are considered important for effective non-linear content access by a broad audience. Compared to the traditional indexing approaches that aimed at typical use cases considered in the TRECVID context, our approach is devised to handle the complex content of live concert registrations that is characterized by immense diversity of artists, music, visual effects and concert dynamics in general, that contains numerous improvisations and has highly unpredictable temporal structure and event (e.g. instrumental section, vocal solo) realization. We will show that, addressing such an unconstrained use case will require robust and generic solutions.

### 1.3.3 Video Summarization

The fundamental problem in creating usable video summaries is how to take into account the subjective and context-dependent aspects determining the relevance and informativeness of a summary. It is rather unlikely to find a single part of a video that will be perceived as the most important in every possible use case. Fur-

thermore, many different summaries of a given video may be perceived as equally good. Since the ambiguity described above makes it impossible to obtain an insight about the suitability of an existing summarization method for a given use case, learning about the possibilities for generating usable summaries based on the state-of-the-art methods only becomes possible in a broad evaluation effort where different algorithms are tested on the same representative data set and evaluated in view of a single real-life use case. To enable such evaluation, the TRECVID introduced a new video summarization task in 2006. The material selected for the tests consists of unedited (raw) video material that includes lots of retakes and other redundant and irrelevant parts. The task is to automatically create a short summary of this material, which excludes as much of the redundant and irrelevant material as possible, and which is maximally informative about the actual content and the story line. In view of the discussion from Section 1.2.2, extrinsic evaluation criteria are applied to assess how helpful the created summary is for a professional editor in the process of selecting video material for the montage. Chapter 5 of this thesis first presents a novel solution we developed to address this TRECVID task, and then discusses this solution by comparing it with other methods tested in the same context.

## Chapter 2

### Shot Boundary Detection based on Spatiotemporal Video Data Blocks

In this chapter we present and evaluate a simple but effective method for simultaneously detecting shot boundaries of various types by means of an analysis of spatiotemporal video data blocks. This method resulted from our research aiming at jointly improving the detection performance of gradual shot boundaries and reducing the computational complexity of the detection process. While the material presented in this chapter focuses on the rationale, technical details and evaluation of the proposed method, we expand the discussion on this method in Chapter 3, where we compare it with other related and representative approaches and draw conclusions about whether and under which conditions the shot boundary detection problem can be considered solved.

#### 2.1 Introduction

A video shot can be defined as a single cinematic take, during which the camera run is not interrupted. A shot boundary is the connecting region between two consecutive shots, and can be either abrupt or gradual. In gradual transitions, a graphi-

---

This chapter is based on the following publication:

U. Naci, A. Hanjalic, "A unified framework for fast and effective shot transition detection based on analysis of spatiotemporal video data blocks", *Proceedings of the Content-Based Multimedia Indexing (CBMI) Conference*, 2005

cal effect is superimposed to the last frames of the previous shot and the first frames of the following shot so as to generate a smooth transition. While the diversity of possible graphical effects used for this purpose is immense, they can generally be classified into the following three main categories:

- Dissolve,
- Fade (fade-in and fade-out),
- Wipe

In a *dissolve*, the content of the first shot disappears while at the same time the content from the next shot emerges (Figure 2.1a). A *fade* is characterized by darkening of the scene gradually to black and/or recovering of a new scene from a black frame. Figures 2.1b and 2.1c depict these effects. As opposed to a dissolve or fade, which are characterized by a gradual content change in every pixel of a video frame, a *wipe* introduces local abrupt changes in the frame content distributed over time. The effect starts by replacing the old content by the new one in some frame regions and continues until the entire frame contains the material of the new shot. Examples of graphical effects used to realize the wipes are shown in Figure 2.2. If no graphical effects are applied to separate two shots, then the boundary is abrupt. We will refer to this type of boundary as a *cut*.

Detecting the boundaries between consecutive shots in a video is a standard first step in the approaches targeting automatic video content analysis and indexing. Since a cut is the most common boundary type and its detection is a well-defined problem, many methods have been proposed in recent literature, reporting reasonable results in terms of the cut detection performance. However, the problem of detecting gradual boundaries has not successfully been solved yet by the approaches developed so far. Namely, although a vast diversity of methods for detecting various types of gradual boundaries exist (in the next chapter we give a detailed review of the state of the art for gradual transitions), and one could tend to assume that there is sufficient potential for definitely solving this problem, there are many deficiencies that are inherent to the existing methods and that prevent their effective usage in the practice of video content analysis. We elaborate in the following on two major deficiencies.

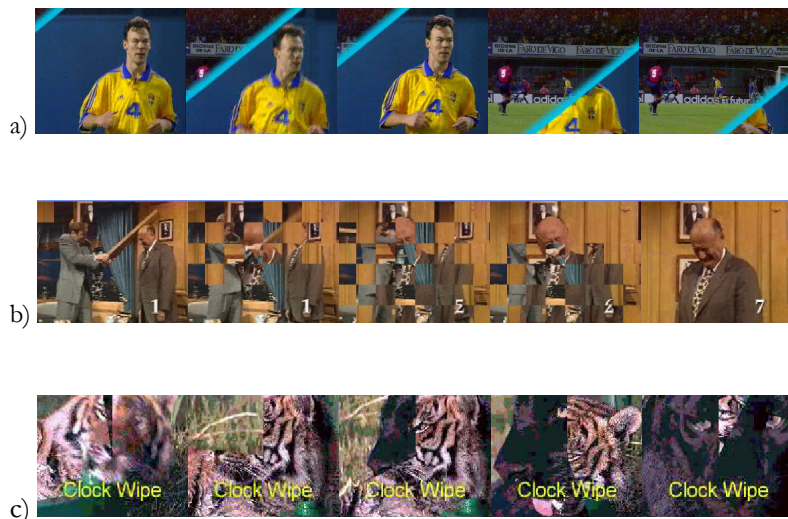
First, the existing systems are not sufficiently capable of coping with the great diversity of measurable signal behavior around and within gradual shot transitions. The origin of this diversity is threefold:

- practically unlimited variation in transitions types, which is due to a vast variety of possible editing effects,
- varying video-directing styles, which is particularly related to the length of a transition, and

- multiple superimposed effects, e.g. the case of a gradual transition accompanied by an object or camera motion.



**Figure 2.1** Examples of smooth gradual scene transitions: (a) dissolve, (b) fade in, and (c) fade out.



**Figure 2.2** Examples of graphical effects: In (a), we see a classical wipe effect from upper left corner to the lower right one. In (b), there is an example for the chessboard effect, and in (c) a clock wipe.

Another important deficiency of the existing methods is that the desired high efficiency of shot transition detection can hardly be matched with a high reliability of detection performance. This results either in fast but unreliable methods or in methods where severe concessions are made with respect to the efficiency in order to improve the reliability.

The research on shot boundary detection reported in this thesis aimed at:

- proposing a shot boundary detection method that is likely to contribute to resolving the deficiencies described above, and
- drawing conclusions regarding the extent to which the shot boundary detection problem can be considered solved in view of the proposed method and the state-of-the-art in the field.

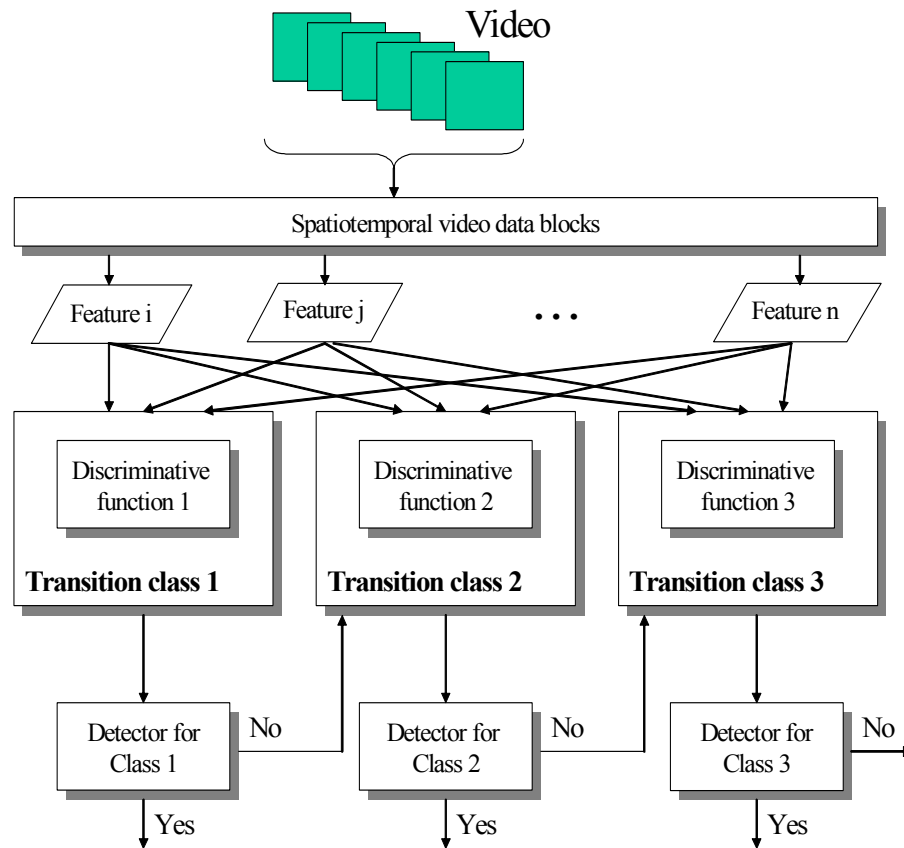
In this chapter we focus on the first item listed above and present the underlying idea and the technical details of our proposed detection method based on the concept of spatiotemporal video data blocks. With this method we aim at jointly reducing the computational complexity and improving the performance of gradual boundary detection both regarding precision and recall and the preciseness of graphical effect extraction in terms of the beginning and ending time stamps. We also discuss the performance of the proposed method on a standard TRECVID test data set. The aim formulated in the second item is addressed in Chapter 3, where we present a comparative analysis that involved our method, the representative methods found in literature and the methods evaluated in the TRECVID context. Based on this analysis we derive conclusions on the limits of individual shot boundary detection algorithms, the possibilities to further increase the detection performance and the conditions under which these possibilities become valid.

## 2.2 The Concept of Spatiotemporal Video Data Blocks

Video is a three-dimensional signal and its properties can best be revealed by simultaneously exploiting all three dimensions of its information flow. Two of these dimensions reveal the visual content flow in horizontal and vertical frame directions, and the third one reveals the variations in this flow over time. The shot boundary detection method we elaborate on in this chapter is based on the extraction of the relevant features from spatiotemporal video data blocks marked by these three dimensions and on modeling of those features to detect and identify a vast range of transition types including cuts, dissolves, fades, wipes and an abundance of different graphical effects used to realize these different boundary classes. The extracted features are mainly related to the behavior of luminance values of pixels in the blocks and form the basis of our unified framework for detecting various transition types. The detection performance is independent of the variations in the form and length (speed) of a transition. Further, as the features used

and the processing steps performed are rather simple, our proposed method is computationally inexpensive. Finally, we are able to detect the beginning and ending time stamps of the transitions with a high level of reliability.

The scheme of our proposed method is shown in Figure 2.3. The features extracted from spatiotemporal video data blocks serve to provide elementary evidence on the presence of a shot transition in the observed time interval. We search for this evidence by investigating local properties of the visual content flow that can help differentiate between the shot transitions and other phenomena in this flow, like those caused by camera and object motion or lighting changes.



**Figure 2.3** Scheme of the proposed shot-boundary detection method.

The feature values collected from a number of neighboring blocks are used to compute the values of *discriminative functions* [Bescos05] for three major classes in which we group all transition types. The discriminative function value serves as an indication for the occurrence of a shot transition from the corresponding class within the observed time interval. The transition class 1 contains cuts. Dissolves and fades are, due to the similar underlying principle, grouped into the same transition class 3, while graphical effects covering most of the remaining transition types (wipes) belong to the transition class 2.

Based on the values of the discriminative functions we compute the probability values for finding a shot transition from a particular class in the observed time interval. The cascade of detectors uses these probability values as input to detect the shot transitions.

We start the technical part of this chapter with a detailed explanation of the feature extraction process in Section 2.3. The actual detection of different transition types based on the computation of discriminative function values and their mapping onto probabilities are explained in Section 2.4. In Section 2.5, we elaborate on the performance of our method based on an experimental evaluation. We conclude this chapter with a discussion in Section 2.6.

### 2.3 Feature Extraction

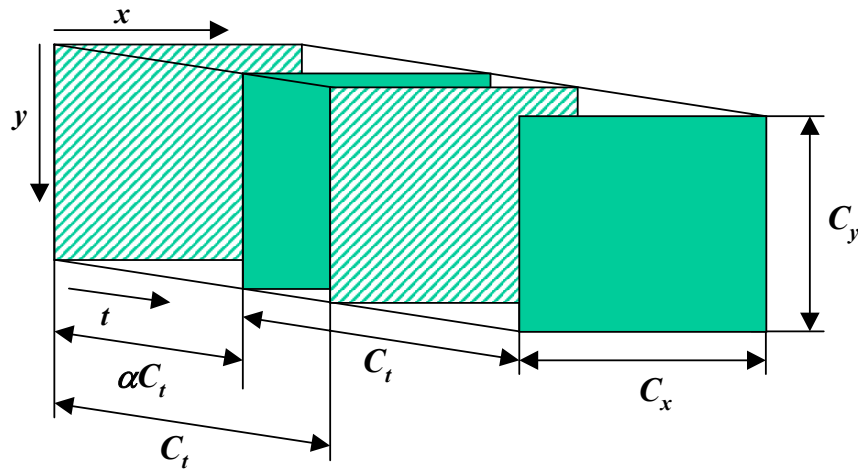
Let video data be defined as a three dimensional discrete function of luminance (intensity) values  $I(x,y,t)$  where  $0 \leq x < X$ ,  $0 \leq y < Y$  and  $0 \leq t < T$ . Here,  $X$ ,  $Y$  and  $T$  represent the horizontal and vertical frame dimensions and the length of the video, respectively. To perform a 3D analysis on the data, we define time-overlapping spatiotemporal data blocks of dimensions  $C_x$ ,  $C_y$  and  $C_t$  and the temporal overlap factor  $\alpha$ . An illustration of these blocks is given in Figure 2.4. We represent each block by the set of luminance values  $I_{i,j,k}(m,n,f)$  of its pixels, that is

$$I_{i,j,k}(m,n,f) = I(m+i \cdot C_x, n+j \cdot C_y, f+k \cdot \alpha \cdot C_t) \quad (2.1)$$

Here,  $0 \leq m < C_x$ ,  $0 \leq n < C_y$ ,  $0 \leq f < C_t$ , and  $0 < \alpha \leq 1$ , while the triplet  $(i,j,k)$  serves to index a block in the totality of video data. To further clarify the variables in Equation 2.1, we emphasize that  $x, y$  and  $t$  correspond to the global indices of a pixel in a video data stream, while  $m, n$  and  $f$  are the local indices in a single spatiotemporal data block. In other words, a pixel at location  $(m,n,f)$  in the spatiotemporal data block  $(i,j,k)$  has the global indices  $(x,y,t)$  computed as  $x = m+i \cdot C_x$ ;  $y = n+j \cdot C_y$  and  $t = f+k \cdot \alpha \cdot C_t$ .

We observed that within a single data block it is sufficient to analyze the changes in the luminance along the time dimension to be able to detect various shot transitions types. In case of a cut, in a block comprising the data from two consecutive shots the majority of pixel luminance tracks will show a large disconti-

nuity at the time stamp of a cut. As partly visible from the examples in Figure 2.2, a wipe is characterized by a series of local abrupt content changes in different frame regions and at different discrete time stamps because of a limited temporal resolution of video. Therefore, the same types of discontinuities in the pixel luminance tracks can be expected in individual data blocks as in the case of a cut. The major difference between a cut and a wipe is that in the case of a wipe the discontinuities are spread over a time interval corresponding to the wipe duration, as opposed to cuts, where the pixel luminance discontinuities are aligned in time, that is, they share the same time index  $t$ . Compared to cuts and wipes, dissolves and fades are characterized by monotonously changing luminance values in spatiotemporal data blocks over the duration interval of a dissolve/fade.

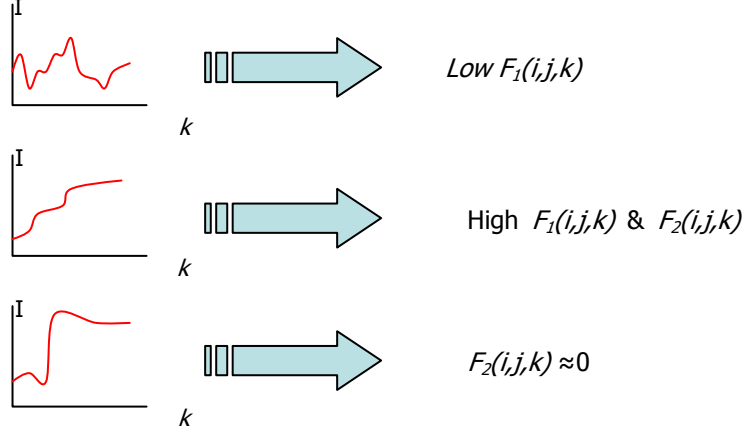


**Figure 2.4** Illustration of two overlapping spatiotemporal video data blocks with an indication of the corresponding parameters.

We now translate the above observations related to the behavior of luminance within a block  $(i,j,\kappa)$  into a quantitative evidence of shot transition occurrence per block by defining the following feature set characterizing each block as a whole:

- $F_1(i,j,\kappa)$ , which evaluates the monotonousness of the luminance flow in the block  $(i,j,\kappa)$  along the time dimension,
- $F_2(i,j,\kappa)$ , which is the measure of abruptness (gradualness) of a change in the luminance flow in the block  $(i,j,\kappa)$  along the time dimension,
- $F_3(i,j,\kappa)$ , which evaluates how simultaneous the changes in the luminance flow in the block  $(i,j,\kappa)$  are at different video frames  $f$  of the block.

Examples shown in Figure 2.5 illustrate the relation between the temporal behavior of the intensity value over consecutive blocks and the features  $F_1(i,j,k)$  and  $F_2(i,j,k)$  that we wish to model. Features  $F_1(i,j,k)$  and  $F_2(i,j,k)$  will be used for detecting dissolves and fades while  $F_3(i,j,k)$  will serve for detecting all other transition types (class 1 and 2).



**Figure 2.5** Targeted relation between the intensity value behavior over consecutive data blocks and features  $F_1(i,j,k)$  and  $F_2(i,j,k)$ . A low  $F_1$  value represents a high fluctuation of the intensity value and indicates that there is no shot boundary in spite of intensive motion in the frames. A high  $F_1$  value indicates, on the other hand, monotonically increasing or decreasing pixel intensity values. Finally, a low  $F_2$  value suggests an abrupt change in the intensity value.

To model the features introduced above, we first search for the derivative values of the function  $I_{i,j,k}(m,n,f)$  along the time dimension. This derivative is estimated as:

$$\nabla_{\mathbf{k}} I_{i,j,k}(m,n,f) = I_{i,j,k}(m,n,f+1) - I_{i,j,k}(m,n,f) \quad (2.2)$$

where  $\mathbf{k}$  is the unit vector in time direction. Then, we calculate two different measures from this derivative per block, namely *the absolute cumulative luminance change*:

$$\nabla_{\mathbf{k}}^a I_{i,j,k} = \frac{1}{C_x \cdot C_y \cdot (C_t - 1)} \cdot \sum_{m=0}^{C_x-1} \sum_{n=0}^{C_y-1} \sum_{f=0}^{C_t-2} |\nabla_{\mathbf{k}} I_{i,j,k}(m,n,f)| \quad (2.3)$$

and the average luminance change:

$$\nabla_{\mathbf{k}}^d I_{i,j,k} = \frac{1}{C_x \cdot C_y \cdot (C_t - 1)} \cdot \sum_{m=0}^{C_x-1} \sum_{n=0}^{C_y-1} \sum_{f=0}^{C_t-2} (\nabla_{\mathbf{k}} I_{i,j,k}(m,n,f)) \quad (2.4)$$

In addition to calculating the values in Equations 2.3 and 2.4, we keep track of the maximum derivative value per pixel track of a block. For each spatial location  $(m, n)$  in the block  $(i, j, k)$ , we search for the frame  $f_{i, j, k}^{\max}(m, n)$ , at which the maximum luminance change takes place, that is:

$$f_{i, j, k}^{\max}(m, n) = \arg \max_f \left( \left| \nabla_{\mathbf{k}} I_{i, j, k}(m, n, f) \right| \right) \quad (2.5)$$

After the frames in Equation 2.5 are determined for each pair  $(m, n)$ , we average the maximum time derivative values found at these frames for all pairs  $(m, n)$ , that is:

$$\nabla_{\mathbf{k}}^{\max} I_{i, j, k} = \frac{1}{C_x \cdot C_y} \sum_{m=0}^{C_x-1} \sum_{n=0}^{C_y-1} \left| \nabla_{\mathbf{k}} I_{i, j, k}(m, n, f_{i, j, k}^{\max}(m, n)) \right| \quad (2.6)$$

The first two of the features we introduced above can now be defined as follows:

$$F_1(i, j, k) = \frac{\left| \nabla_{\mathbf{k}}^d I_{i, j, k} \right|}{\left| \nabla_{\mathbf{k}}^a I_{i, j, k} \right|} \quad (2.7)$$

and

$$F_2(i, j, k) = 1 - \frac{\nabla_{\mathbf{k}}^{\max} I_{i, j, k}}{\nabla_{\mathbf{k}}^a I_{i, j, k}} \quad (2.8)$$

The value of  $F_1(i, j, k)$  equals to 1 if the function  $I_{i, j, k}(m, n, f)$  is monotonically increasing or decreasing, and gets closer to zero as the fluctuations in the function values increase. The higher the value of  $F_2(i, j, k)$  (i.e. close to 1), the more gradual (smooth) are the variations in the function  $I_{i, j, k}(m, n, f)$  over time.

The block points  $(m, n, f_{i, j, k}^{\max}(m, n))$  marking the maximum time derivative values per pixel track in a spatiotemporal video data block are also useful for detecting cuts and wipes. To do this, we calculate the feature  $F_3(i, j, k)$ , which measures whether the dominant changes in the luminance flow occur simultaneously for all pixel tracks, that is, whether the points  $(m, n, f_{i, j, k}^{\max}(m, n))$  form a plane vertical to the time direction. For this reason this feature is obtained as a vector

$$F_3(i,j,k) = \{F_3^f(i,j,k) \mid 0 \leq f < C_t\}$$

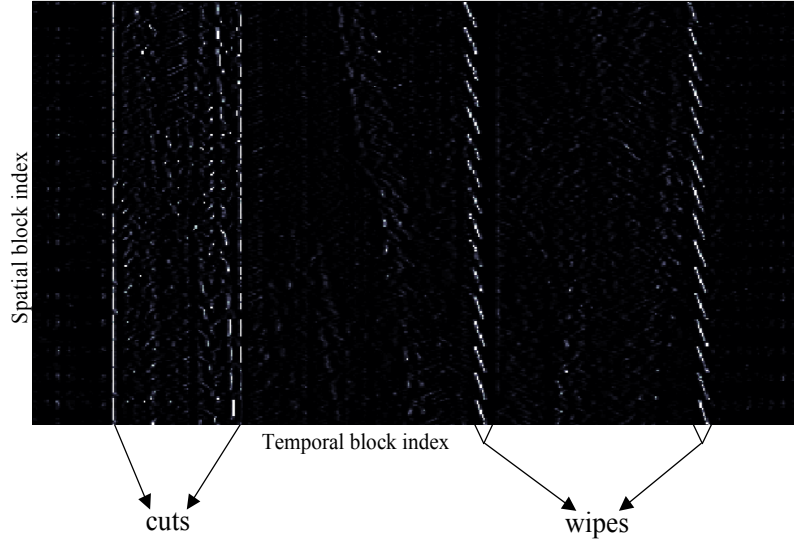
where each component  $F_3^f(i,j,k)$  corresponds to a plane approximation error at the frame  $f$  of a block:

$$F_3^f(i,j,k) = \frac{1}{C_x \cdot C_y} \cdot \sum_{m=0}^{C_x-1} \sum_{n=0}^{C_y-1} \frac{(f_{i,j,k}^{\max}(m,n) - t_{\max dist})^2}{(f_{i,j,k}^{\max}(m,n) - f)^2 + \varepsilon} \quad (2.9)$$

for  $0 \leq f < C_t$  and

$$t_{\max dist} = \begin{cases} 0 & \text{if } f < C_t/2 \\ C_t - 1 & \text{otherwise} \end{cases}$$

Furthermore,  $\varepsilon$  is a small number, introduced to avoid division by zero in case of a perfectly planar distribution of the maximum-derivative points. We emphasize here that in the case of an overlap between consecutive blocks (defined by the factor  $\alpha$ ), Equation 2.9 may be used several times for one and the same frame  $t = kaC_t + f$  of a video, as this frame may correspond to different value pairs  $(f, k)$ . In such cases, the value of the feature component  $F_3^f(i,j,k)$  is computed as the mean of all values in Equation 2.9 computed for the same frame  $t$ .



**Figure 2.6** An illustration of  $F_3^f(i,j,k)$  values along the time dimension.

The matrix in Figure 2.6 depicts the  $F_3^f(i, j, k)$  values for an eight-second sports video that contains two cuts and two wipes. Each column contains the values of  $F_3^f(i, j, k)$  collected row by row from all blocks sharing the same time index  $k$ . The brightness level of matrix elements directly reveals the values of  $F_3^f(i, j, k)$ . We observe that in case of a cut, high values of this feature are time-aligned, that is, they form a plane vertical to the time axis. On the other hand, a wipe is characterized by high feature values, which are not time-aligned, but distributed over a limited time interval. The characteristic regular patterns found for the wipes in Figure 2.6 correspond to the specific wipe type illustrated in Figure 2.2b. One can also observe accidental high feature values between the transitions. These values mainly result from object or camera motion. For instance, the “cloud” of high feature values between two cuts in Figure 2.6 corresponds to a camera following a running player after scoring a goal. In the following section we define criteria for successfully distinguishing between such “clouds” and the patterns corresponding to cuts and wipes.

## 2.4 Detecting Shot Boundaries

### 2.4.1 Cut Detection

To detect cuts, we first integrate the elementary evidence found in the individual blocks and represented by the values  $F_3^f(i, j, k)$ , into the discriminative function  $\psi_1(t)$ , for  $0 \leq t < T$ , which serves as an indicator of a cut at the frame  $t$ :

$$\psi_1(t) = \psi_1(k \cdot \alpha \cdot C_t + f) = \frac{1}{\left[ \frac{X}{C_x} \right] \cdot \left[ \frac{Y}{C_y} \right]} \sum_{i=0}^{\left[ \frac{X}{C_x} \right]} \sum_{j=0}^{\left[ \frac{Y}{C_y} \right]} F_3^f(i, j, k) \quad (2.10)$$

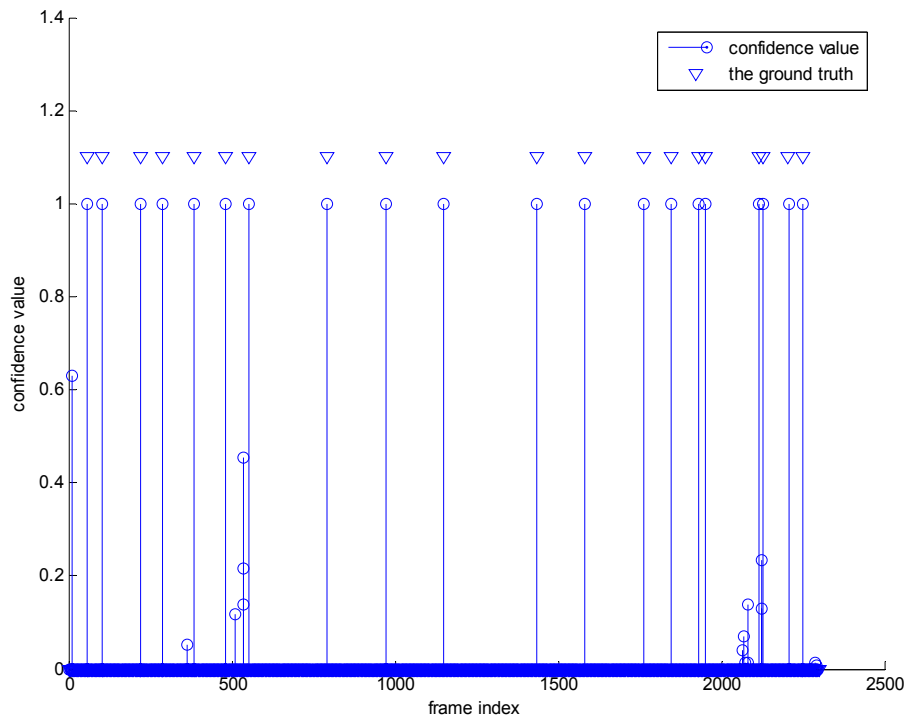
In the next step we apply a piecewise linear mapping of  $\psi_1(t)$  values to the interval  $[0, 1]$  to obtain the probability (confidence) of finding a cut at the observed frame  $t$ :

$$p^{abrupt}(t) = \begin{cases} 0 & , \text{ if } \psi_1(t) \leq A \\ \frac{1}{B-A} \cdot \psi_1(t) - \frac{A}{B-A} & , \text{ if } A < \psi_1(t) < B \\ 1 & , \text{ if } \psi_1(t) \geq B \end{cases} \quad (2.11)$$

Here, the parameters  $A$  and  $B$  are selected based on observing the distribution of the function  $\psi_1(t)$  for cut and non-cut regions in a number of representative

video sequences. Due to a rather clear separation of these regions, the selection of the parameters appears not to be critical for the detection performance and can be kept constant for an arbitrary video being analyzed. For the same reason, a simple fixed threshold can be applied to filter out the cuts.

The mapping in Equation 2.11 is useful for enabling more intuitive selection of the detection threshold than when working with the function  $\psi_1(t)$  directly. This threshold can namely be interpreted as the minimum acceptable probability that a detected cut will not be false. Although this thresholding mechanism is relatively simple, it proves sufficient to obtain a detection performance, which is more than satisfactory compared to the state-of-the-art. This is mainly due to the high discriminative power of the features used. Figure 2.7 illustrates this power on the example of a sample sequence from our test set. All cuts and no false cuts are detected for any threshold ranging from 0.65 to 1. More information about the detection performance and a discussion of problematic cases are given in Section 2.5.



**Figure 2.7** Probability (confidence) values computed for a sample video sequence and aligned with ground truth positions of cuts

### 2.4.2 Fade/Dissolve Detection

Referring to our discussion in Section 2.3, the elementary evidence within blocks for detecting dissolves/fades is contained in the values of the features  $F_1(i,j,k)$  and  $F_2(i,j,k)$ . As opposed to cuts, the locations of which are checked per frame  $t$ , the gradual transitions are investigated using block index  $k$ . We check whether the blocks sharing the same temporal index  $k$  belong to a transition or not. To do this we combine the available evidence from all time-aligned blocks for a given  $k$  into a discriminative function  $\psi_3(k)$  indicating whether the observed temporal video “slice” is a part of a dissolve/fade. We define this function as the average of the feature-based evidence values from all blocks belonging to the observed video slice:

$$\psi_3(k) = \frac{1}{\left\lfloor \frac{X}{C_x} \right\rfloor \cdot \left\lfloor \frac{Y}{C_y} \right\rfloor} \sum_{i=0}^{\left\lfloor \frac{X}{C_x} \right\rfloor} \sum_{j=0}^{\left\lfloor \frac{Y}{C_y} \right\rfloor} (F_1(i,j,k) \cdot F_2(i,j,k)) \quad (2.12)$$

Ideally, the function in Equation 2.12 shows high values for all consecutive video slices belonging to a dissolve, and low values elsewhere. However, to maximize the reliability of function values, we apply median filtering to the function in Equation 2.12 to eliminate its accidental (noisy) value fluctuations. We adopt the result of this operation as the probability that the time interval given by the index  $k$  is captured by a dissolve/fade, that is:

$$p^{\text{dissolve / fade}}(k) = \mathbf{median}(\psi_3(k)) \quad (2.13)$$

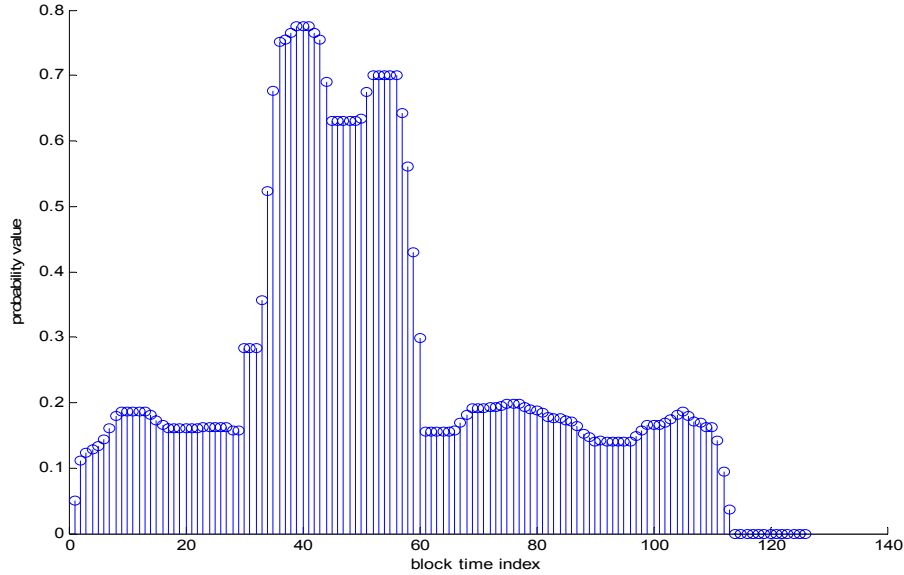
where  $\mathbf{median}(\cdot)$  stands for a median filter. Figure 2.8 illustrates the ranges of probability values corresponding to both detection hypotheses. Similarly as for the cuts, a simple thresholding mechanism can be applied for reliable dissolve/fade detection.

Finally, we find the starting and ending video slices, defined as  $u$  and  $v$ , respectively, of a series of detected consecutive fade/dissolve intervals and choose the frames in the middle of the blocks surrounding the detected block series as the approximate starting and ending frames of a transition:

$$\begin{aligned} t_{start}^{\text{dissolve / fade}} &= u \cdot \alpha \cdot C_t - (C_t / 2) \\ t_{end}^{\text{dissolve / fade}} &= v \cdot \alpha \cdot C_t + (C_t / 2) \end{aligned} \quad (2.14)$$

By artificially extending the dissolve/fade length to the surrounding blocks we generate more confidence that the entire transition is indeed captured by the bor-

ders in Equation 2.14. Due to this, the starting and ending time stamps of a dissolve/fade are found with a maximal error of  $C_s$ .



**Figure 2.8** Probability values computed for a sample video sequence around and within a dissolve region. The dissolve region is clearly indicated by high probability values, which facilitates the detection of the beginning and end time stamps of the transition.

### 2.4.3 Wipe Detection

The detection of wipes in our system is fundamentally the same as the cut detection. Because of the limitations of the video frame rate, the wipes correspond to consecutive abrupt changes in different frame regions that are captured by spatially non-overlapping blocks. We detect a wipe if the blocks at different spatial locations contain abrupt changes in their pixel luminance tracks at different time points, but within a limited time interval.

We first apply Equation 2.9 to calculate the significance of an abrupt change in the pixel luminance track at a frame  $f$  in block  $(i, j, k)$ . Since we assume that the blocks change abruptly only once along a wipe, we relate the obtained result to the sum of the values in Equation 2.9 computed at the neighboring  $2N$  frames surrounding the frame  $f$ , requiring that this sum can not exceed the value  $F_{\mathcal{J}}(i, j, k)$ . Finally, we normalize the result of the comparison with respect to the neighboring frames, as defined in Equation 2.15, to calculate the probability of a wipe-related discontinuity at the frame  $f$ .

$$p_f^{wipe}(i, j, k) = \frac{\max \left( 0, F_3^f(i, j, k) - \sum_{\substack{q=-N \\ q \neq 0}}^{+N} F_3^{f+q}(i, j, k) \right)}{\sum_{q=-N}^{+N} F_3^{f+q}(i, j, k)} \quad (2.15)$$

Here,  $p_f^{wipe}(i, j, k)$  has a high value close to 1 only if the value in Equation 2.9 at frame  $f$  is considerably higher than at all other frames in the neighborhood. As an implementation detail, if the value of  $f+q$  exceeds the block margins, the frames should be taken from the previous or the next block. For example, if  $f+q > C_t$  then

$$F_3^{f+q}(i, j, k) = F_3^{f+q-C_t}(i, j, k+1) \quad (2.16)$$

Just like in the case of cut detection, we now define the discriminative function which indicates whether the frame  $t$  is a part of a wipe. This function integrates the elementary evidence contained in the probability in Equation 2.15 as follows:

$$\psi_2(t) = \psi_2(k \cdot \alpha \cdot C_t + f) = \frac{\sum_{i=0}^{\lfloor \frac{X}{C_x} \rfloor} \sum_{j=0}^{\lfloor \frac{Y}{C_y} \rfloor} p_f^{wipe}(i, j, k)}{\left\lfloor \frac{X}{C_x} \right\rfloor \cdot \left\lfloor \frac{Y}{C_y} \right\rfloor} \quad (2.17)$$

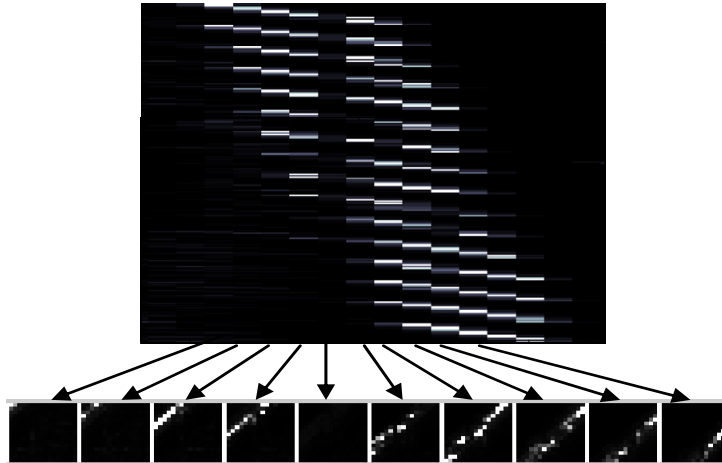
Following the same reasoning as in Section 2.4.1, we map the discriminative function onto the probability of having a wipe at frame  $t$ :

$$p^{wipe}(t) = \begin{cases} 0 & , \text{ if } \psi_2(t) \leq D \\ \frac{1}{E-D} \cdot \psi_2(t) - \frac{D}{E-D} & , \text{ if } D < \psi_2(t) < E \\ 1 & , \text{ if } \psi_2(t) \geq E \end{cases} \quad (2.18)$$

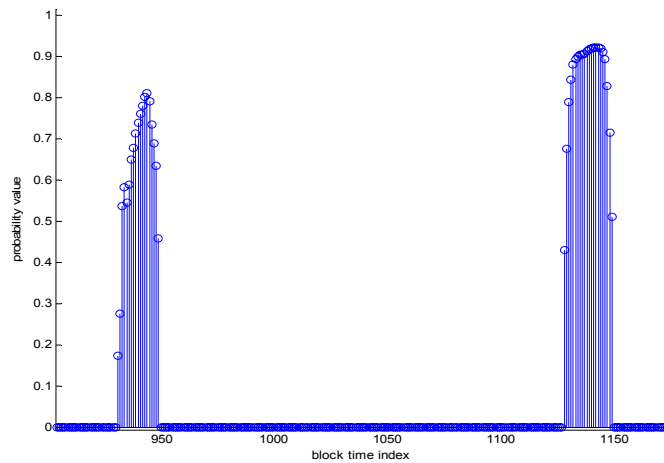
Again, the parameters  $D$  and  $E$  are selected based on observing the distribution of the function  $\psi_2(t)$  for wipe and non-wipe regions in a number of representative video sequences. Figure 2.9 illustrates a zoom-in on a wipe region and probability values for a sample video sequence can be seen in Figure 2.10.

For a series of high probability values calculated in Equation 2.18, we determine the starting and ending time stamps of a wipe in the similar way as in Equation 2.14. Here, however,  $u$  and  $v$  are the frame indices, marking the beginning and ending frame of the detected wipe frame series:

$$\begin{aligned} t_{start}^{wipe} &= u - (C_t / 2) \\ t_{end}^{wipe} &= v + (C_t / 2) \end{aligned} \quad (2.19)$$



**Figure 2.9** A zoom-in on a wipe region from Figure 2.6. When we recombine the column data into 2-D images, the local abrupt changes propagating in the direction of a wipe become clearly visible.



**Figure 2.10** Probability values computed for a sample video sequence around and within two wipe regions.

## 2.5 Evaluation

For a long time, the evaluation of shot boundary detection algorithms has been performed on different isolated datasets not available to a broad scientific community. Although good detection results on such datasets have been reported regularly, a reliable assessment of each individual algorithm was difficult due to the impossibility to test different algorithms on the same data set and measure their relative performance. Driven by the need for more insight into the real performance of the many existing algorithms being developed in the community, the TRECVID benchmarking effort has emerged, which, after its introduction in 2001, has become the most important benchmarking platform for assessing not only the shot boundary detection algorithms but also many other categories of methods in the broad field of multimedia content analysis and indexing.

While the value of TRECVID lies in particular in the availability of a standard large data set on which all algorithms can be evaluated, testing an algorithm in the TRECVID context is in our view not sufficient for obtaining a complete insight into the performance of a particular method. This is because the TRECVID evaluation targets the maximization of the detection performance (e.g. in terms of precision and recall) and does not distinguish between individual algorithms and detection systems that may be built as combinations of many different shot boundary detection methods.

For this reason, the evaluation of the method presented in this chapter is organized as follows: First, we test the method on a smaller but dedicated local data set containing sequences from different genres. We generated this data set specifically to test different aspects of the features and detectors introduced in Section 2.4. In addition to various edit effects, the videos in this test set contain many fast camera movements, which are likely to mislead the process of shot transition detection. Then, we also perform a large-scale evaluation of our method on a standard TRECVID data set purely in terms of the absolute performance. In Appendix A1, we address the implementation of our algorithm in a real-life experimental platform for multimedia content analysis and an objective and subjective evaluation of its performance on many hours of video processed by the platform. Finally, a more thorough comparative study in the TRECVID context aiming at the conclusions not only regarding the relative performance of our method but also the extent to which the shot boundary detection problem can be considered solved is provided in Chapter 3.

### 2.5.1 Evaluation on own Dataset

We created a test database comprising a balanced set of videos from different genres (football, films, documentaries and news videos). The test set contains cuts and different types of gradual transitions. During the experiments we observed that the  $C_r$  value should be kept small so as to secure a continuity of the luminance behav-

ior in the block. We chose  $C_r=5$  for the experiments, with the overlap factor  $\alpha=1/5$ . The latter means that the overlap consists of 4 frames. Similarly, small spatial block sizes should be chosen to be able to capture local luminance changes in video frames. We also observed that especially for detecting wipes it is crucial to use small spatial block sizes. For these purposes, we set the parameters as  $C_x=X/16$  and  $C_y=Y/16$ , where  $X$  and  $Y$  are the width and height values for the video frame respectively. We worked with MPEG-1 video format with frame dimensions of 352x288 pixels. To reduce noise in our data we applied 3x3-mean filter on all video frames in our test set. The values of other parameters introduced in the chapter were set as follows:  $A=100$ ,  $B=500$ ,  $D=0.4$ ,  $E=0.8$  and  $N=7$ .

As shown by the results in Table 2.1, the system succeeded in detecting the shot transitions with a relatively high precision and recall. As expected, the best results were obtained for cuts. We observed only a few missed and falsely detected cuts in cases of highly motion intensive shots and as a consequence of sudden lighting changes, respectively.

	Number	Detected	False positive	Recall (%)	Precision (%)
Cuts	256	253	5	98.8	98.1
Dissolves	55	45	9	81.8	83.3
Wipes	12	11	7	91.7	61.1

**Table 2.1** Performance figures for the proposed shot transition detection algorithm from our data set.

The results for dissolves/fades were obtained when using the same small block size as for detecting cuts and wipes. It is, however, important to note that the performance of fade/dissolve detection increased by about 7% in our experiments, both in the precision and recall, when larger block sizes were used. The consequence is an increase in computational complexity due to the need to repeat some of the detection steps for two different block sizes.

The results on wipe detection are very encouraging as we are able to detect highly diverse graphical effects without implementing algorithms specifically tailored for different effect types. This is one of the major advantages of our method. The relatively low precision rate is a result of objects moving in front of the camera and “simulating” a wipe. An example of such a phenomenon is illustrated in Figure 2.11. Some of these cases show exactly the same properties as wipes and are not likely to be detected without involving higher-level (semantic) video content analysis.

Finally, we evaluated the preciseness of measuring the beginning and ending time stamps of the detected gradual transitions. Most of these time stamps are found within several blocks distance from the true ones. The test results about

detection of transition borders are shown in Table 2.2. The first column contains the deviation expressed in the number of frames from the actual transition border. Negative numbers indicate that the border is detected inside the transition region, which means that we detected the starting time stamp too late or the ending time stamp too early. The results are promising: for both gradual transition classes more than 90% of all borders were found within the tolerance of 10 frames (less than a half a second) from the actual borders. For wipes, 95% of the found borders were even within the 5-frame tolerance interval.



**Figure 2.11** Example showing an object moving in front of the camera. This sequence was falsely detected as wipe.

	Dissolves	Wipes
<-10	7.2%	0
(-10) – (-6)	14.5%	4.5%
(-5) – (-3)	34.5%	18.2%
(-2) – (-1)	20.0%	22.7%
(0) – (3)	9.1%	45.5%
(4) – (5)	10.9%	9.1%
(7) – (10)	3.6%	0
>10	0	0

**Table 2.2** The results of detecting the beginning and the ending time stamps of gradual transitions.

### 2.5.2 Evaluation on the TRECVID Dataset

TRECVID 2005 experiments were quite useful not only for comparing our method with the state-of-the-art in the field, but also for expanding the insights regarding the possible detection problems when tested on a new, independent test platform. We made three main observations based on the tests:

- Our block-based method shows the potential of handling a few common problems that the standard frame-based methods continuously fail to address,
- The gradual transition detection suffers from the single-directional analysis in 3D data blocks,
- The compatibility between TRECVID rules and the output of our detectors was suboptimal, which created artificial deterioration in performance figures.

The first observation revealed that the proposed method is able to handle naturally the fast motion and complicated graphical effects in general, although no specific action, like for instance motion estimation/compensation, has been taken to address any individual potential problem (except for a simple flash detector).

The second observation concerns a relatively low recall rate in gradual transition detection (Table 2.3). The system currently takes into account the evolution of data in the spatiotemporal blocks only in the time direction. This causes a problem when the dissolves are combined with motion. Most of the misses in the gradual transition detection (especially in dissolves) stems from this fact.

	Recall (%)	Precision (%)
Abrupt	91.8	82.3
Gradual	39.8	81.1

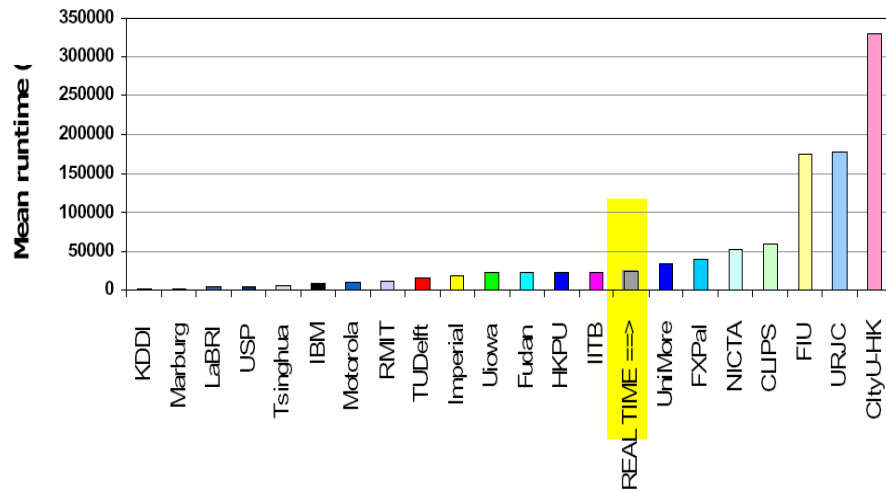
**Table 2.3** Performance figures for the proposed shot transition detection algorithm.

Finally, we observed that there exist excessive amount of extremely short gradual transitions in the test data (i.e. 3 frames in length) that were detected by our algorithm consistently with one frame lag. As an example, if there is a transition between frames 21-23 we detect it as a transition between frames 22-24. Because this is considered in TRECVID as one false detection and one miss, our cut detection results decreased artificially by 5 to 10% both in terms of recall and precision (Table 2.4). Another source of error was that we consider fade in-outs as two separate transitions. More precisely, if the screen turns to black and then dissolves into the following shot, we consider the black screen (or the graphical effect in between) as a separate shot and announce 2 transitions.

	Recall (%)	Precision (%)
Abrupt	94.2	94.5
Gradual	49.4	82.0

**Table 2.4** Performance figures after the corrections in short gradual transition detection and fade in-outs.

Finally, as no complex, specialized video or image processing operation is employed, the method is also highly computationally efficient. On a 2GHz Pentium M PC with 1 GB of RAM it takes only 40 seconds to process 10 minutes of video. In Figure 2.12, one may see that our algorithm does quite well in computational efficiency compared to the other TRECVID systems.



**Figure 2.12** Mean system runtimes for the systems participating in TRECVID compared to the real-time.

## 2.6 Discussion

In this work we explored the possibilities of utilizing an approach based on spatio-temporal block-based analysis of video for constructing a unified framework for detecting and identifying different types of shot transitions. Our proposed method generally performed well, which can be seen from both objective and subjective evaluations, on both the dedicated and widely used video data sets. Although the algorithm showed some deficiencies under very fast object and camera motion and sudden illumination changes, it is unlikely that these deficiencies could be improved without a higher-level (semantic) analysis of video. Furthermore, the recall performance of the algorithm in the TRECVID context was lower than expected. However, this can be explained to a large extent by the specificities of the TRECVID evaluation process.

The biggest contribution reported in this chapter is the provision of a unified framework for detecting a vast diversity of shot transitions with a reasonably high performance and at a relatively low computational cost. Considering a steady increase in the diversity of graphical effects in broadcasting, the proposed method is promising as it can handle this diversity with minimum restrictions and assumptions. As we will explain in Chapter 3, the systems achieving a higher performance are typically based on a combination of multiple systems trained to handle various possible cases in graphical effects, which, however, comes at the expense of a much higher complexity and computational cost.



## Chapter 3

### Shot Boundary Detection: Problem Solved?

The solution for the shot boundary detection problem we proposed in Chapter 2 was inspired by the deficiencies of the existing methods and our aim to jointly improve the detection performance regarding gradual shot boundaries and reduce the computational complexity of the detection process. In this chapter we expand the analysis of the performance, applicability scope and complexity to a larger set of the existing representative shot boundary detection methods with the goal to investigate to which extent the shot boundary detection problem could be considered solved.

#### 3.1 Introduction

As already stated in Chapter 2, shot boundary detection is the first necessary step in many video processing chains targeting content-based video analysis and semantic inference. Developing shot boundary detection methods and continuously improving their performance, applicability scope and the computational efficiency has belonged to the *Multimedia Content Analysis* (MCA) research practice already for twenty years. In fact, it can be said that the MCA research field was initiated in the late 1980's by the first attempts to detect boundaries between consecutive video shots.

In view of many significant results obtained so far in solving the shot boundary detection problem, and the observable saturation in the performance increase over the past years, due to which the evaluation of this problem is not pursued any

more in the scope of the TRECVID benchmark (starting from TRECVID 2007), a valid question to raise at this point is whether and under which conditions the shot boundary detection problem could be considered solved. To address this question, we pursue in this chapter an extensive comparative evaluation of a number of representative state-of-the-art methods, including out method presented in Chapter 2. We perform both qualitative and quantitative evaluation, the qualitative one addressing the applicability scope and complexity of the methods and algorithms considered, and the quantitative one focusing on the detection performance figures, either those reported in literature or those obtained in the TRECVID evaluation context.

Since the detection of abrupt shot boundaries is a reasonably well defined problem, the performance of the modern cut detectors is not likely to vary much under the changing content and context conditions. However, this cannot be said for the gradual shot boundary detectors. There, various editing and style effects, referred to in [Hanjalic02] as extreme factors and typically including strong object/camera motion and lighting changes, still have large influence on the performance and applicability scope of the detectors. Therefore, we focus in the following section on the qualitative analysis of a number of representative gradual shot boundary detectors in order to understand better the phenomena underlying their performance, complexity and application scope. This is then followed in Section 3.3 by an analysis of the detection performance. Finally, based on the material from sections 3.2 and 3.3, we discuss in Section 3.4 the current solution status of the shot boundary detection problem.

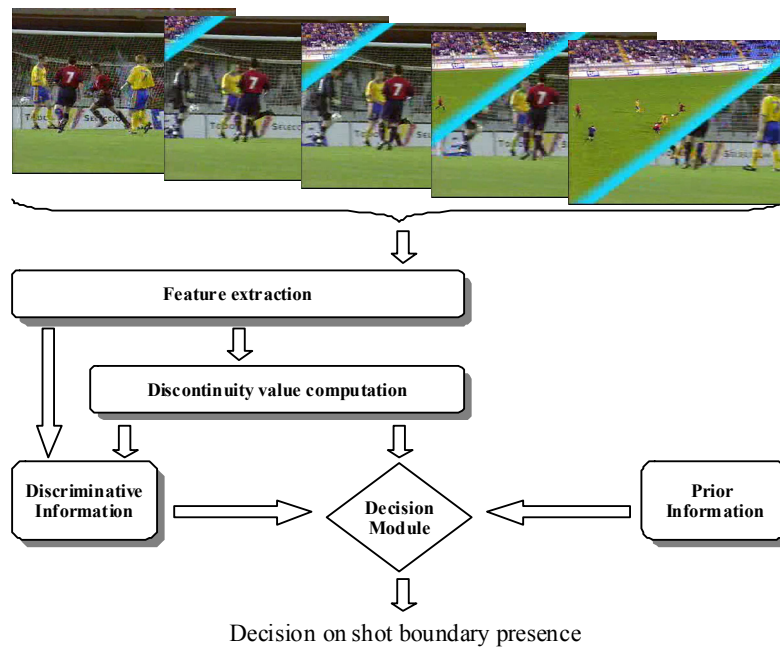
## 3.2 Qualitative Analysis of Gradual Shot Boundary Detectors

A gradual shot boundary detection system is generally composed of three major building blocks, namely,

- feature extraction,
- modeling, and
- decision-making.

Since the features extracted from the frames enable the computation of *discontinuity values* revealing the changes in the visual content flow in video, they provide the basic information for successfully detecting gradual transitions. Ideally, the changes in the visual content flow will be larger within a transition than within a shot, which makes the discontinuity value an important clue in the detection process [Hanjalic02]. However, solely relying on the discontinuity values is not sufficient in most cases. Because the extreme factors, such as fast motion or sudden lighting variations, can also produce high discontinuity values, in the range that is similar to the one produced by gradual shot boundaries, additional clues are

needed to obtain more confidence regarding the presence or absence of a gradual boundary at a given video segment. An example of such additional clues is the information about the pattern created by the discontinuity values and its match with the expected pattern corresponding to a particular gradual transition. This information can also be referred to as *discriminative information* [Hanjalic02]. For instance, Meng et al. [Meng95] make use of the typical parabolic pattern of pixel intensity variance along the frames within the boundary to improve the detection of dissolves. This type of discriminative information is, just like the discontinuity values, based on direct feature measurement. In addition, we also mention here the *structural* type of discriminative information. This type expands the information about the range of the discontinuity value at a given time stamp by the relative information about this range based on the comparison with the discontinuity values from the neighboring time stamps. Moreover, the *prior information* about the video being analyzed can be helpful in the detection process [Hanjalic02]. Both the discriminative and the prior information are typically considered by developing various *boundary models*, the output of which is processed together with the discontinuity values in a detection (decision-making) module. An illustration of the process flow of a typical gradual boundary detector is given in Figure 3.1.



**Figure 3.1** Block scheme illustrating the process flow of a general gradual shot boundary detector.

Clearly, a classification of methods for gradual transition detection can be based on the feature sets that are employed, on whether and which prior or discriminative information is used, and on the type of the detection mechanism used. While the detection mechanisms (e.g. statistical ones) are rather general, and since not all methods use discriminative and prior information, the most fundamental classification of these methods can be done in terms of features. With this respect, there are four fundamental categories of gradual shot boundary detection systems:

- pixel-based systems,
- histogram-based systems,
- edge-based systems, and
- motion-based systems

The simplest category, the pixel-based systems, decide on the presence of a gradual shot boundary based on collecting the information on the pixel values in the frames, their statistical properties and temporal changes. Shot boundaries may also be detected by tracking the changes in the histograms of successive frames. As histograms are more robust against camera and object motion, this approach has proven to be more effective in reducing the number of (motion induced) false detections. Even more robustness to motion effects can be obtained by applying the systems measuring the motion field between the consecutive video frames, and especially those systems that apply motion compensation. However, compared to the histogram-based systems, higher robustness to motion typically goes together with a considerable increase in computational complexity. Finally, another class of approaches is based on observing the changes in the edge statistics between the consecutive video frames. We refer to [Hanjalic04] for a detailed overview of the shot boundary detection methods classified based on feature selection.

For the discussion in the context of this chapter, we selected a representative group of methods to reflect all four classes of approaches. We selected the methods that are tested more than once and/or on sufficiently large video collections. We also considered some promising hybrid and conceptually atypical methods, including the one we introduced in Chapter 2. All selected methods (except the one from Chapter 2) are explained in more detail in the following section.

### 3.2.1 A Selection of Representative Methods

The first approach we choose to discuss is the *Variance Curve Approach* by Meng et al. [Meng95]. This approach examines the pixel intensity variance along the frames for the detection of dissolves. Since a downward parabolic pattern is expected to indicate a dissolve along a group of consecutive frames, the detection of a dissolve is approached either by checking the main parameters of the parabolic pattern (e.g.

the height and the width) or by fitting the mathematical model of the parabolic curve to the series of computed variance values.

In the *Chromatic Edit Model* proposed by Song et al. [Song98], the first and the second partial derivatives of pixel intensities with respect to time are investigated for gradual transition detection. The assumption in [Song98] is that during a linear gradual transition, the first partial derivative of the pixel intensities with respect to time remains constant, while the second partial derivative with respect to time should be close to zero. Because of the noise created by the motion during the transition, Song et al. [Song98] used the proportion between the first and the second partial derivatives as the metric. If the second partial derivative is less than a fraction of the first derivative along a number of frames, a gradual transition is detected.

The *Plateau Approach* by Yeo and Liu [Yeo95] is based on computing the discontinuity values as the differences  $D_i^k$  between the DC coefficients of all frame pairs  $i$  and  $i+k$ . Due to the temporal lag  $k$  the discontinuity values  $D_i^k$  are expected to form a “plateau” during the gradual boundary, that is, a series of relatively higher discontinuity values is assumed to characterize a boundary. Then, two criteria are applied to detect a boundary in a plateau region. First, the variation of the discontinuity values along the plateau is required not to be too large. While Cheong [Cheong00] allowed up to 20% variation, Kobla et al. [Kobla99] set the tolerance as 10%. The second criterion,

$$D_i^k \geq l \times D_{i-k/2-1}^k \quad \text{or} \quad D_i^k \geq l \times D_{i+k/2+1}^k \quad (3.1)$$

is used to check whether the difference values in the plateau region are considerably (factor  $l$ ) higher than those measured between the plateau region and the points before the rise and after the fall of the plateau. However, both Cheong [Cheong00] and Kobla et al. [Kobla99] used the differences  $D_{i-k-1}^k$  and  $D_{i+k+1}^k$  measured over a longer frame range, instead of  $D_{i-k/2-1}^k$  and  $D_{i+k/2+1}^k$ , claiming that this was better for the detection of long-lasting transitions. Various implementations of this method (e.g. [Cheong00], [Kobla99], [Ankush02]) also used different values of  $l$  in Equation 3.1.

Ankush et al. [Ankush02] combined the *Double Chromatic Difference* (DCD) approach [Yu97] and the plateau method [Yeo95] for an improved gradual transition detection performance. The DCD sequence is defined as

$$DCD(t) = \sum_{x,y} F \left( \left| \frac{g(x,y,t_0) - g(x,y,t_N)}{2} - g(x,y,t) \right| \right) \quad (3.2)$$

With  $t_0 \leq t \leq t_N$ , where  $t_0, t_N$  are two points within a dissolve interval,  $g(x, y, t)$  is a frame, and  $F(\cdot)$  is a threshold function. Just as in the variance curve approach,  $DGD(t)$  shows a roughly parabolic shape along a dissolve and the detection mechanisms are similar in both methods. The plateau algorithm [Yeo95] is integrated to detect wipe regions and fades. Besides detecting fades, the plateau algorithm is claimed to detect the beginning portion of wipes. The end portions of the wipes are detected using a statistical wipe detector whose details are presented in [Alattar98].

The *Twin Comparison Method*, first proposed by Zhang et al. [Zhang93] is a typical example of a histogram based detection system. In this method successive frames are compared using a histogram difference metric. Global histograms are evaluated on the hue color channel with 256 bins for each histogram. Two thresholds are used for boundary detection; a higher threshold  $T_b$  and a lower threshold  $T_s$ . A cut is detected when the difference between two consecutive frames is greater than the higher threshold value,  $T_b$ . If the difference is smaller than  $T_b$  but greater than  $T_s$ , this point is marked as a potential start of a gradual transition ( $F_s$ ). The potential end frame of the transition ( $F_e$ ) is marked once the difference between the consecutive frames falls below  $T_s$ . Finally a gradual shot boundary is detected between  $F_s$  and  $F_e$  if the frame difference between these two points exceeds the higher threshold  $T_b$ .

Zabih et al. [Zabih95] and Porter et al. [Porter03] discussed the limitations of the abovementioned methods in terms of vulnerability to motion. Zabih et al. [Zabih95] proposed a motion-tolerable method in which edge change ratios (ECR) are employed to compute the discontinuity values. ECR can be interpreted as the change in the amount of the detected edges from one frame to the next one and is calculated as

$$ECR_n = \max\left(\frac{X_n^{in}}{\sigma_n}, \frac{X_{n-1}^{out}}{\sigma_{n-1}}\right) \quad (3.3)$$

where  $X_n^{in}$  and  $X_{n-1}^{out}$  are the number of entering and exiting edge pixels in frames  $n$  and  $n-1$ , respectively, and  $\sigma_n$  and  $\sigma_{n-1}$  are the numbers of edge pixels in both frames. Finally, the value (3.3) is compared to a threshold to detect shot changes. The method was tested by Porter et al. [Porter03] and Lienhart [Lienhart01a]. Contrary to Zabih et al. [Zabih95], Porter et al. [Porter03] did not aim at compensating for motion but at detecting and modeling camera and object movements. After dividing frames into blocks, they select some of the blocks as *regions of interest* (ROI) such that blocks with low pixel variance values are discarded. Then they track the block motion and search for gradually changing feature similarity values to detect fades and dissolves.

The statistical shot boundary detector approach by Hanjalic [Hanjalic02] is a good example of a motion-based approach. It combines the discontinuity values

obtained after motion compensation with a model for prior probability of a shot transition at a given time stamp and the models for these two types of discriminative information (demonstrated on the example of a dissolve) into a statistical detection framework to optimize the detection performance with respect to the minimization of the detection error probability.

While following a similar strategy as in [Hanjalic02], Bescos [Bescos04] aims at generating rules that are able to handle different pattern models accounting for various possible gradual transition types. There, two criteria are defined for analyzing a series of discontinuity values along a potential boundary region:

- The maximum value of the series should be located in the center.
- The difference between two consecutive discontinuity values cannot be higher than the half of the central value.

To investigate whether these conditions are satisfied for a series of  $2n+1$  discontinuity values, two functions are defined :

$$\lambda_1[k] = \delta[k] - \text{Median}(\delta[k-n], \delta[k], \delta[k+n]) \quad (3.4)$$

and

$$\lambda_2[k] = \frac{\delta[k] - \delta[k-n]}{\sum_{m=0}^{n-1} |\delta[k-m] - \delta[k-m-1]|} \cdot \frac{\delta[k] - \delta[k+n]}{\sum_{m=0}^{n-1} |\delta[k+m] - \delta[k+m+1]|} \quad (3.5)$$

Here,  $k$  is the index of the center of the series, and  $\delta[k]$  is the corresponding discontinuity value. The function  $\lambda_1[k]$  is claimed to highlight the central discontinuity value as it is computed between the frames from different shots, while  $\lambda_2[k]$  measures whether the discontinuity value function is an increasing/decreasing function along the first/last  $n$  elements of the series. Based on the features  $(\delta[k], \lambda_1[k], \lambda_2[k])$  extracted from the training data set and the ground truth boundary annotations, the threshold values are learned, that can be applied to the functions (3.4) and (3.5) to perform boundary detection on new test sequences.

More complex methods have been proposed as well. The first one we mention here is a hybrid method called *The Video Trails* method by Kobla et al. [Kobla99]. They utilize Faloutsos and Lin's *FastMap* method [Faloutsos95] for mapping each frame of the video onto a lower-dimensional space. The feature used here was extracted as a weighted sum of both pixel and histogram differences between frames. The second complex method we include is the *Color Anglogram* technique by Zhao et al. [Zhao03]. Color anglograms are extracted by decomposing an image into non-overlapping blocks and assigning each block a feature vector composed of spatial location and average hue and saturation values. Then a Delaunay triangu-

lation is constructed [Zhao03] and feature point histogram is computed using the two largest (or smallest) angles in the Delaunay triangles. They combined *latent semantic indexing* [Berry99], a singular value decomposition based dimensionality reduction technique, together with color anglograms, which are claimed to outperform the color histograms.

The last method we include in our survey – the *Transition Synthesizer Method* proposed by Lienhart [Lienhart01a, Lienhart01b] - is an interesting alternative to the abovementioned “classical” approaches. A transition synthesizer that is able to synthesize a specified number of transition examples of the specified kinds from a video database is built. They used the synthesized transitions to train a real valued, feed-forward neural network with hyperbolic tangent activation function to serve as the detection module. The system is implemented and trained only for dissolves. Then the tests are run on real video sequences for performance evaluation.

### 3.2.2 Analysis of the Applicability Scope

We determine the applicability scope of a gradual transition detection method based on the following two criteria:

- the number of gradual transition types it is capable of handling, and
- the amount of information a method provides on a particular transition (e.g. type, length, starting and ending time stamps, etc.)

With respect to the first criterion, the methods introduced above can be classified into the methods focusing on particular types of gradual transitions or the methods detecting the occurrence of a gradual transition in general. Examples of the methods from the first group are the variance curve method [Meng95] and the method based on the chromatic edit model [Song98], which are designed for detection of dissolves only. Another method from this group but with a slightly larger applicability scope is the motion based approach [Porter03], which detects and identifies fades and dissolves.

The second group of methods is represented in our survey by the plateau detection [Yeo95], the twin comparison [Zhang93] method, *The Video Trails PixHist* method [Kobla99] and the Color anglogram method [Zhao03]. These methods look for the regions in video where continuous rapid changes compared to in-shot regions accumulate to a significant change in image content over a short period of time. Thus, these approaches focus on detecting a gradual transition without attempting to identify the precise transition type. Although the pattern adaptation method [Bescos04] generally belongs to this group as well, the assumptions set on the feature pattern may restrict the claimed unconstrained recognition of any transition type (e.g. a fade-in not followed by a fade-out).

The system we developed and discussed in Chapter 2 [Naci06] has been designed to fulfill the two criteria defined above to the maximum possible extent.

The underlying generic mechanism evaluating the behavior of pixel intensities in spatiotemporal pixel blocks reacts to the typical phenomena characteristic for a dissolve, fade or a general visual effect used to realize a wipe. However, at the same time, it does not impose any constraints to these phenomena by which its applicability scope would be limited to specific types of visual effects.

The remaining methods considered in our survey fall between the groups mentioned above, as their application scope is extendible. For instance, the transition synthesizer [Lienhart01a, Lienhart01b] is capable of detecting and identifying the transitions that the system is trained for beforehand. So we can say that the application scope of this method is determined by the scope of the training data set. Furthermore, a generic underlying principle makes the statistical shot boundary approach [Hanjalic02] capable of detecting and classifying various transitions types, although it is currently developed and evaluated for the specific problem of dissolve detection. A similar argument holds for the edge-based approach [Zabih95]. Integrated DCD + Plateau approach [Ankush02], on the other hand, is already capable of and tested for detecting and identifying dissolves, fades and wipes.

In terms of the amount of information provided about a transition, it can be observed that the accuracy of the boundaries of a detected transition is often an overlooked criterion in evaluating the shot boundary detection methods. Considering its importance for further steps in content-based video analysis, such as scene segmentation [Hanjalic04], the accuracy criterion should be considered in addition to standard performance measures like precision and recall. Since the boundary accuracy is not frequently evaluated, not many conclusions can be drawn regarding this performance aspect from the evaluations of the methods discussed earlier in this chapter. What can be said is, for instance, that the methods based on the variance curve [Meng95], plateau [Yeo95] and integration of DCD and plateau [Ankush02] (except for detecting boundaries of wipe regions, for which a statistical wipe detector is utilized) are far from being precise, as the patterns they are searching for are rather noisy. A more explicit treatment of this problem can be found in the pattern adaptation method [Bescos04], which attempts to handle the transition boundary detection problem using pattern windows of different lengths. Finally, by focusing on local spatiotemporal phenomena, the method [Naci06] proposed in Chapter 2 is capable of estimating the beginning and ending time stamp of a gradual transition quite precisely, which makes this method rather unique in the field.

We summarized our qualitative evaluation of the applicability scope of the selected shot boundary detection methods in the third column of Table 3.1. Due to the reasons mentioned in the previous paragraph, we based the scores assigned to each method on the first criterion mentioned above. In other words, we referred to the methods detecting only one and two transition types as *Poor*. Methods detecting and identifying two types of transitions are referred to as *Fair*. Detecting multiple types is rated *Good* if identification is not possible and as *Excellent* otherwise. When assigning the scores, we looked at the set of boundaries that were actually considered for evaluation. For instance, while some of the methods, like [Lien-

hart01a, Lienhart01b, Hanjalic02] are conceptually capable of detecting and identifying any transition type, they were evaluated for cuts and dissolves only. For this reason, we graded them as *Good*.

METHOD	DETECTED BOUNDARY TYPES	APPLICABILITY SCORE	COMPLEXITY SCORE
Variance curve [Meng95]	Dissolves	Poor	Excellent
Chromatic edit model [Song98]	Fades + dissolves	Poor	Excellent
Plateau detection [Yeo95]	All (no identification)	Good	Excellent
Integrated DCD+Plateau [Ankush02]	All (identification)	Excellent	Good
Twin comparison [Zhang93]	All (no identification)	Good	Excellent
Edge based [Zabih95]	All (identification)	Excellent	Fair
Motion based [Porter03]	Fades + dissolves (identification)	Fair	Fair
Transition synthesizer [Lienhart01a, Lienhart01b]	All (no identification)	Good	Fair
Pattern adaptation [Bescos04]	All (no identification)	Good	Excellent
Statistical shot boundary detector [Hanjalic02]	All (identification)	Good	Fair
VT-PixHist [Kobla99]	All (no identification)	Good	Poor
Color anglogram [Zhao03]	All (no identification)	Good	Poor
Spatiotemporal blocks [Naci05a]	All (identification)	Excellent	Excellent

**Table 3.1** Qualitative rating based on applicability scope and complexity. *Identification* in the table refers to the identification of the boundary types.

### 3.2.3 Complexity Evaluation

The detection methods based on the variance curve [Meng95], chromatic edit model [Song98], plateau detection [Yeo95] and twin comparison [Zhang93] can be considered the basic (simplest) methods in terms of computational complexity. If we take the complexity of these methods as a reference, we can develop some ideas about the complexity of other methods considered in this chapter. For instance, the pattern adaptation method [Bescos04] is expected to perform very well in terms of computational complexity because it employs simple pixel-level methods. In the same way, the integrated DCD + Plateau method [Ankush02] is also an efficient one because it combines two simple methods with some improvements. However, considering that a statistical wipe detector is also utilized for the detection of the end points of wipe regions, combining so many simple methods may cause degradation in computational efficiency.

Motion-based methods [Hanjalic02, Porter03] are more complex mainly because of motion compensation operations. The same can be said for edge-based approaches which require complicated image processing operations on every video frame. The edge based approach [Zabih95] was shown in [Leferve03] to be more than 200 times more expensive than simple pixel-based methods. Since the transition synthesizer method [Lienhart01a, Lienhart01b] also utilizes edge based contrast property, it can also be classified as an expensive method. Even more complexity can be found in the video trails PixHist [Kobla99] and color anglogram [Zhao03] methods, which require the FastMap [Faloutsos95] clustering method and the latent semantic indexing (LSI), respectively. These complex steps make these methods rather computationally expensive.

Reducing the computational complexity was one of the main criteria we consider while developing our system in Chapter 2. We managed this by relying on simple but effective features and detectors. As one can observe from Figure 2.12 (Chapter 2), our system is around three times faster than real time and it is one of the most efficient algorithms evaluated in the TRECVID context.

In the subjective scoring shown in the last column in Table 3.1, we take the simple methods mentioned in the beginning of this section as a reference and rate them *Excellent*. Since the pattern adaptation method [Song98] also uses these simple methods without any expensive additional operations, it is also awarded the *Excellent* label regarding the computational complexity. Due to a high complexity of motion estimation and compensation steps in the motion-based methods [Hanjalic02, Porter03] and the need for advanced image processing operations in edge-based methods, we rate them as *Fair*. The transition synthesizer is also rated as *Fair* because of some complexity-related drawbacks of its learning based approach. Finally, the *Video Trails PixHist* [Kobla99] and Color anglogram [Zhao03] methods require both dimension reduction and classification operations, which are expensive to perform in the high-dimensional feature space used. Also color anglograms

are computationally expensive features to extract. Therefore, we rated these two methods as *Poor*.

### 3.3 Quantitative Evaluation

Following up on the qualitative analysis of the methods and algorithms for gradual shot boundary detection, we now investigate the quantitative aspects of their performance. We first perform a comparative study of the methods introduced above in terms of the detection performance for gradual boundaries based on the evaluation provided in literature by the authors themselves or by other authors who implemented and tested their methods. Then, we focus on the TRECVID evaluation benchmark and analyze the detection performance obtained for a large, standard and representative video data set. We would like to emphasize that the analysis of the TRECVID-based detection performance will be done on a set of methods different than the one presented in Section 3.2.1. This is because the representative methods we selected before were not all evaluated in the TRECVID context, and not all TRECVID-evaluated methods were representative enough or were described well enough to be used for the qualitative analysis.

#### 3.3.1 Literature-based Performance Evaluation

In Table 3.2, we show the performance results from the tests of the abovementioned methods together with a subjective rating. We rated the methods performing on average with more than 80% precision and recall as *Excellent*. Those between 70% and 80% are referred to as *Good*, those between 50% and 70% as *Fair*, and those below 50% as *Poor*. If more than one assessment is available on a method, we employed the majority voting. For the chromatic edit model approach [Song98], the results obtained in [Cheong00] are adopted as the base for the score. The reasons for this choice are explained in the paragraphs below, where we briefly address the most typical results from Table 3.2.

The best results were reported for the methods from [Zhao03], [Hanjalic02] and [Porter03]. Considering that two of these methods ([Hanjalic02] and [Porter03]) are based on motion features, we may conclude that it is critical to take the motion information into account when building a shot boundary detection system. We made this observation when building our own system in Chapter 2 as well.

The results obtained by Cheong [Cheong00] and Kobla et al. [Kobla99] for the variance-curve based method [Meng95] are reported in the first row of Table 3.2. In addition to dissolves, their test set also contained some other transition types. However, as the variance curve approach is specific for the dissolve type of transitions, the transition diversity in the data set is the main reason for a relatively low recall rate computed for all transitions together. At the same time, a relatively low

precision rate indicates that the system is also prone to errors caused by many other effects producing parabolic patterns similar as those in the dissolves.

The method	Tested in	Recall	Precision	Rating
Variance curve [Meng95]	[Cheong00]	67.0%	48.6%	Fair
	[Kobla99]	63.4%	48.4%	
Chromatic edit model [Song98]	[Cheong00]	83.7%	72.0%	Good
	[Kobla99]	18.3%	63.2%	
Plateau detection [Yeo95]	[Cheong00]	86.7%	74.0%	Fair
	[Ankush02]	46.9%	38.8%	
	[Kobla99]	49.4%	55.1%	
Integrated DCD+Plateau [Ankush02]	[Ankush02]	75.4%	74.7%	Good
Twin comparison [Zhang93]	[Cheong00]	85.4%	70.3%	Good
	[Ankush02]	56.9%	64.5%	
	[Kobla99]	61.2%	53.7%	
	[Porter03]	58%	85%	
	[Lupatini98]	76.3%	64.8%	
Edge based [Zabih95]	[Porter03]	45%	33%	Poor
	[Lienhart01a]	65.1%	3.5%	
Motion based [Porter03]	[Song98]	88%	77%	Excellent
Transition synthesizer [Lienhart01a, Lienhart01b]	[Lienhart01a]	61.6%	50.6%	Fair
Pattern adaptation [Bescos04]	[Bescos04]	87%	66%	Good
Statistical shot boundary detector [Hanjalic02]	[Hanjalic02]	79%	83%	Excellent
VT-PixHist [Kobla99]	[Kobla99]	69.7%	62.3%	Fair
Color anglogram [Zhao03]	[Zhao03]	72.6%	88.3%	Excellent
Spatiotemporal blocks [Naci05a]	[Naci05a]	83.6	77.8	Good
	TRECVID	49.4	82.0	

Table 3.2 Rating based on the performance as reported in literature.

The results obtained for [Meng95] indicate the potential of the variance pattern to help discover the dissolves, but also the need to combine this pattern with other information in order to improve the precision [Hanjalic 02]. Motion compensation is also likely to be helpful for the method of [Bescos04], where the recall rate is relatively high but at the same time considerably higher than the precision rate.

For the test results for [Yeo95] we observe varying performance figures from three different implementations. Because the size and the variety of the test material used seem to be similar in all cases, the differences in the reported results are likely a result of different parameter tuning. This might suggest that the method is fragile to the selection of parameter values and/or the small variances in the test material. Similar conclusions could be drawn for the chromatic edit model for dissolve detection proposed in [Song98]. The performance figures reported by the implementations of this model by Cheong [Cheong00] and Kobla et al. [Kobla99] differ considerably. Kobla et al. [Kobla99] suggested that the obtained low scores in their evaluation may be a result of working with DC images instead of full images. Another difference between these two implementations is that Cheong [Cheong00] and Kobla et al. [Kobla99] set different threshold values for the ratio of the derivatives.

The method of [Zhang93] is tested by Cheong [Cheong00], Ankush et al. [Ankush02], Kobla et al. [Kobla99], Lupatini et al. [Lupatini98] and Porter et al. [Porter03]. The results relevant to Twin Comparison method in Table 3.2 also indicate a variance from 60% to 85% in recall and 55% to 85% in precision. But in general, the low recall rate in some tests are compensated by relatively higher precision, which can be seen as a consequence of choosing different operating points and which suggests that this method could provide an equal error rate of 70% to 80% on a general data set.

The results shown in Table 3.2 are taken from Table 2.1 by integrating the performance figures of dissolves and wipes into an overall performance figure for gradual boundary detection. Our shot boundary detection system introduced in Chapter 2 shows a good performance compared to the systems discussed above. Table 2.1 indicates that the recall and precision rates for dissolve detection obtained by our method are around 80%. For the wipes, however the precision of the system drops to around 60%. However, we emphasize again that this result is still encouraging as we are able to detect highly diverse graphical effects without implementing algorithms specifically tailored for different effect types. The main reason for suboptimal performance figures was found in complex motion effects like the one in Figure 2.11 (Chapter 2). Such effects appeared frequently in the TRECVID database, based on which the results in Table 2.4 were obtained. We observed that in order to keep the precision around 80% under these conditions, the recall rate of our system needed to drop considerably. We discuss this and other aspects of the performance of our system in the TRECVID context in more detail in the following sections.

### 3.3.2 TRECVID-based Evaluation

Testing a method or an algorithm under realistic conditions is a necessary step for investigating its true quality and practical applicability. Such testing may reveal various problems that would not come to surface in the standard test environment and if limited and insufficiently diversified test data are used. For shot boundary detection, the TRECVID evaluation benchmark has served as a widely recognized evaluation platform for several years. Although the TRECVID test data set did not show a large diversity of video content (the main content is the news video), it was rich in graphical effects, fast object and camera motions and the number of annotated shot boundaries. All participating systems are evaluated using the same evaluation criteria and the common test set.

Although the set of systems evaluated in TRECVID is not completely identical to the set used for the qualitative analysis in the previous section, the methods discussed before can be said to cover most of the principles underlying the TRECVID-evaluated systems as well. In this sense, the conclusions drawn in Section 3.2 regarding complexity and applicability scope can be applied here as well to interpret the absolute and relative performance of most of the TRECVID-evaluated systems.

We now briefly discuss the most representative systems that participated in TRECVID and received top scores in the shot boundary detection task. The details of all systems can be found in the TRECVID proceedings [TRECVID]. The system from Tsinghua University [Cao06, Yuan07] is one of the best performers over the past years. This system deploys multiple detectors for different gradual transition effects existing in the dataset. It employs motion vector based spatial features and it detects the different types of motion existing in the video in the first place. Then the different fade in / fade out detectors, gradual transition detectors and cut detectors are combined into a unified framework that functions based on a set of collaboration rules for these individual detectors. The system is highly complex and based on many different methods developed for shot boundary detection in the past. All elements in the system are tuned in such a way that, in a specific situation the best performing algorithm can make a decision.

Another representative system evaluated in the TRECVID context is the IBM's *CueVideo System* [Campbell06] which also performed consistently well over a number of years. This algorithm is based on a Finite States Machine (FSM) processing one frame at a time in a single pass over the video. It uses RGB color histograms, localized edge intensity histograms and thumbnails comparisons to evaluate the difference between pairs of frames at 1, 3, 5, 7, 9 and 13 frames apart. Adaptive thresholds are computed using rank filtering on pairs differences statistics in a window of 61 frames around the processed frame. A different threshold is computed for pairs of frames at different time-differences.

The system from National ICT Australia [Zhenghua05] was another successful system in detecting both the abrupt and the gradual shot transitions. We observe

that this system makes use of a combination of machine learning and video analysis methods and is therefore quite computationally expensive.

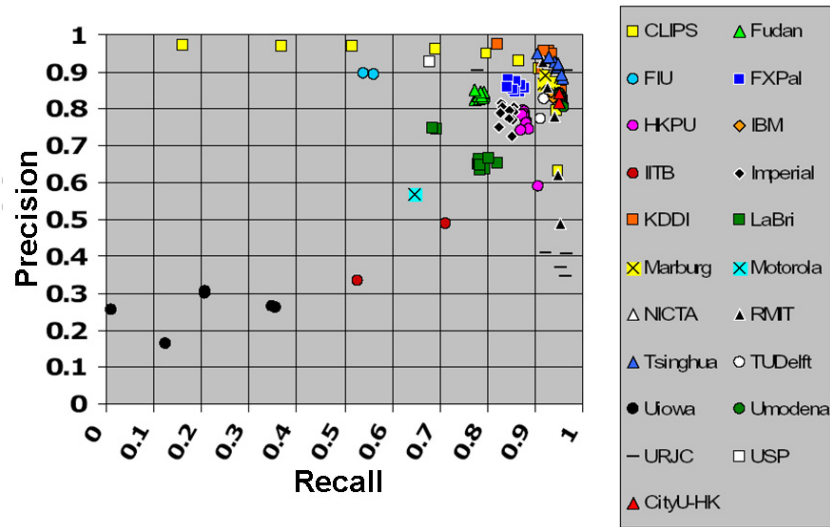
The system from FX Palo Alto Laboratory is another top TRECVID performer. This system explained in [Cooper04] works on channel color histogram features extracted both at the global frame level and from 4x4 pixel blocks of a video frame. An inter-frame similarity matrix is created for each group of consecutive frames of length 5 and 10. The inter-frame similarity features are classified into groups as non-boundary, cut and gradual boundary using a kNN classifier.

When we look at these top performing systems and focus on the gradual transition detection task, we see that they are mostly based on supervised classification. Considering the amount of special graphical effects used in the transitions, it is therefore not a surprise to see that only the methods that develop dedicated classifiers to detect individual effects reach high performance figures. As an example shown in Figure 3.2, *the flying CNN logo* animation that is abundantly used in news videos of CNN, is one of the reasons that our generic system [Naci05a] has performed poorly with respect to its recall rate.

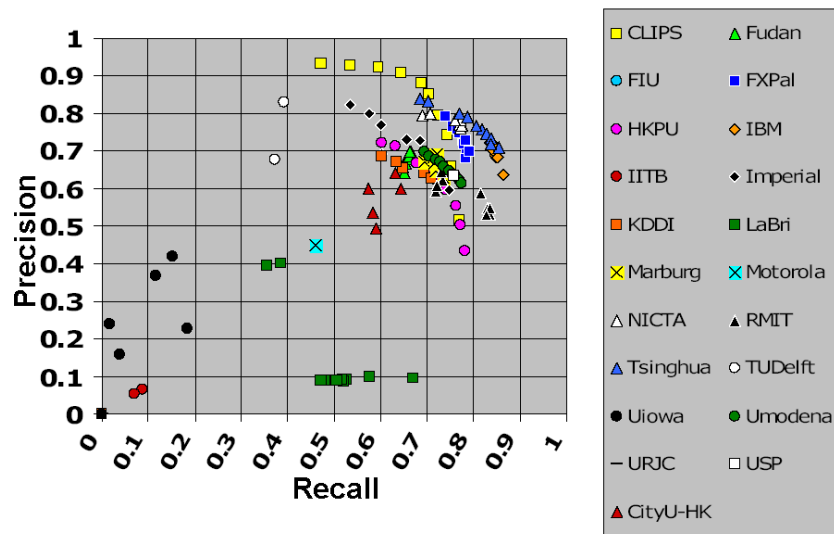
While details of the evaluation results of the shot boundary detection tasks and the links to the publications describing the participating systems can be found in TRECVID reports [Smeaton06], we provide an overview of the recent performances of all systems for both the abrupt and gradual shot detection in Figures 3.3 and 3.4, respectively, to provide insight into the current possibilities to resolve this detection task. The results shown in these figures are from the 2005 edition of TRECVID. Although the shot boundary detection has been tested in two subsequent years as well, we chose to focus on the results of this year because they are the results of our method are also included.



**Figure 3.2** The CNN logo animation, frequently used between the shots in many test videos, is an example for the gradual transitions that can be detected by employing a dedicated transition detector.



**Figure 3.3** The Precision and recall figures for abrupt scene change detection task in TRECVID 2005.



**Figure 3.4** The precision and recall figures for gradual scene change detection task in TRECVID 2005.

### 3.4 Shot Boundary Detection: Problem Solved?

Based on the qualitative and quantitative analysis of the shot boundary detection performance, we now investigate to which extent the shot boundary detection problem could be considered solved. Here we also consider the requirement we impose on a shot boundary detector that is to be employed in a practical MIR system, namely the optimal trade-off regarding the detection performance, applicability scope and computational complexity. In view of this, we are interested in drawing conclusions whether further research would be needed regarding this problem, and if yes, what the objectives of such research should be.

The qualitative analysis showed what range of different shot boundaries could be detected, at what level of transition boundary accuracy and at which computational costs. The quantitative analysis, on the other hand, provided an indication what performance in terms of precision and recall could be expected using the state-of-the-art methods or compound approaches employing several methods in parallel. We may conclude that many individual methods may give good results for specific gradual transition types, since large majority of these use dedicated, separately trained detectors. These methods are especially useful for handling complicated situations where sudden changes in motion or illumination are likely to confuse the detection system. As an alternative to these methods, compound approaches have been proposed, such as [Cao06, Yuan07], built by integrating different methods, each of which meant to address a part of the scope of the shot boundary detection problem. Although such “brute force” detection systems have indeed shown good results in the TRECVID evaluation context, the integration of different methods has been rather ad hoc while still requesting considerable computational resources.

In view of the high computational complexity of the compound methods discussed above, we do not consider them suitable for deployment in practical applications. And indeed, the results reported in Section 3.3 show that methods based on simpler features may also give good detection and identification results, if designed cleverly and combined with discriminative and prior information. Taking this fact into account, we can conclude that building a system based on such an inexpensive method, but comprising specialized functional modules (e.g. taking into account the prior and discriminative information) that are able to cope with more complicated situations whenever necessary may be the key of designing a close-to-perfect system in terms of performance, applicability scope and computational efficiency.

Specifically regarding the reachable detection performance, the results shown in Figure 3.3 indicate that the detection of cuts could be considered a solved problem. Compared to this, and as indicated in Figure 3.4, no existing method is able to handle the gradual shot transition problem at its entirety and at a sufficient level of efficiency and performance. As mentioned before, specific visual effects, a good example of which is shown in Figure 3.2, remain the main source of detection

errors. The complex nature of such effects makes them largely undetectable by the existing methods using generic approaches. The effects of this kind could be addressed only if the shot boundary detection problem is moved to a higher abstraction level and addressed in combination with more complex problems like video scene segmentation and classification [Hanjalic 04]. If moved to this abstraction level, the shot boundary detection problem can be optimized and evaluated per case. This is because the reliable detection performance in critical cases, that is where shot boundaries provide fundamental input in further MCA steps, is much more important than obtaining an average good performance over a large test data set.

Our qualitative and quantitative analysis in this chapter has revealed enough potential for reaching a good average performance both regarding the detection performance but also the trade-off with the computational complexity and applicability scope. Future research efforts in this field should address dedicated optimizations of the detection performance to address specific critical cases in the MCA processing chain.



## Chapter 4

### Content-based Indexing of Live Concert Registrations

In this chapter we address the problem of indexing live concert TV registrations. The challenge lies mainly in the high amount of both signal- and content-level noise, immense diversity of music genres, instruments and vocals involving unclear content structure and structure variations over different registrations. Clearly, the robustness of an indexing mechanism regarding all of these issues becomes critical for a successful automation of the indexing process. We approach this challenge first by introducing a new family of audio features which enables us to detect semantic concepts under strongly varying conditions. Then, we also investigate to which extent the inclusion of the visual information would improve the audio-based indexing process on this type of content.

#### 4.1 Introduction

Making the live concert registrations downloadable on the Internet is becoming increasingly popular. This is visible not only on the sites like *YouTube.com* but also on the emerging commercial providers, like *concert-online.com* and *Fabchannel.com* that offer the content using a video-on-demand model and that attract thousands of users on a daily basis. To provide a better service to on-line users, a non-linear

---

This chapter is based on the following publication:

U. Naci, A. Hanjalic, "Content-Based Indexing Of Music Concert Recordings Based On Crossing-Rate Features", *6th International Workshop on Content-Based Multimedia Indexing (CBMI 2008)*, London, UK, June 2008.

access to concert videos is desired. In this way, in addition to the possibility to download an entire concert registration, the user could also be able to view selected segments, either by searching for particular instrumental or vocal sequences, or simply by obtaining a shorter compilation of a concert containing the most exciting parts. More specifically, the users are likely to be interested in the following selection of content elements for which non-linear access should be provided [Snoek05]:

- Instrument or vocal solos,
- Songs, and other thematic parts of the concert (e.g. also including speaking parts or interaction with the audience), and
- Concert highlights.

While there is considerable previous work addressing some of the issues listed above, like for instance the detection of different instruments in music signals [Deng08, Kitahara07], or segmentation of general audio signals (e.g. speech, music, noise or any combination of these) [Sounders97, Wyse98, Zang01, Lu02, Xu05], hardly any approach was developed for or tested on live concert registrations. The content therein is typically noisy and unstructured due to improvisation, and strongly varying across different concerts. Under these conditions, the robustness and broad applicability of the algorithms for extracting structural and semantic content elements become critical for a successful automation of the indexing process on this type of content. While attempts have been made to expand the widely proposed audio content analysis methods towards more complex systems to improve their robustness and applicability [Zang01, Lu02], it was found that the high accuracies typically reported in literature significantly get deteriorated under the influence of environmental noise and deviations from the pre-defined models [Xu05].

Inspired by the above, we developed a novel indexing approach for live concert registrations that we present in this chapter. In particular, we focus on the extraction of instrument solos and detection of song transitions. Since the high level of noisiness and the lack of structure in this type of content prevent a successful segmentation using the analogies of the approaches known from text and video document segmentation theory [Hanjalic04], we approached this problem by developing a method for detecting applause.

As described in Section 4.2, our approach first focuses on the analysis of the audio signal of the concert registration, for the purpose of which we introduced and evaluated a new family of audio features, the *crossing rate features*. These features proved to be invariant and robust enough to cope effectively with the peculiarities of this type of content. This can be concluded based on a good indexing performance obtained for a diverse data set being a representative subset of the Fabchannel.com online concert collection. Then, we investigate in Section 4.3 to which

extent the inclusion of the visual information would improve the audio-based indexing process in this specific application case. Both the mono- and multi-modal indexing options are evaluated experimentally in Section 4.4. We conclude this chapter by the discussion in Section 4.5.

## 4.2 Audio-based Indexing

Before we introduce the new family of crossing rate features, we first discuss some typical conventional features that have been used frequently for content-based audio indexing purposes in recent literature. For some of these features we also propose an adaptation that proved to be particularly useful when dealing with the difficult use case of live concert registrations. Both the conventional and the new feature set will be used to train supervised semantic concept detectors for instrument solos and applause, the performance of which will be evaluated in a comparative study involving a large collection of live concerts.

### 4.2.1 Feature Extraction: A Common Approach

The number of approaches addressing content-based indexing of live concert recordings is rather limited. One may refer to [Ferguson03] as one of the few representative attempts in this direction. In [Ferguson03], the audio features of spectral flux, Zero Crossing Rate (ZCR), Mel-Frequency Cepstral Coefficients (MFCC) and energy (and in particular the Short Time Energy (STE)) are employed for the detection of various types of audio segments in concert videos. These features, together with Band Energy Ratio (BER), pitch and harmonics-related features are considered most common in the practice of content-based audio analysis [Nielsen07, Eronen01, Cai05, Lu06, Wang00]. In [Nielsen07], the performance evaluation of spectral features led to a conclusion that MFCCs should be preferred for the task of automatic instrument recognition. This conclusion supports the results from an earlier but more extensive work reported in [Eronen01]. In view of this extensive prior knowledge and the abovementioned reviews addressing the usability of audio features for the use cases related to the one addressed in this chapter, we choose to focus on a set of features described in detail in the paragraphs below.

Prior to feature extraction, the analyzed audio signal is first converted into a format characterized by the 44.1 kHz sampling frequency and 16-bit audio sample representation. Furthermore, mono-channel signals are used. As common in audio signal processing, we divide a signal into elementary segments or *audio frames*, from which features are extracted. Contrary to the common practice in speech processing wherein the audio frame length is typically around 30-50 ms, we extracted the features from the frames of 100 ms duration and 50% overlap, which resulted in better performance figures in our use case.

#### 4.2.1.1 Energy-related Features

The Short Time Energy (STE) is a simple but fundamental property of an audio signal and appears in virtually all content-based audio analysis approaches. It represents the magnitude of signal amplitude variations over the duration of an audio frame or, equivalently, the overall spectral power of a frame. Given the audio frame  $n$ , the signal samples  $x(i)$ , and the frame duration  $N$ , the STE value for that frame can be modeled as

$$E(n) = \frac{1}{N} \sum_i x(i)^2 \quad (4.1)$$

To further exploit the energy information, we also introduce the Band Energy Ratio (BER) feature, which can be defined as the fraction of the energy contained in the lower spectral components in the total energy of an audio frame, that is

$$BER(n) = \frac{\int_0^{f_H} |S_n(f)|^2 df}{E(n)} \quad (4.2)$$

Here,  $S_n(f)$  is the short term frequency component of the frame  $n$  and  $f_H$  is the upper bound of the spectral components considered. Despite being a rather simple feature, the effectiveness of BER in content-based audio analysis practice was shown in various existing approaches, such as for instrument detection [Nielsen07, Eronen01] and also for general audio event detection systems [Cai05].

#### 4.2.1.2 Zero Crossing Rate

The Zero Crossing Rate (ZCR) is defined as the number of times the audio signal crosses the zero line. For the audio frame  $n$  the value of this feature can be modeled as

$$Z_0(n) = \frac{1}{2} \sum_i |\text{sgn}(x(i)) - \text{sgn}(x(i-1))| \quad (4.3)$$

where

$$\text{sgn}(x(n)) = \begin{cases} 1, & x(n) \geq 0 \\ -1, & x(n) < 0 \end{cases}$$

The ZCR feature is one of the most referenced audio features and is used in many audio processing applications, including speech processing [Haque06], audio segmentation [Lu02], and audio based event detection [Lu02, Ferguson03].

Since the ZCR values show rapid fluctuations in most applications and use cases, it is a common approach to smooth these values to facilitate further content analysis and indexing steps. A smoothed form of the ZCR feature curve can be obtained using the following expression:

$$\bar{Z}_0(n) = \frac{1}{N} \sum_i Z(i) \quad (4.4)$$

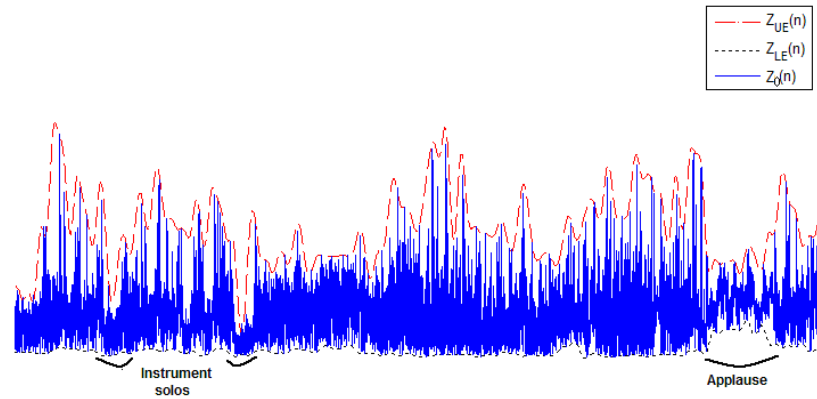
We observed, however, that smoothing of the ZCR curve using a simple filtering method may result in a critical information loss. As an alternative, we introduced in [Naci07a] the idea to work with the upper and lower envelopes of the ZCR function that we denote by  $Z_0^{UE}(n)$  and  $Z_0^{LE}(n)$ , respectively and that can be extracted using the following expressions taking the smoothed ZCR as input:

$$Z_0^{UE}(n) = \max(\{\bar{Z}_0(n-W) \bar{Z}_0(n-W+1) \dots \bar{Z}_0(n)\}) \quad (4.5)$$

$$Z_0^{LE}(n) = \min(\{\bar{Z}_0(n-W) \bar{Z}_0(n-W+1) \dots \bar{Z}_0(n)\}) \quad (4.6)$$

where  $W$  defines the length of the window for min-max calculations.

The effect of using upper and lower envelopes is visible on the example in Figure 4.1. This figure comprises a ZCR curve extracted from a 10 minute segment of a concert video. The upper and lower envelopes extracted using Equations 4.5 and 4.6 are superimposed on the figure as the upper and lower dashed curves, respectively. The characteristic behavior of upper and lower envelopes is clearly visible for certain audio events. For the applause and other noise-like segments in the concerts the lower envelope of the ZCR increases. On the other hand, in the case of instrument solos the values of the upper ZCR envelope become lower than usual. These observations can be explained as follows: An applause or cheering results in a non-harmonic, largely uncorrelated audio signal and continuous and frequent zero crossings. As a result of the lack of harmonics in the signal, the audio data sequence continuously oscillates between negative and positive values and the zero crossing rate is therefore sustained at a higher value and therefore an elevated lower envelope results. On the other hand, along the instrument solo sections one specific frequency value and its harmonics are dominant. This causes relatively less zero crossings and, consequently, lower values of the upper envelope. Due to these properties, the ZCR envelopes have shown the potential to contribute significantly to a robust and efficient analysis of live concert registrations [Naci07b].



**Figure 4.1** ZCR curve ( $Z_0(n)$ ) together with the extracted upper ( $Z_{UE}(n)$ ) and lower ( $Z_{LE}(n)$ ) envelopes for a 10-minute segment of a concert. The specific behavior of the upper and lower envelopes at instrument solos (low upper envelope) and applauds (high lower envelope) is clearly visible.

#### 4.2.1.3 Mel-Frequency Cepstral Coefficients

Mel-Frequency Cepstral Coefficients (MFCC) [Rabiner93] were inspired by the perceptual characteristics of the human auditory system and provide a representation of an audio signal based on the mel-scaled signal spectrum. To extract the MFCC for a given audio frame  $n$ , the logarithms of spectral amplitudes of the frame are mapped onto the perceptual, logarithmic mel-scale using a triangular band-pass filter bank to emphasize the perceptually important frequency components. In the last step, the discrete cosine transform (DCT) is applied to the output of the filter bank to compute the MFCC.

MFCC are among the most important feature sets in the audio field [Mermelstein76], and especially for speech analysis and speaker recognition systems [Naci03]. However, more recently, MFCC also established themselves as a reliable feature set in general (content-based) audio processing tasks [Chung07, Knees07].

#### 4.2.2 Crossing Rate Features

In this section we introduce a new family of features for the content-based audio analysis that reflect various “crossing” properties of audio signals. This family consists of two sets of crossing rate features. The first one, referred to as *Higher Order Crossing Rates* (HO CR), has first been proposed by Kedem and Slud [Kedem81]. Here, higher order crossings correspond to the zero crossing rates of the signal and its derivatives. While the HO CR have been successfully applied in many practical signal processing problems, we demonstrate in this chapter the applicability of this feature set in the context of content-based audio analysis.

Although a satisfactory performance of the HOOCR feature set could be observed in the use case addressed in this chapter, on some parts of our development set the indexing process constantly failed to detect the right semantic concept. We concluded that the main reason behind this failure lies in the fact that although the HOOCR successfully model the distribution of zero crossings in time, they fail to take into account the amplitude (energy) behavior of the signal. This limitation of the HOOCR feature set motivated us to expand the crossing rate feature family by defining the *Band Crossing Rates* (BCR) features, which more explicitly model the changes in signal amplitude. The BCR features correspond to the zero crossing rate of the center-clipped version of a signal at various amplitude levels.

As we will demonstrate later in this chapter, the HOOCR and BCR features, when combined, proved to be powerful enough to cope with difficult properties of the data in live concert registrations. In addition, the fact that they form one feature family, the members of which can all be extracted using the relatively simple counting of signal crossings, makes these features a much more efficient alternative compared to the common set of audio features discussed before. We describe both sets of crossing rate features in more detail in the following subsections.

#### 4.2.2.1 Higher-Order Crossing Rates

Despite the popularity and proven advantages of the ZCR feature in the audio processing practice, the basic ZCR feature as defined before exploits only a fraction of the information that can be extracted from the analysis of the “crossing” behavior of a signal. While the mathematics behind most of the standard audio features, like MFCCs, is well analyzed and this analysis is then used for the theoretical understanding of the behavior of these features under different circumstances and in different use cases (e.g. music, speech, noise), the ZCR has been deployed in a rather heuristic way, and mainly for the purpose for which it has proven to be one of the most powerful features, namely to differentiate between voiced and unvoiced sounds. This deployment of the ZCR was driven mainly by experimental observations of its behavior for different types of signals, but no deeper analysis has been performed yet to explore the full potential of this feature. This can also be seen from another fact, namely that the ZCR has been used as a single-dimensional feature without considering the fact that it can actually be seen as a member of a multidimensional feature family. As we will show in this section, more information can be extracted related to the crossing behavior of a signal, which can lead to a feature family that can be used to create a more reliable basis for feature-based signal representation in view of the goals and objectives approached in this chapter.

Despite the largely ad-hoc applications of ZCR in the audio processing domain, theoretical approaches towards understanding the underlying idea and potential of ZCR are known from other fields. The crossing behavior of the signals belonging to various types of stochastic processes has been investigated in the field

of mathematics since the pioneering work of Rice [Rice45] and Kac [Kac43]. While initially addressing the purely theoretical problem of defining the expected number of zero crossings of a given stochastic process [Kratz06], the practical importance of zero crossings has first been discovered by Kedem and Slud [Kedem81] and applied in the analysis of speech and seismic data. Later on, their usability has been demonstrated in other problems as well, such as time-series data analysis, e.g. 2D data reconstruction [Zakhor90] and characterizing clipping noise in filtered signals [How07].

Focusing again on the audio processing context and considering the fact that the features in this context are generally extracted from nearly-stationary audio frames, an important property of the ZCR feature in this context is that it can be related to the first-order autocorrelation coefficient [Kedem80]:

$$Z_0(n) = \frac{1}{\pi} \cos^{-1}(\rho_1) \quad (4.7)$$

Another important property of ZCR is the *Dominant Frequency Principle* [Higgins80]. According to this principle the ZCR of a signal converges to the crossing rate of the dominant component (if it exists) of the signal spectrum. Based on this principle, Kedem [Kedem86] showed that the spectral distribution of a signal can be perfectly reconstructed from the zero crossings of a signal and its derivatives [Kedem87]. Namely, given the relationship in Equation 4.7, one can show that under the Gaussian assumption, we can write

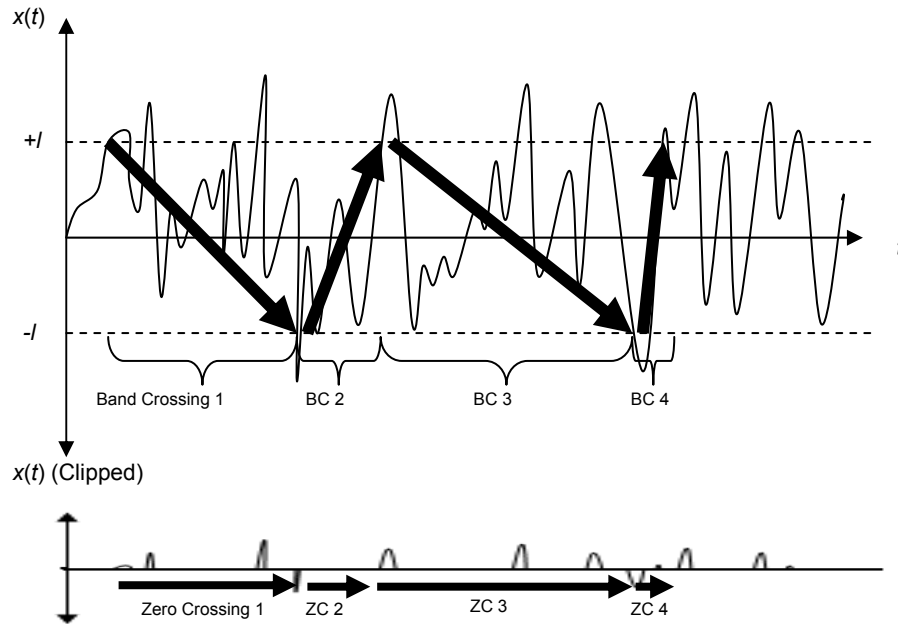
$$Z_i(n) = \frac{1}{\pi} \cos^{-1}\left(\frac{\nabla^{2i} \rho_{i-1}}{\nabla^{2i} \rho_i}\right) \quad (4.8)$$

where  $Z_i(n)$  is the zero-crossing rate of the  $i^{\text{th}}$  derivative of the signal  $X(n)$ ,  $\rho_i$  is the  $i^{\text{th}}$  autocorrelation coefficient and  $\nabla$  is the gradient operator. According to Equation 4.8, one can recursively estimate all autocorrelation coefficients of a signal [Kedem87]. For example, rewriting Equation 4.8 for  $i=1$  gives

$$Z_1(n) = \frac{1}{\pi} \cos^{-1}\left(\frac{-1 + 2\rho_1 - \rho_2}{2(1 - \rho_1)}\right) \quad (4.9)$$

After estimating  $\rho_1$  from Equation 4.7,  $\rho_2$  can be estimated from Equation 4.9. Continuing the recursive operation, one may estimate all the correlation coefficients for the process  $x(n)$ . This suggests that the feature vector  $Z_i(n)$  for  $0 \leq i \leq \infty$  represents all aspects of a signal, except the energy component of the signal. We refer to the feature vectors  $Z_i(n)$  as the HOOCR features. The ability of HOOCR features to represent all temporal aspects of a signal's behavior, combined with the simplicity of extracting these features from a signal, make them an attractive alter-

native to common feature sets used in audio processing field and, in particular, in the field of content-based audio analysis.



**Figure 4.2** The occurrences of *Band Crossings* on a sample signal. Once the signal is center clipped between the levels  $+l$  and  $-l$ , the *Zero Crossings* in the new signal correspond to the *Band Crossings* in the original signal.

#### 4.2.2.2 *Band Crossing Rates*

Although the theory suggests that HOCR features perfectly represent a stationary signal, in practice only a limited number of members of this feature family are extracted and it is not possible to sufficiently cover the signal space and allow a meaningful signal classification. Furthermore, although we assume a stationary signal behavior within an audio frame, this assumption does not always hold perfectly.

To compensate for the above, we introduce the set of *Band Crossing Rates* (BCR), which combine the information about the frequency values and the amplitudes of the dominant frequencies in the signal. We define the *band* as the amplitude region in the normalized signal between the amplitude levels  $-l$  and  $+l$ . A “downwards” *band crossing* occurs when the signal value drops under  $-l$  for the first time after the value was higher than  $+l$ . The following (“upwards”) band crossing

occurs only when the signal value again exceeds the  $+l$  value. In practice, the BCR value corresponds to the zero crossing rate of the centre-clipped version of the original signal, as illustrated in Figure 4.2. The pseudo-code in Figure 4.3 demonstrates the computation of these features.

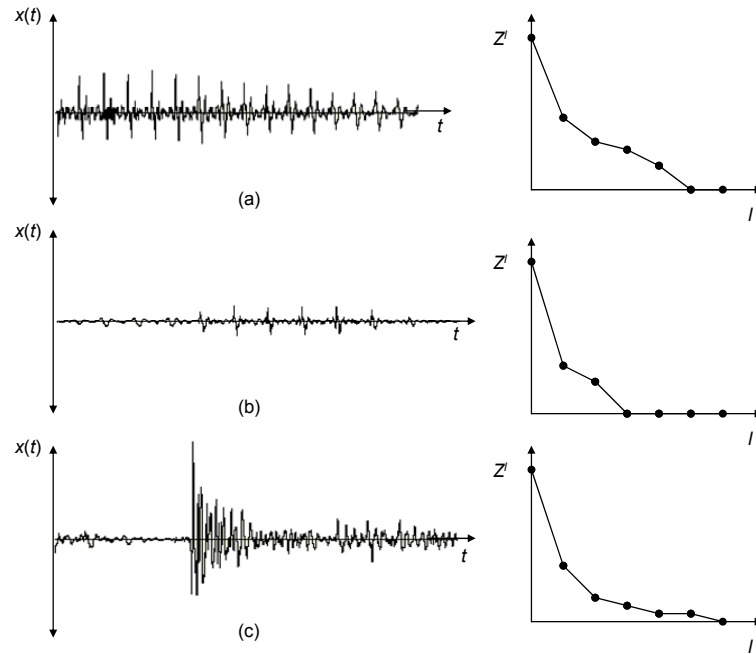
```

FOR  $l=0:\Delta:L$ 
  Flag = 0;
  Count = 0;
  FOR  $t=0:T$ 
    IF  $x(t) > +l$ 
      IF Flag = DOWN
        Count = Count+1;
      END
      Flag = UP;
    END
    ELSE IF  $x(t) < -l$ 
      IF Flag = UP
        Count = Count+1;
      END
      Flag = DOWN;
    END
  END
   $Z^l = \text{Count} / T$ ;
END

```

**Figure 4.3** The algorithm for extracting the BCRs. Initially, when the signal amplitude is over  $+l$  (under  $-l$ ) the flag is set to UP (DOWN). Then every time the amplitude drops under  $-l$  (gets over  $+l$ ) the flag is changed to DOWN (UP). The average number of flag changes corresponds to the BCR at level  $l$ , which we denote with  $Z^l$ .

We illustrate the contribution of the BCR features to the semantic inference process by the examples shown in Figures 4.4a-4.4c. In Figures 4.4a and 4.4b, two signals are depicted showing similar behavior except for the amplitude level. Highly similar crossing rate characteristics of these two signals make the HOCC features alone incapable of discriminating between them, which has caused practical problems, such as the failure to discriminate between applause sections and some silence or other noise-like sections. The BCR features bring additional differentiation between these signals, which helps the indexing process. Another example where the BCR feature positively affected the system performance is depicted in Figure 4.4a-c. While the band crossing rates will remain consistent due to a consistent signal behavior in Figure 4.4a, the signal in Figure 4.4c shows an impulse-like behavior, due to which the BCR for higher amplitude levels will exhibit a sharp decrease over time.



**Figure 4.4** Three examples signals where the BCR features particularly ameliorate the system performance. The HOCR features corresponding to the audio sections in (a) and (b) show similar characteristics although the first data sequence (a) is taken from a music section and the second data sequence (b) is a silent/noisy section. The case in (c) corresponds to signal with an impulse-like behavior, which we observe in certain audio events (e.g. along a drum section). Compared to a consistent signal behavior in (a) (and consistent BCR features), some of these features will drop after a certain band range in (c). The different BCR feature distributions for the three cases can be seen in the diagrams in the second column.

### 4.3 Video-assisted Audio Content Analysis

In the previous sections, we attacked the problem of indexing live concert registrations by solely relying on the audio modality. This approach is motivated by the fact that the events being of interest for indexing (instrument solos, song breaks) are well-defined in the audio domain, and even so well that without considering audio domain information it would be virtually impossible to detect (and annotate) these events with an acceptable level of success. As can be observed from the performance of the system developed in [Snoek07] and using solely visual cues, relying on visual information only is not likely to produce satisfactory results on the

use case of content-based indexing of live concert registrations. However, the information from the visual modality may still support the audio-based decision making mechanism and help improve the indexing reliability in some cases. In order to investigate the possibilities and limits of such support, we explore in this section the challenge to solve the indexing problem formulated in this chapter by a multi-modal approach relying on both the audio and visual modality.

Multimodal approaches integrating the information that is extracted from audio, speech, visual and text modality or various combinations of these, constitute an important research topic in multimedia indexing [Snoek05, Kokaram06, Carmichael08], speech processing [Potamianos03, Papandreou07], and pattern recognition (e.g. face recognition [Bowyer06, Kakadiaris05]). Depending on in which phase the information from different modalities is combined to reach a decision related to the problem considered, the multi-modal techniques can be divided into two major classes [Kittler98, Jain00]:

- *Feature fusion (or early fusion) techniques*, which are based on combining the features extracted from different modalities in one feature vector, possibly after preprocessing the features (i.e. by normalizing or weighting them). The resulting (compound) feature vector is then used to design and deploy a single (multi-modal) decision-making module.
- *Decision fusion (or late fusion) techniques*, where the features extracted from different modalities are first taken separately to design individual decision-making modules per modality. Then, the intermediate decisions made by individual modules are combined together in the second phase to reach the final decision.

Focusing again on music content analysis, as opposed to previous multi-modal approaches (e.g. [Gillet05]) where the audio and visual modalities are treated with equal importance, we recognize the dominance of audio features in our specific use case. We therefore treat the modalities with different importance and employ them in different ways in the indexing process. In particular, we adopt the notion of a *primary modality* as a domain which defines the event to be detected, and a *support modality* that may be correlated to the events characterized in the primary modality, but is not a direct representative of these events. In our approach, the audio modality is the primary modality and we perform the audio-based indexing first. Then, we propose a method that exploits the visual modality to refine the audio-based indexing results.

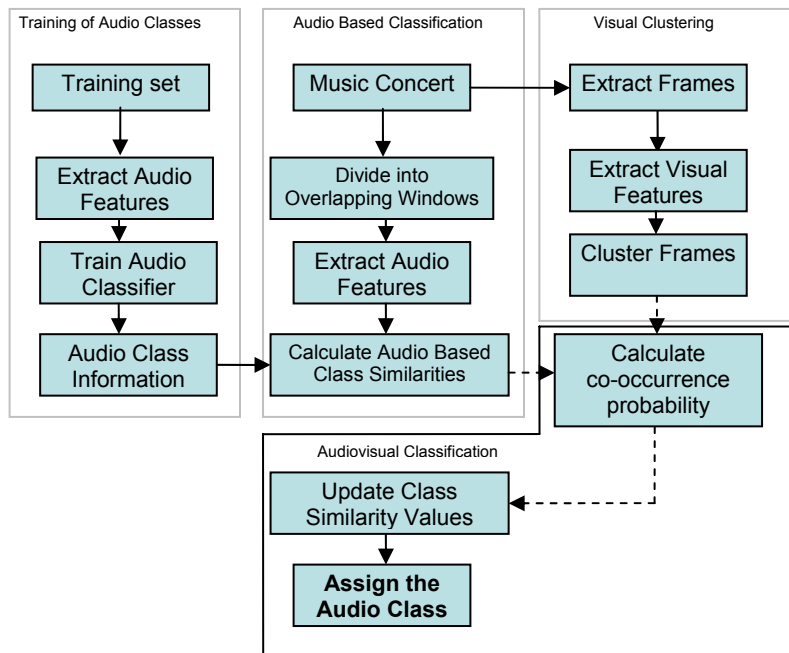
### 4.3.1 Algorithm Design Choices

Certain properties of the support modality impose that this modality should be deployed in a different manner than the primary modality in order to maximize the

indexing effectiveness. A weak, indirect relationship between the support modality and the targeted events in our case makes namely the definition of a reliable feature representation and the collection of a suitable training data set intractable. Therefore, any solution based on supervised classification can be considered impractical. To illustrate this, one may consider the problem of detecting instrument solos in live-concert videos. While one could assume that a zoom-in onto a specific instrument or the close-up of the person playing the instrument is an indication of an instrument solo, our experience shows that these camera operations are neither necessary nor sufficient conditions for detecting this type of events. While a solo is often accompanied by the abovementioned camera operations, it also happens frequently that no camera operation is applied during a solo or that a camera operation is not time-aligned with the solo, for instance in the case where the person performing that solo is zoomed-in after the solo is finished. Also in general, we observed that different directors use different approaches and editing techniques to generate live-concert registrations. While one director prefers to zoom-in on an instrument in the case of the solo performance of that instrument, others may prefer to focus on the face of the artist. This also suggests that each concert registration should be treated separately to limit the influence of the variability in the directing style on the indexing performance.

In view of the above, we developed a multimodal approach to the problem of detection of instrument solos and applause in the concert video recordings, which consists of a supervised classification step in the (primary) audio domain and a supporting unsupervised clustering step in the visual domain. As shown in Figure 4.5 we first apply a clustering algorithm to group the visually similar video frames. The number of clusters is set automatically by the algorithm for each video. In the second step, the system looks for a correlation between the visual clusters extracted from video and the events detected purely from the audio modality. For the classification step in the audio domain we adopt the crossing-rate features described in Section 4.2. Based on the correlation information extracted, the audio classes are updated.

The methods for extracting visual features, creating visual clusters and utilizing them in updating the audio classes are explained in the following sections. The experimental results obtained for the multi-modal indexing approach are given in Section 4.4.3. There, we also compare the multi-modal approach with the audio-only indexing option.



**Figure 4.5** Integration of the information from visual clusters and the classes extracted from audio modality. After the visual clusters are created, the correlation between the audio classes and visual clusters is calculated. The audio class similarity value at time  $t$  is then updated based on the visual class appearing at and around that time point and their correlation with the corresponding audio classes.

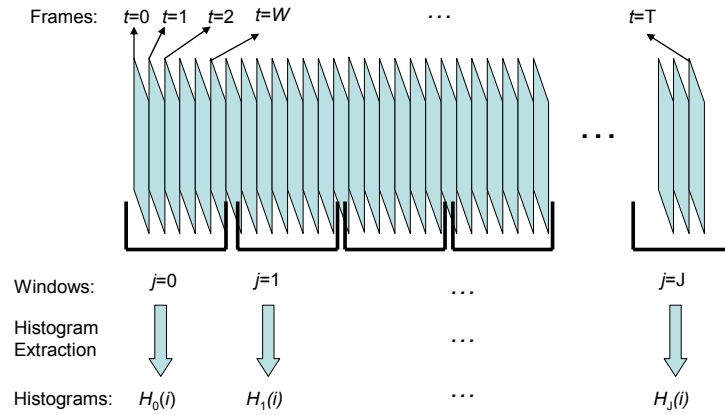
### 4.3.2 Visual Feature Extraction

We start the description of the scheme in Figure 4.5 by defining the visual feature extraction module. Considering a large visual content diversity of live concert registrations and a vague relation between visual features and the content segments of interest, we seek for a clustering method that is as general as possible. Since a music concert takes place on a single location and under relatively constant general conditions throughout the concert (e.g. light show, concert area composition), color histograms can be considered as the features being general enough to model different viewpoints of the concert area. Besides, they do not need to be defined and trained a priori, as compared to more complex alternatives, such as object and event models. In addition, in order to be able to capture the changes in the consecutive frames in time together with the content in individual frames, we have

extracted the color histograms from the groups of consecutive frames, instead of calculating them for each frame separately.

We formalize the histogram extraction process by considering a concert video of  $T$  frames, where each frame is represented by  $F(x,y,t)$ , where  $(x,y)$  are the pixel coordinates in a frame of the dimensions  $X$  and  $Y$ , and where  $t$  is the temporal frame index with  $0 \leq x < X$ ,  $0 \leq y < Y$  and  $0 \leq t < T$ . We represent the data from the red, green and blue color channel as  $F_R(x,y,t)$ ,  $F_G(x,y,t)$  and  $F_B(x,y,t)$ , respectively. As mentioned above, the histogram extraction is not performed on individual frames. The consecutive video frames are first grouped in non-overlapping sequences (windows) of length  $W$ , as depicted in Figure 4.6. If  $J$  denotes the total number of frame groups in the video, then each frame group can be defined as

$$L(j) = \{F(x,y,t) \mid j \cdot W \leq t \leq (j+1) \cdot W - 1\}, \quad 0 \leq j \leq J-1 \quad (4.10)$$



**Figure 4.6** Extraction of the non-overlapping frame groups and their histograms, through which the feature vector is generated that will be used in the clustering phase.

For each color channel and for each group of frames  $L(j)$  we create the normalized color histograms  $H_R(i,j)$ ,  $H_G(i,j)$  and  $H_B(i,j)$ , where  $0 \leq i < H_{\max}$ , and where  $H_{\max}$  is the number of bins in the histogram. This leads to our visual feature vector as the histogram triplet  $H_j(i)$  defined as

$$H_j(i) = [H_R(i,j) \quad H_G(i,j) \quad H_B(i,j)] \quad (4.11)$$

Since the features are extracted from frame groups of length  $W$ , after the clustering step, all frames in the same group will be assigned to the same visual cluster.

### 4.3.3 Creating Visual Content Clusters

The clustering algorithm to be selected for grouping the short windows of consecutive frames needs to satisfy some basic conditions in order to comply with the requirements set in Section 4.3.1. Firstly, the preferred algorithm should not require manual setting of many (critical) parameters. Second, a top-down clustering method is needed that starts from the whole video, detects the perceptually significant groups first and leaves the details and variations in cluster structure to be treated in later stages. In view of these conditions, the *normalized cut* algorithm [Shi00] known from the graph theory and first introduced by Shi and Malik in 1997 appears to be a good choice and therefore we adopt it in the realization of our clustering step. This algorithm, which has been used extensively in image/video partitioning and clustering systems (e.g. [Shi98, Odobez03]) in the past years, aims at using the features to come up with a hierarchy of possible data partitions sequentially and in a top-down fashion, instead of targeting one “correct” partitioning only.

In the first step, an undirected graph  $G=(R, E)$  is constructed. Here,  $R$  is the set of nodes, each of which represents one sample. In our case, this sample corresponds to one frame group  $L(j)$  represented by the feature vector  $H_j(i)$ . Furthermore,  $E$  is the corresponding edge set. Each edge connects two nodes, e.g.  $L(p)$  and  $L(q)$ , and is represented by the weight  $w_{pq}$  which defines the similarity between the nodes. We define the weight as

$$w_{pq} = e^{-\frac{d(L(p),L(q))^2}{\sigma^2}} \quad (4.12)$$

where  $d(L(p),L(q))$  is the Euclidean distance between the nodes with respect their histograms and  $\sigma$  is a scaling parameter. Grouping the data into  $N$  clusters is then equivalent to partitioning the graph  $G=(R, E)$  into  $N$  disjoint sets, which can be done by removing the edges connecting these sets. The decision to separate two sets from each other can be done by investigating the dissimilarity between these sets represented by the weights  $w_{pq}$ .

Taking the case  $N=2$  as example, the dissimilarity between two disjoint sets  $C$  and  $D$  can be measured by computing the total weight of all edges that have been removed, which in graph theory is also referred to as a *cut*:

$$cut(C, D) = \sum_{p \in C, q \in D} w_{p,q} \quad (4.13)$$

While obtaining the optimal bipartitioning of a graph can now be seen as the problem of minimizing the *cut* value, Wu and Leahy [Wu93] showed that this minimum cut approach tends to cut out unnaturally small sets of isolated graph

nodes. The *normalized cut* alternative removes this bias by not only trying to minimize the disassociation between two sets, but also to maximize the association within the sets at the same time [Shi00]. The normalized cut partitioning criterion can be defined as

$$Ncut(C, D) = \frac{cut(C, D)}{assoc(C, R)} + \frac{cut(C, D)}{assoc(D, R)} \quad (4.14)$$

with

$$assoc(C, R) = \sum_{p \in C, q \in R} w_{p,q} \quad (4.15)$$

representing the total connection from nodes in  $C$  to all nodes in the graph, and with  $assoc(D, R)$  defined in the same way for the set  $D$ .

Although the optimization problem based on the partitioning criterion in Equation 4.14 is an NP-hard problem, efficient spectral techniques like the one based on a generalized eigenvalue problem [Shi00] have been proposed to optimize this criterion. Here, eigenvectors and eigenvalues are extracted from the graph similarity matrix  $w = [w_{pq}]$ . Partitioning of the graph  $G = (R, E)$  into  $N$  sets can be either initiated by a graph bipartitioning and then continued recursively, or performed simultaneously for  $N$  sets. We adopted the recursive approach in this chapter. Here, in each step of the recursion, the graph is bipartitioned such that the  $Ncut$  is minimized. Then, a decision is made whether the current partition should be subdivided by checking whether the partitioning is stable and whether a predefined limit for the cluster number is exceeded. The stability of a partition is investigated by observing the degree of smoothness in the eigenvector values [Shi00]. As the criterion for eigenvector stability, we used the *eigengap* measure. The eigengap of a matrix  $w$  is defined as

$$\delta(w) = 1 - \frac{\lambda_2}{\lambda_1} \quad (4.16)$$

where  $\lambda_1$  and  $\lambda_2$  are the two largest eigenvalues of  $w$ . This measure is proved to be related to the *tightness* of clusters [Ng01, Vempala00] and successfully employed in spectral clustering algorithms for determining the number of clusters [Odobez03]. We used the threshold value 0.15 for  $\delta(w)$  to stop further clustering. Although this method works well in most of the cases, the eigengap method considers only the inter-cluster information and if some clusters are not clearly defined, there is the risk that the clustering will not be stopped until the cluster sizes are considerably reduced and, consequently, a very high number of clusters are created. For this reason, we also limited the maximum number of clusters to 12 to control the cluster-granularity.

#### 4.3.4 Updating Audio Content Classes Using Visual Clusters

In the previous sections we first described the procedure, using which the audio frames extracted from the soundtracks of live concert registrations are assigned class memberships with respect to a set of predefined semantic concepts. Then, we introduced an approach for grouping series of consecutive video frames of a live concert registration into clusters based on their general visual similarity. In this section we define a method for updating the classification results obtained in the audio domain using the information about visual clusters.

Since the video frame groups and the audio frames are not necessarily aligned, we consider from now on a single video frame that can be represented by the pair  $(a_t, v_t)$ . Here,  $a_t$  and  $v_t$  stand for the audio class and the visual cluster, respectively, that the frame is assigned to using the audio classification and visual clustering steps. The class updating procedure can now be defined as the mapping

$$(a_t, v_t) \Rightarrow \hat{a}_t \quad (4.17)$$

where  $\hat{a}_t$  is the definitive class assignment for the video frame considered. If we assume that there are  $S$  audio classes  $A_i$  and  $N$  visual clusters  $V_n$ , then the value and index ranges of  $a_t$  and  $v_t$  can be defined as

$$a_t = A_i, \quad v_t = V_n, \quad 0 \leq n < N, \quad 0 \leq i < S, \quad 0 \leq t < T$$

where  $T$  is the length of the concert video in frames.

Using the audio class probability  $p(a_t = A_i)$  obtained in the audio classification step as a prior, we now search for the probability of having  $\hat{a}_t = A_i$ . This probability can be expressed as the posterior probability of  $a_t = A_i$  taking into account the prior and conditioned by the visual cluster membership  $v_t$ . Assuming that the visual cluster at time  $t$  is  $V_n$  (i.e.  $v_t = V_n$ ), this conditional probability can be formulated as

$$p(\hat{a}_t = A_i) = p(a_t = A_i | v_t = V_n) = \frac{p(v_t = V_n | a_t = A_i) \cdot p(a_t = A_i)}{p(v_t = V_n)} \quad (4.18)$$

where,  $p(v_t = V_n) = p(V_n)$  and,

$$p(v_t = V_n | a_t = A_i) = p(V_n | A_i) \quad (4.19)$$

can be computed using the following expression:

$$p(V_n | A_i) = \frac{\sum_{t=0}^{T-1} EQV \{ (v_t = V_n) \wedge (a_t = A_i) \}}{\sum_{t=0}^{T-1} EQV \{ a_t = A_i \}} \quad (4.20)$$

here EQV is a binary Boolean operator giving the value 1 when its argument is true and 0 otherwise.

The visual cluster probabilities in Equation 4.18 can also be calculated from the statistics of the video sequence as

$$p(v_n) = \frac{\sum_{t=0}^{T-1} EQV \{ v_t = V_n \}}{T} \quad (4.21)$$

Finally, the audio class at time  $t$  taking account the analysis in both the audio and visual domain can be inferred as the one maximizing the conditional probability in Equation 4.18, i.e.:

$$\hat{a}_t = \arg \max_{a_t} (p(a_t | v_n)), \quad a_t \in \{A_1, \dots, A_S\} \quad (4.22)$$

We now demonstrate our approach on a simple example illustrated in Figure 4.7a–4.7c. In the first step, we represent a frame sequence as depicted in Figure 4.7a using the audio class and visual cluster labels per frame. We create a hypothetical sequence of initial audio classes and visual clusters, which, we assume, are assigned to each frame separately using the mechanisms explained above. In this example we assume that there exist two audio classes ( $A_0, A_1$ ) and five visual clusters ( $V_0, \dots, V_5$ ). In the second step, the result of which is depicted in Figure 4.7b, we calculate the conditional probabilities of visual clusters given an audio class using the statistics from the sequence in Figure 4.7a and Equation 4.20. Finally, in the last step, we update the audio class probabilities extracted from audio modality alone (given in rows 2 and 3 of Figure 4.7c) using the correlation information from visual clusters and Equation 4.21. As we can observe from the results reported in rows 6 and 7 of Figure 4.7c, in our example three of the assigned classes changed ( $a_3, a_4$  and  $a_8$ ).

The approach proposed above allows us to utilize the indirect information from the support modality for each video individually and is robust against the likely differences in the editing techniques, lighting conditions and other possible variations that can be applied by different directors and in different concert settings. We observed that while including visual information did not affect the audio-based result in a negative way, the visual information proved to be useful especially for the borderline cases, where at a certain time point  $t$ , the audio based class probabilities are close to each other and, thus, the confidence of the audio-based decision module is low. In this sense, the effect of the visual support on the overall system performance was perceived as positive, as can also be seen from the results in Section 4.4.3.

$T=0-19$   $A_0 A_0 A_0 A_1 A_1 A_1 A_1 A_1 A_0 A_1 A_1 A_1 A_1 A_1 A_1 A_1 A_0 A_0 A_0 A_0$   
 $V_0 V_2 V_2 V_2 V_2 V_5 V_5 V_0 V_5 V_5 V_5 V_3 V_0 V_5 V_5 V_5 V_2 V_2 V_3 V_2$

$T=20-39$   $A_0 A_0 A_1 A_1 A_1 A_0 A_0 A_0 A_0 A_0 A_1 A_1 A_1 A_1 A_1 A_1 A_1 A_0 A_1 A_1$   
 $V_2 V_2 V_0 V_5 V_5 V_3 V_2 V_2 V_2 V_3 V_0 V_3 V_5 V_5 V_5 V_3 V_5 V_2 V_1 V_0$

$T=40-59$   $A_1 A_1 A_0 A_0 A_0 A_0 A_0 A_0 A_0 A_0 A_0 A_0 A_0 A_1 A_1 A_1 A_1 A_1 A_1 A_1$   
 $V_5 V_5 V_0 V_2 V_2 V_2 V_4 V_4 V_0 V_0 V_2 V_2 V_2 V_2 V_4 V_0 V_5 V_5 V_5 V_4$

$T=60-79$   $A_1 A_1 A_0 A_0 A_0 A_1 A_1 A_1 A_1 A_1 A_0 A_0 A_0 A_0 A_0 A_0 A_0 A_0 A_0 A_0$   
 $V_5 V_5 V_4 V_2 V_1 V_0 V_5 V_5 V_5 V_4 V_0 V_2 V_1 V_2 V_2 V_2 V_2 V_4 V_4 V_2$

$T=80-99$   $A_0 A_0 A_1 A_1 A_1 A_1 A_1 A_1 A_1 A_1 A_0 A_0 A_0 A_0 A_0 A_0 A_1 A_1 A_1 A_1$   
 $V_4 V_2 V_0 V_3 V_3 V_2 V_5 V_5 V_5 V_1 V_0 V_2 V_2 V_2 V_3 V_2 V_2 V_5 V_0 V_5$

(a) A sample series of audio and visual clusters.

$p(V_j   A_i)$	$V_0$	$V_1$	$V_2$	$V_3$	$V_4$	$V_5$
$A_0$	6/48	2/48	29/48	4/48	6/48	1/48
$A_1$	9/52	2/52	5/52	5/52	3/52	28/52

(b) The conditional probabilities of observing the visual clusters for the sequence in (a) given an audio class.

$a_t$	$\Rightarrow$	$a_0$	$a_1$	$a_2$	$a_3$	$a_4$	$a_5$	$a_6$	$a_7$	$a_8$	$a_9$
$P(a_t = A_0)$	$\Rightarrow$	0.82	0.73	0.53	0.41	0.38	0.21	0.27	0.16	0.55	0.46
$P(a_t = A_1)$	$\Rightarrow$	0.18	0.27	0.47	0.59	0.62	0.79	0.73	0.84	0.45	0.54
Detected class based on audio modality	$\Rightarrow$	$A_0$	$A_0$	$A_0$	$A_1$	$A_1$	$A_1$	$A_1$	$A_1$	$A_0$	$A_1$
Corresponding Visual Class	$\Rightarrow$	$V_0$	$V_2$	$V_2$	$V_2$	$V_2$	$V_5$	$V_5$	$V_0$	$V_5$	$V_5$
$p(a_t = A_0   v_t)$		0.75	0.96	0.88	0.79	0.83	0.01	0.02	0.01	0.11	0.08
$p(a_t = A_1   v_t)$	$\Rightarrow$	0.25	0.04	0.12	0.21	0.17	0.99	0.98	0.99	0.89	0.92
Detected class based on audio and visual modalities	$\Rightarrow$	$A_0$	$A_0$	$A_0$	$A_0$	$A_0$	$A_1$	$A_1$	$A_1$	$A_1$	$A_1$

(c) The updated audio classes.

Figure 4.7 The update procedure of the audio classes using visual clusters in three steps.

## 4.4 Experimental Evaluation

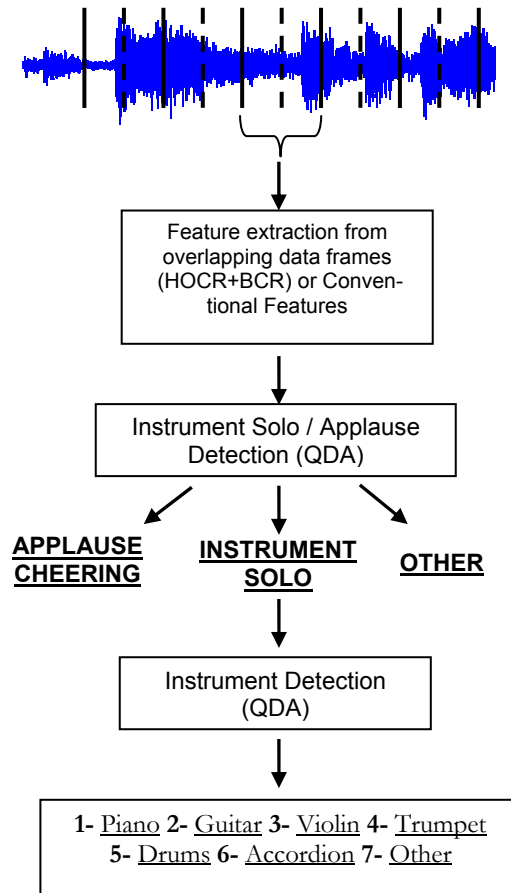
The evaluation of the indexing schemes presented in this chapter is performed on a concert database provided by an online concert provider. The database is composed of live recorded rock, pop and rap music concerts in different concert halls. We used approximately 10 hours of annotated audio data for evaluating the indexing performance. In this data set, instrument solos cover 13.2% and applause or cheerings take 7.8% of the material. 77.9% of the test data belongs to the “non-solo music” segment. The rest (1.1%) corresponds to silent parts, either at the beginning or at the end of some of the concerts, or between the songs in case of long breaks. We refer to the non-solo music and the silence segments as “other” class. 20% of the data is used for training the system and the tests are performed on the remaining data. We performed multiple tests for cross validation.

To evaluate the system performance, the manually annotated test material is compared to the segments indexed automatically. To quantify the test results, we created a confusion matrix summarizing the results for each experimental case. Each entry in the matrix corresponds to the overlap between the annotated and indexed segments mentioned in the corresponding row and column of the matrix, respectively, as depicted in Tables 4.1-4.2.

In Section 4.4.1, we first present the experimental setup we will use to evaluate the indexing scheme operating in the audio domain using the new crossing-rates feature family introduced in Section 4.2. Then, we evaluate separately two types of the results of audio-based indexing, namely the identification of instrument solos and applauses (in Section 4.4.2) and the identification of different instrument classes (in Section 4.4.3). Finally, in Section 4.4.4, we investigate the possibility to improve the audio-based indexing results by considering the visual modality using the method presented in Section 4.3.

### 4.4.1 Audio-based Indexing; Evaluation Setup

To evaluate the quality of the introduced crossing-rates features regarding the indexing problem posed in this chapter, we performed a comparative study involving two systems. The first one is based on the HOCR and BCR features and the other one, the baseline, uses the state-of-the-art combination of traditional features we described in Section 4.2.1. While we adopt the common definitions of BER, MFCC and STE, we employ the ZCR in a novel fashion, as introduced in Section 4.2.1.2, in order to maximize the performance of the reference system and so to maximize the validity of our comparative analysis. For the extraction of the HOCR features, first eight coefficients in Equation 4.8 are included in the feature vector as suggested in [Kedem86].



**Figure 4.8** Overview of the Instrument Solo / Applause and Instrument Detection systems. Features are extracted from 100 ms audio frames with 50 ms overlap. In both cases the decision is made per audio frame and based on majority voting over 5 second windows.

For both systems, the same general approach is used to classify audio frames into semantic concepts. This approach is illustrated schematically in Figure 4.8. After extracting the features, indexing is performed in two steps. In the first step, each audio frame is classified either as Applause, Instrument Solo or Other. In the

second step the frame is further assigned to one of the pre-defined instrument classes. The first classification step is particularly suitable to provide insight into how good each feature set performs on representing perceptually highly scattered data. In contrast to this, the instrument classes in the second classification are better defined and homogeneous. For both classification steps the systems were trained and tested using the same dataset and methodology described above. The final results are obtained after applying a majority voting filter to the indexing output of each step in order to filter out the remaining, short, falsely detected segments. The majority voting filter is applied on series of audio frames in the total duration of five seconds. The features are derived from 100 ms windows with an overlap of 50 ms, so each series comprises 100 windows.

For the classification purposes we used the *Quadratic Discriminant Analysis* (QDA) method [Srivastava07]. QDA is one of the popular and standard tools in supervised classification systems. It has been widely used in audio event detection systems [Agostini03] and its efficiency has been proved for similar applications in recent literature [Weihs07, Morchen06, Starzacher08].

#### 4.4.2 Instrument Solo / Applause Detection

The confusion matrices containing the results of the two audio-based indexing systems are given in Tables 4.1 and 4.2. From Table 4.2 we observe that the performance of the proposed system based on the crossing-rates features is satisfactory, except for a number of missed instrument solo parts. The main reason for this was found in the presence of solos performed by percussion instruments (drums), which are difficult to treat in the same way as other “regular” instruments. However, the detection rate for other instrument solos was shown to be satisfactory.

When compared to the results of an optimized system utilizing the conventional set of audio features in Table 4.1, the proposed system shows an improvement in the detection rates. A detailed analysis of the results showed that ZCR Envelope and BER are the most critical features in the conventional system, which confirmed the assumption made in Section 4.2 and based on the analysis of Figure 4.1. On the other hand, relying only on ZCR is not sufficient to create a solid base for an indexing decision. Although, intuitively, the ZCR envelope should do a good job in differentiating between the applause/cheering segments and instrument solos, in some cases it fails to correctly classify these events. As an example, the sound with a rather high pitch created by cymbals, if dominant, is likely to mislead this feature and confuse it with the sound of applause/cheering. We observed that the BER feature is able to help dealing with such problems related to the high-pitch instruments dominating the music.

		Applause/ Cheering	Instrument Solo	Other
DETECTION	Applause/ Cheering	<b>93.2%</b>	1.2%	5.6%
	Instrument Solo	3.0%	<b>67.8%</b>	29.2%
	Other	9.9%	2.6%	<b>87.5%</b>

**Table 4.1** Confusion matrix for solo/applause detection using standard audio features.

		Applause/ Cheering	Instrument Solo	Other
DETECTION	Applause/ Cheering	<b>91.2%</b>	0.9%	7.9%
	Instrument Solo	2.3%	<b>77.4%</b>	20.3%
	Other	7.8%	3.9%	<b>88.3%</b>

**Table 4.2** Confusion matrix for solo/applause detection using HOCC and BCR features.

#### 4.4.3 Instrument Identification

In the instrument identification part, we measured the performance of the crossing rates features when differentiating among the music instruments playing in the concerts. For this purpose, we annotated the parts in our dataset where a single instrument is playing in solo or dominating all other instruments. In order not to limit ourselves to a short list of instruments and become biased towards a limited number of concert registrations, we randomly scanned the concerts in our entire collection and selected and annotated the single (or dominating single) instrument parts. We found six instruments that were sufficiently present in the collection in a solo context and we therefore focused on their detection in our experimental evaluation. These instruments include piano, guitar, violin, trumpet, drums and accordion. The annotated instrument sections are randomly assigned to training

and testing sets and the reported results were obtained by taking the average of multiple cross-validation tests. We also added the “*other*” class to the training and test sets taking into account the concert sections that cannot be related to a single instrument in order to evaluate the performance of detecting desired classes against unclassified random data.

The results obtained using our proposed system are shown in Table 4.4. Just as in Table 4.1 and Table 4.2, each row corresponds to one of the target instruments and each column corresponds to an instrument detected by our method. As an example, in the row corresponding to instrument 2 (Guitar), we observe that there are 442 sections in the test set annotated as guitar. 373 of these sections are detected accurately as guitar. However, 60 of these sections are recognized as Piano (instrument 1), two of them as Violin (instrument 3) and seven of them as Trumpet (instrument 4). Compared to the results in Table 4.3 that were obtained using the common audio features, Table 4.4 suggests that the crossing rate features have potential in modeling the instruments with high accuracy despite noisy data and a relatively high diversity of instrument solo realizations, even by the same class of instruments, over different parts of the collection.

The common audio features have failed to create effective class models on our dataset. The main reason for this is that the features like pitch, ZCR or BER give clues about general properties of the audio data but fail to classify the events that are not directly related to any of these general properties, particularly in the difficult use case considered in this chapter. Furthermore, the MFCCs, which proved to be an important feature set in audio content analysis, require a time series modeling before classification. MFCCs can be expected to give good results when used together with a proper time series model like Hidden Markov Model (HMM). The main difficulty here is that MFCCs are rather fragile in the presence of noise and insufficient training data, like in our use case.

Moreover, training a HMM requires large amounts of data to model a class properly. On a closed-set problem like speech recognition where one can limit the classes to a finite number of phonemes to be detected, this is a reasonable requirement. However, instrument detection in concert videos is an open-set problem and for finding a reliable solution fast converging and adaptable training methods are required. As shown by the experimental results, the crossing-rate features can be considered a reasonable alternative to common features for open-set problems and difficult data sets.

	1	2	3	4	5	6	7
1	<b>304</b>	5	4	35	31	0	19
2	290	<b>36</b>	0	15	78	0	53
3	76	9	<b>1</b>	8	3	0	59
4	24	0	1	<b>28</b>	0	1	56
5	8	2	0	0	<b>31</b>	0	5
6	22	0	0	4	0	<b>12</b>	7
7	85	30	11	24	59	3	<b>1405</b>

(1- Piano 2- Guitar 3- Violin 4- Trumpet 5- Drums 6- Accordion 7- Other)

**Table 4.3** Confusion matrix for instrument detection using the standard audio features.

	1	2	3	4	5	6	7
1	<b>297</b>	91	7	0	0	0	4
2	60	<b>373</b>	2	7	0	0	22
3	22	25	<b>104</b>	0	0	0	4
4	0	7	2	<b>100</b>	0	1	0
5	0	0	0	0	<b>42</b>	0	2
6	1	3	0	2	0	<b>39</b>	2
7	87	19	4	0	6	0	<b>1501</b>

(1- Piano 2- Guitar 3- Violin 4- Trumpet 5- Drums 6- Accordion 7- Other)

**Table 4.4** Confusion matrix for instrument detection using the HOCC and BCR features.

#### 4.4.4 Visually-assisted Indexing

In order to investigate the benefit of including the visual modality into the indexing process, we evaluated the method proposed in Section 4.3 on the same dataset as

the one used in Sections 4.4.2 and 4.4.3. We analyzed the effect of visual support for both audio classification schemes, i.e. based on both the conventional and the crossing-rates features.

		Applause/ Cheering	Instrument Solo	Other
DETECTION	Applause/ Cheering	<b>95.3%</b>	2.5%	2.2%
	Instrument Solo	4.1%	<b>71.8%</b>	24.1%
	Other	10.3%	4.6%	<b>85.1%</b>

**Table 4.5** Confusion matrix for solo/applause detection results based on audio classes using standard audio features and visual clustering.

		Applause/ Cheering	Instrument Solo	Other
DETECTION	Applause/ Cheering	<b>95.8%</b>	2.3%	1.9%
	Instrument Solo	3.2%	<b>78.8%</b>	18.0%
	Other	9.1%	4.0%	<b>86.9%</b>

**Table 4.6** Confusion matrix for solo/applause detection results based on audio classes using HOCC and BCR features and visual clustering.

A comparison between Table 4.5 and Table 4.1, and between Table 4.6 and Table 4.2 shows a small but consistent improvement in both the applause and instrument detection results. Although the visual modality seems to be able to affect the already good results of audio based indexing only marginally, our proposed conservative use of the support modality has proven to give better results than any conventional indexing method based on supervised training of semantic

concept detectors in visual domain [Snoek05]. The main reason for a modest contribution of the visual modality can be found in a high visual diversity of instrument solo realizations, which makes the creation of meaningful visual clusters a difficult task in this use case. However, our approach proved to be able to use the limited relevant information extracted from the visual stream in an effective manner so the results of audio-based analysis have not degraded, but improved wherever possible.

## 4.5 Conclusions

In this chapter, we introduced a new approach for automatic audio-based content indexing using essentially the audio modality and also exploiting the visual modality in a primary/support modality framework. We chose the live concert records as the target data type since they constitute a good base for testing the robustness of indexing mechanisms in a realistic use case. The objective of the research reported in this chapter was twofold.

As the first objective, we wanted to evaluate the effectiveness of crossing-rate features in audio indexing. The crossing rates features are well studied in mathematics. These studies revealed many interesting connections between stochastic processes and their crossing rate features. We observed that these findings may make the crossing rate features a promising tool in audio indexing as well, which was also confirmed by the experimental results. These results suggest that these features have the capacity to generalize enough to address the high diversity and noisiness of the content in the considered use case. On a variety of cases from detecting instrument solos in general or applause and cheering from the audience to indexing individual instruments, the proposed feature set performed robustly, and consistently better than a state-of-the-art system using a common audio feature set that we devised as a baseline. Besides, the simplicity of extraction makes the proposed crossing-rate features certainly a more efficient alternative to classical features employed in the indexing context.

As a second objective, we investigated the potential of the visual modality to improve the audio-based indexing performance. The results obtained by comparing the multi-modal indexing approach with a mono-modal (audio-based) one confirmed the assumption that a high diversity of the visual content related to a particular semantic concept does not allow a significant positive contribution of the visual modality to the overall performance. As a matter of fact, “blindly” relying on the information from the visual channel when training semantic concept detectors may even prove counter-productive, even leading to a degradation of the results obtained in the audio domain. The problem of indexing live concert registrations is therefore another good example emphasizing the need for carefully selecting the modalities to work with and the way how they are employed in a

---

given MIR application and use case. Our approach treating the visual modality in a conservative fashion - as a support modality only - proved to be effective in extracting useful information from the visual channel and utilizing it in the way to still produce modest but consistent improvement compared to the mono-modal case.



## Chapter 5

### TRECVID BBC Rushes Summarization

This chapter presents the results of our participation in the TRECVID evaluation benchmark 2008. We focus here on the framework we developed to automatically create summaries of raw footage referred to as “BBC Rushes”. Because of the subjective nature of the video summarization problem, the common evaluation frameworks like TRECVID, characterized by clearly defined summarization requirements, evaluation criteria and representative data sets, are very important for obtaining a realistic insight into the reliability of the generated summaries. The presented BBC Rushes summarization framework was developed through a joint effort within the EU COST292 Semantic Multimodal Analysis of Digital Media, where a number of new and existing MIR technologies were integrated in an innovative fashion in the attempt to solve the corresponding TRECVID task.

#### 5.1 Introduction

The problem of condensing a long video document into a compact but comprehensive summary has been one of the major challenges in the MIR field. The need for such condensed video content representations range from professional applica-

---

This chapter is based on the following publication:

U. Naci, U. Damjanovic, C. Kaes, B. Mansencal, M. Corvaglia: The COST292 experimental framework for RUSHES task in TRECVID 2008, *Proceedings of the ACM Multimedia 2008 Conference*, TRECVID/BBC Rushes Summarization Workshop 2008.

tions involving the editors that seek a quick insight into enormous amounts of raw video footage, via automated generation of movie trailers for enhancing the Electronic Program Guides (EPG) in personal video recorders (PVR), to smart delivery of selected video excerpts like sport highlights or breaking news on mobile devices.

Considerable amount of previous work can be found in recent literature where different approaches to video summarization were proposed [Truong07, Barbieri07]. As already introduced in Chapter 1, video summarization is a general term used to describe any method or technique aiming to present all important aspects of the video content to the user in a compact fashion. This representation should be realized in a way that it is not longer than required in a given use case (e.g. 10% of the original video document length), but at the same time, maximally informative of the content of the original video document. In other words, a user should be able to obtain a full impression about the content of the video document, including all elements of the story and the story line itself, by watching the summary alone.

Two general types of video summaries can be distinguished, namely the *static* and the *dynamic* one. Static summaries of a general video are typically obtained by representing the video content using a limited number of carefully selected video frames, or *keyframes*. Different methods for keyframe extraction and keyframe-based video representation have been proposed in literature. Many video streaming services (e.g. YouTube) represent the content of their videos with a single frame, which can be considered as the simplest form of keyframe extraction. Another simple way of extracting keyframes is uniform sampling, that is, extracting video frames with a given fixed time interval in-between. Uniform sampling does not take the content information in the video into account and therefore does not always produce meaningful and compact summaries. This approach may, for instance, make a long and static scene be represented by multiple keyframes of the same content, while some other more dynamic scenes may become underrepresented if the frame extraction interval is too coarse in view of the content dynamics.

In the search for better options, methods were proposed to select the keyframes reducing the redundancy in the visual content in the overall keyframe set. In these methods, a new keyframe is created if the visual content is sufficiently different than the last extracted keyframe. Color histograms were typically used as the frame features in this approach, as proposed in [Yeung95, Zhang97 and Kang99]. In addition, image block displacements [Zhang03] and geometric properties of the extracted video objects [Kim02] were often used as the features. A further class of techniques with similar objectives examines the motion intensity in the video and extracts keyframes at segments of sufficient changes in the visual content flow [Xiao06]. A disadvantage of the approaches described above is that they are not able to handle the scenes repeated along the video. To deal with this problem, frame clustering based methods were proposed. Different clustering techniques, like sequential clustering in [Zhuang98], hierarchical clustering in

[Girgensohn99] or fuzzy c-means clustering in [Yu04], can be employed to group video frames, after which most relevant clusters are selected and each of them is represented by a keyframe. The main challenge in this type of methods is how to define a clustering criterion that best reveals the grouping phenomena in the visual content flow of a video. To avoid dealing with this challenge, further classes of keyframe extraction methods were proposed, like those modeling the “interestingness” of the visual content and then selecting keyframes at the peak values of the interestingness time curve [Liu03], [Dufaux00].

The keyframes extracted using any of the methods described above can be used to present the video content to the user in a compact manner, e.g. using *video poster* [Yeung97], but also to facilitate video browsing. Examples of browsing solutions proposed in the past are a storyboard display [Komlodi98, Taskiran06], which is one of the most straightforward yet efficient ways of presenting the video content to the user. To further improve the video viewing and browsing experience, hierarchical keyframe representation is proposed [Sull01], wherein first a whole video is represented by a single keyframe and then further keyframes are extracted down in the tree structure.

Independent of the technique used for keyframe extraction, static video summaries suffer from the fact that two fundamental sources of information about the content flow in a video, namely the motion and the sound, are not considered in the summary. For this reason, considerable effort has been invested in developing theories and methods for *dynamic video summary* generation. This work, also referred to as *video skimming*, is about selecting certain dynamic video segments that are representative of the entire video content and concatenating them into a short version of the video – a video skim. Information from different modalities can be used for this purpose. Some systems analyze the visual scene and its dynamics, transform this analysis into a complexity measure that is used to remove the redundant segments and keep the essence of the video document [Sundaram02]. Regarding the audio domain, [Wactlar96] used automatic speech recognition transcripts to guide the summary generation process. Multimodal summarization systems have been developed as well, like the one in [Ma02] which relies on the audio, visual and text modality.

In general, the process of video skimming consists of three main steps. The first step is the segmentation of the video into short coherent sections. While these sections mostly correspond to shots, other segmentation approaches have been investigated as well. Some of them even proved to be more advantageous than shots in certain applications and use cases. For instance, [Taskiran01] used pauses in the speech as the segmentation criterion. [Peyrard03], on the other hand, used the abrupt changes in the motion field. Higher-level (semantic) criteria have been employed for this purpose as well. [Ariki03], for example, partitioned the video into event and non-event segments.

The segmentation step is followed by a selection process, where a decision is made on which segments or their parts are suitable to be considered for the sum-

mary. In addition to the complexity measure used in [Sundaram02] to select a minimum-length sequence being suitable for a summary (explanation given in Chapter 1), a popular method to do this is to perform clustering on the segments to group the similar segments together. By reducing a cluster representation to a single segment from the cluster and using the segments selected this way to form a summary, the redundancy in the final summary can be minimized [Gong03]. As a more complex alternative to clustering, video can be indexed first in terms of dynamic semantic concepts (events) and then the segments corresponding to events can be selected in the summary [Ariki03]. Similarly, [Babaguchi00] calculated an event significance (impact) measure and ranked the segments based on this measure. Further domain-specific choices can be made as well to improve the quality of a summary in a given use case. For example, [Miura03] eliminated all scenes with face images when summarizing cooking videos.

The last step in summary creation is to compose the final summary based on the selected segments. In many cases, a simple concatenation of these video segments results in a summary which is too long. In this case, it is necessary to select certain portions of the segments. For instance, in simple video skimming systems based on keyframe extraction, video segments can be selected around the keyframes in a suitable length using the audio and visual break points as parameters to tune the summarization algorithms for continuity and for a more pleasant viewing experience [Wu00, Ouyang03].

The many efforts described above, targeting the generation of meaningful video summaries and abstracts have had one major deficiency, namely, the lack of a common evaluation platform including a large and representative data set and a collection of well-defined and realistic evaluation criteria. Although they were evaluated on separate datasets, some of which were quite substantial [Liu03, Ngo03], due to the missing thorough comparative analysis under realistic conditions it was not possible to obtain a clear idea about the true performance and usability of the state-of-the-art methods in real-life applications.

In view of the needs described above and building on its success regarding the evaluation of shot-boundary detection, video indexing and search tasks, the TRECVID evaluation benchmark was expanded in 2007 to also address the task of video summarization [Over08a]. The task was built on the use case defined by the professional editors from BBC. It aims at collecting proposals for video summarization mechanisms that are capable of generating meaningful summaries of raw documentary footage referred to as *BBC Rushes*. The term “meaningful” stands for a summary that fulfills a number of predefined criteria. The meaningfulness of the submitted summaries is evaluated under the constraint of a pre-specified summary duration. In the remainder of this chapter, we will refer to BBC Rushes simply as *rushes*.

In this chapter we present the rushes summarization framework that we developed and tested in the scope of TRECVID. After presenting the data collection and the requirements posed on the summaries for this use case in Section 5.2, we

provide a detailed explanation of all components of this framework in Section 5.3. Then, we proceed in Section 5.4 with an extensive overview of the TRECVID evaluation results of the rushes summarization systems from 2008, including the one presented in this chapter. Section 5.5 concludes the chapter.

## 5.2 BBC Rushes: Data description and Summarization Requirements

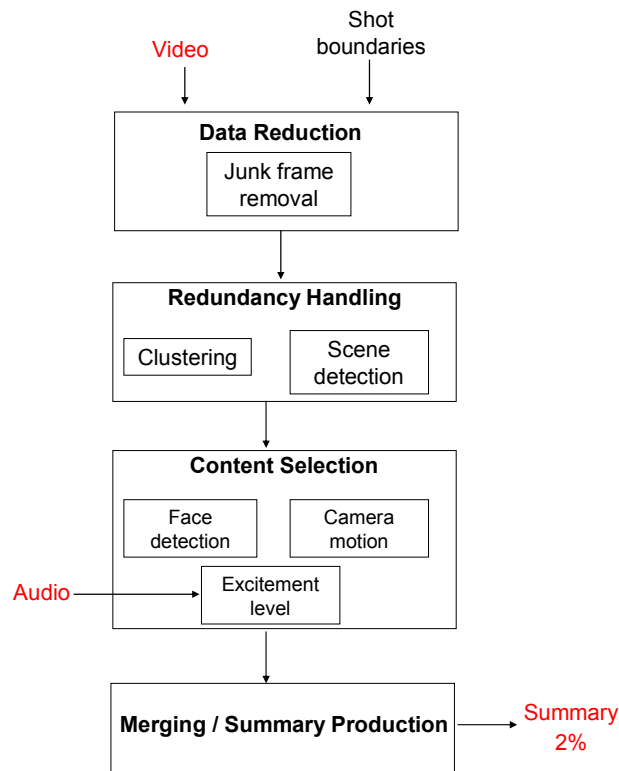
The term “Rushes” stands for the raw material (extra video, B-rolls footage) [Over08a] from which the final video is produced, e.g. in the montage process. Typically, much more material may be shot than actually needed in order to produce enough footage of a sufficient quality to be able to produce the final video. Consequently, rushes often contain many takes of one and the same scene due to various errors or unpredicted effects (e.g. an actor makes a mistake in his text, a plane flies over the scene), which results in many repeating frame sequences. Furthermore, rushes also contain long segments in which the camera is on, but non-active (e.g. before the “official” recording starts) and therefore fixed on a given scene. Finally, the material contains color-bar sections, clap-board, black screens and several other sorts of “junk” material that has no value for the final video.

The main requirement in the rushes summarization task is to automatically create a summary clip, the length of which is maximally 2% of the original rushes video duration. Under this requirement, the amount of “junk” material and the redundancy stemming from multiple takes should be minimized, while the pleasantness of the viewing experience, the easiness of content interpretation and the inclusion of the key objects and events should be maximized. Evaluation of the summaries in view of these criteria is performed manually, by a human judge.

## 5.3 Proposed Rushes Summarization Framework

In view of the peculiarities of the rushes data collection and the requirements of rushes summarization task described in the previous section, we provide here an overview of our developed summarization framework aimed at successfully addressing this task. The data processing in the framework can be said to go through four main stages, that are indicated in Figure 5.1 and that largely follow the general approach to video skimming described in the previous section.

The first *data reduction* stage takes as input the video data and the shot boundaries detected automatically by the module we developed using the methods and approaches discussed in Chapter 2. In this stage, the irrelevant parts of the video are removed. These parts include the “junk” frames that do not show any aspect of the story.

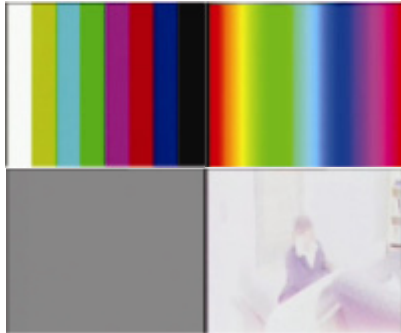


**Figure 5.1** Proposed BBC Rushes summarization framework.

In the second *redundancy handling* stage, similar audiovisual material from different parts of the video is first grouped together in clusters. Due to the specificity of the footage contained in the rushes data set, this similar material typically corresponds to multiple takes of one and the same scene. In this sense, the clustering process can, in this particular case, be seen as a means to perform scene segmentation of the analyzed video.

In the third *content selection* stage, we analyze the content per take to discover semantic content elements that could be of interest for inclusion into the summary. As our intention was to be able to handle any type of content and not only the rushes material, we focused on those semantic concepts that are considered important in a general case, such as faces and motion, since these are the main “ingredients” of the events marking the story of a video. Depending on the presence of relevant semantic content elements, and the assumed or learned importance of these elements, different segments of a scene take are assigned different importance levels that will determine whether these segments will be included into a

summary in the last *summary generation* stage. The four stages and the corresponding processing blocks are explained in detail in the following sections.



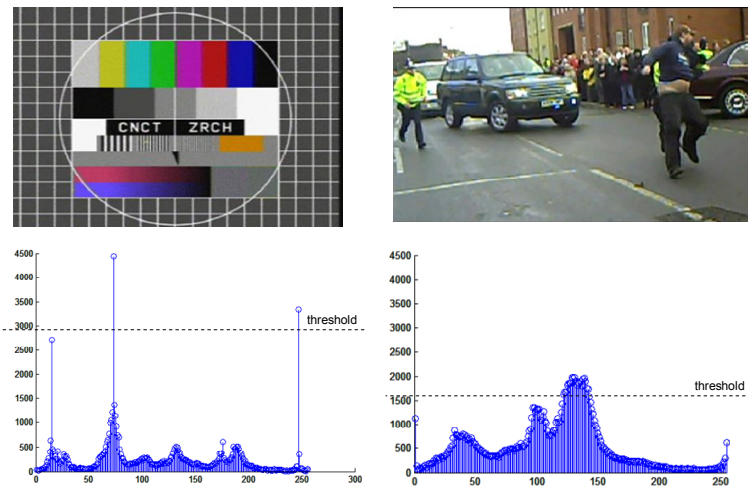
**Figure 5.2** Examples of “junk” frames present in the data set: a) sharp color bars; b) diffuse color bars; c) grey/black frame; d) saturated frame.

### 5.3.1 Data reduction

The processes we developed for this stage aim at reducing the number of non-informative “junk” frames in a summarized video to a minimum. These irrelevant frames typically include color bars, clap boards and single-color frames (e.g. black, white, gray). In our approach we expanded the definition of non-informative frames also to low-quality “regular” frames, such as those saturated ones, which resulted in four classes of junk frames that we addressed in our framework, namely

- sharp color bars (Figure 5.2a),
- diffuse color bars (Figure 5.2b),
- single-color frames (Figure 5.2c), and
- saturated frames (Figure 5.2d).

The method for removing junk frames that we explain in this section can be considered as an independent step in the whole summarization framework. Although the processes in further stages of the summarization framework avoid making assumptions about the processed video content as much as possible and can therefore be considered generic, our junk frame removal method is “engineered” to remove the junk frame types mentioned above. Since this group of frame types is fairly common in general unedited video material and has well-defined properties, a specialized algorithm to detect and remove them in the first place is expected to affect the overall summarization system performance positively in a general case as well.



**Figure 5.3** The difference between the histograms of a single color channel (R-channel) of a synthetic frame (on the left) and a real video frame.

To detect the frames containing one or a few distinct colors, we adopt a simple but effective method based on a thresholded frame histogram computed for each channel of the RGB color space. These synthetic frames are mainly composed of a few fundamental color components. Because of this, the channel histograms of these images have a few impulsive peak values whereas a real video frame has a much smoother histogram shape. We use this fact not only to decide if a video frame is a color test frame (synthetic frame in Figure 5.3), but also to detect diffuse color bars (Figure 5.2b). For the latter case, we apply the same histogram thresholding method, this time to a video frame downsampled to 8x8 pixels, since these images can be considered as having undergone low-pass filtering. This method may falsely detect the parts of scripted scenes with very few colors, such as very dark scenes. This, however, is acceptable as we observed that events in such scenes are generally not suitable to be included in the summary, either.

Another practical problem that repeats in different parts of the rushes videos is the saturated frames. This is an artifact that appears while the camera settings and lighting in the recording area are re-calibrated and is considered as junk material. As can be seen in Figure 5.2d, it is characterized by many saturated pixel values. In our experiments, we observed that the most straightforward methods for detecting the saturated pixels (i.e. comparing mean luminance value or histogram distribution with a threshold to detect highly saturated frames) may result in many false detections. For example, certain light frames with a white background or sky may be detected as “junk sections” and eventually be removed from the video material

to be summarized. For this reason, instead of making the saturated frame decision solely on the frame-by-frame basis, we first check for a fast and steady increase in the saturation level at the beginning of a potential saturated frame group. For this purpose, we adopted the concept of *Accumulating Histogram Difference* (AHD) [Qian06]. This method, which is originally used for the detection of fade-in/outs and flashlights, proved to be successful at detecting the beginning part of a saturation in the video along which a normal frame converts into a saturated frame in a steady manner, as a result of an increase in the light level in the environment or opening of the camera diaphragm. Accumulating histogram difference between two frames is defined as

$$AHD(n) = \sum_{l=x_{min}}^{x_{max}} \Delta H_n(l) = \sum_{l=x_{min}}^{x_{max}} H_{n+1}(l) - \sum_{l=x_{min}}^{x_{max}} H_n(l) \quad (5.1)$$

where  $[x_{min}, x_{max}]$  is the range of the bins  $l$  used to compute the value, and where  $H_n$  is the normalized histogram of frame  $n$ . Each frame is converted into the *Hue, Saturation, Value* (HSV) color space in order to process each channel independently. A saturated frame has typically a low S-value and a high V-value. We selected the number of bins as 256. For the S-channel, the first 35 bins were considered, and for the V-channel the last 20 bins. We used the combined feature vector of length 55 for the AHD calculation.

Once a frame with a significant increase in the saturation is detected in the first step, we compare only the S-channel cumulative histogram of the subsequent frames with a threshold and mark them as saturated frames as long as the cumulative histogram values are over this threshold.

### 5.3.2 Redundancy Handling

Following the detection and removal of the “junk” frames, the second stage in our framework focuses on generating the base for content selection that is to be included in a summary. In view of the specific structure of rushes data characterized by multiple scene takes, we perform this stage by identifying the redundancy in the video content produced by this structure. In other words, it is our objective to group the material from all multiple takes of the same scene together, which will then provide an insight into different “story” parts of the rushes video content that need to be covered in a summary.

In our system we put the emphasis on the visual similarity among the frames to determine the parts in the video that are coherent in terms of the content. The main reason for concentrating on the visual clues is that the visual modality is the most reliable information source to work with in the rushes context. Rushes videos are typically characterized by a lower quality audio, or even by the absence of a soundtrack. Although we also observed in our system that the audio information

(when exists) can significantly contribute to the quality of the summary when utilized properly, building a system using complex audio classification/segmentation or speech recognition methods could pose many problems. Furthermore, a raw rushes video does not contain any textual metadata (e.g. like labels, closed captions or program information data).

To identify the redundancy, we adopt a clustering approach, and more specifically, the recursive spectral clustering based on the normalized cut partitioning criterion [Shi00] that we already employed in Chapter 4. The motivation behind selecting the spectral clustering approach in the rushes scenario is basically the same as in Chapter 4. Like in the concert video analysis case, the complexity and unpredictability of the rushes video content and the lack of base for making assumptions about the cluster distribution in the feature space (e.g. Gaussian assumption for K-means clustering [Duda00]) make spectral clustering more viable than other clustering options.

As an additional step to the utilization of the spectral clustering process in Chapter 4, here we also take into account the temporal proximity between the frames to relate the clusters to the scenes and thus perform *scene segmentation* – the problem we already introduced in Chapter 1. We form the scenes by exploiting the temporal proximity of the frames and frame segments assigned to the same or different clusters.

The use of temporal proximity is motivated by the different characteristics of the rushes video data compared to the music concert videos in Chapter 4. We observe that in the concert videos (as in soccer videos) the event takes place in one environment and the cameras, more or less fixed to their specific places, take the images from different parts of the event hall in a repetitive manner. Therefore, the scenes belonging to a visual cluster (e.g. a zoom-in on the face of the singer) are distributed in the video more or less homogeneously. The rushes videos, however, have sequential characteristics. They are composed of consecutive scene segments, each lasting for a certain time duration and followed by the next one.

For this reason, in a similar manner as in [Odobez03], we introduced the temporal proximity criterion into the weight calculation in Equation 4.12 of Chapter 4 as follows:

$$w_{pq} = e^{-\frac{d(L(p),L(q))^2}{\sigma^2}} \cdot e^{-\frac{d_t(L(p),L(q))^2}{\sigma_t^2}} \quad (5.2)$$

wherin  $d_t(L(p),L(q))$  is the absolute distance between the two nodes expressed in number of frames. The  $\sigma_t$  value, which tunes the effect of the temporal proximity on the final clustering, is set experimentally.

In addition, as a post-processing step, two consecutive segments  $S_1$  and  $S_2$  are put in the same scene if the distance  $d(S_1,S_2)$  between them is smaller than a weighted sum of the lengths of the two segments:

$$d(S_1, S_2) < T \cdot (\text{length}(S_1) + \text{length}(S_2)) \quad (5.3)$$

where  $T$  is an experimentally determined constant.

### 5.3.3 Content Selection

After different scenes and the corresponding takes have been detected, we proceed now with the selection of the segments from each scene that is potentially interesting for being included into a summary. The main rationale behind our approach is that a summary at any level of abstraction (i.e. of any length) should provide insight into the key semantic content elements, such as people, their activities and the actions of the director aiming to emphasize these activities. In our view, those segments of the footage that provide more information on these semantic content elements should count more when creating a summary. This is also visible from the fact that most of the events in the ground truth produced for the evaluation of the BBC Rushes summarization task contain a reference to a human posture or action.

To locate the semantic content elements mentioned above, we focused on face detection, camera motion detection and modeling of the expected excitement elicited in the user when listening to the soundtrack of the rushes (if present). For the latter, we relied on the information from the sound track. We approached the face detection problem by combining two existing detectors, namely Lienhart's extension of the Viola and Jones [Viola04] detector using Haar-like features [OPENCV], and the one that uses skin color appearance model trained on the faces detected by OpenCV. The first implementation of the latter approach was described in [Don05]. Compared to solely relying on the OpenCV based approach, this hybrid method allowed us to increase the recall without degrading the precision.

To detect significant camera motion, we use the algorithm described in [Kraemer06]. First, we estimate the global camera motion and then apply the likelihood significance test to classify specific camera motions. The algorithm allows for classification of the camera motion into physical motion categories, such as *pan/travelling*, *tilt*, *zoom*, *rotation* and other complex motions. The significance criterion enables us to eliminate from the detected motion set those fragments in the footage containing short, noisy camera motions that occur mainly during the scene setup and that may disturb the summary.

Finally, for the modeling of the audio-based excitement level, we used two audio features - energy and pitch - as the basis to estimate the temporal changes of the elicited excitement according to the *arousal* model proposed in [Hanjalic05]. High arousal values are expected to indicate the parts of the footage that are potentially more interesting to be included in the summary than those where arousal is low. An example for the latter segment is where the actors discuss the scene that

is to be recorded. Compared to the actual scene where various dramatic effects may lead to higher arousal values, the discussion segments are typically characterized by lower sound energy and low pitch.

### 5.3.4 Information Fusion and Summary Production

In view of the rushes summarization task specification in Section 5.2, the objective of the last stage of our framework is to build on the results from the previous stages and to generate a summary, the length of which is not longer than 2% of the original length of the rushes video to be summarized. However, in order to be acceptable, the summary still needs to minimize the amount of “junk” material and the redundancy stemming from multiple takes, and maximize the pleasantness of the viewing experience, the easiness of content interpretation and the inclusion of the key objects and events. We developed a system that integrates the output of the previous stages to optimize the quality of a summary in view of the abovementioned requirements.

Since TRECVID does not specify the format of a summary, various summarization approaches are possible, starting from a simple fast-forwarding through the video, via picture-in-picture summaries, to those involving complex content analysis and editing processes. However, in view of the pleasantness of the viewing experience, these options are heavily suboptimal, either because the content is offered at the rate at which it cannot be properly understood, or because the temporally or spatially compressed visual presentation cannot be aligned with the audio track, which may carry important information for understanding the content. For this reason, our system is designed to include the segments not shorter than 2 seconds in the summary.

In terms of content inclusion, our system extracts a video segment from each scene of the original rushes video. The role each segment plays in the summary in terms of its relative duration is determined using the importance value  $I(t)$ , which is measured at each time stamp  $t$  based on the key content elements found in the material at and around that time stamp. The function  $I(t)$  is defined as the weighted sum of the elementary importance indicators  $F(t)$ ,  $C(t)$  and  $E(t)$ , which represent the presence of faces, camera motion and audio excitement level, respectively, that is

$$I(t) = w_F F(t) + w_C C(t) + w_E E(t) \quad (5.4)$$

The weights in Equation 5.4 are empirically set based on the initial tests and observations made on a group of videos belonging to our development set and in view of the ground truth available from the BBC rushes tasks from the previous year. We observed, for instance, that the audio excitement level indicator  $E(t)$  should receive the highest weight, since the arousal information appeared to be the most indicative of the material of interest for the summary. In the absence of audio

data or in the situation where audio was not that informative (e.g. documentaries), the other two indicators still performed rather well in measuring the importance value of the content. For example in the videos where long landscape shots dominate the content, the camera motion feature made the most dynamic shot parts be involved in the summary, which maximized the content richness of the summary.

Once the curve  $I(t)$  is computed, we map individual importance values obtained for different time stamps onto the importance values each characterizing a scene in its entirety, and then use this new value to derive the length of the summary segment representing that scene. To do this we assume that  $N$  scenes,  $S_1 \dots S_N$ , are detected in a video and that the duration of the summary segment taken from each scene can be modeled as a function of the *normalized cumulative scene importance*. We define the normalized cumulative scene importance as the ratio of the sum of the importance values across all time stamps of a scene to the total length of the scene:

$$I(S_i) = \frac{\int_{S_i} I(t) dt}{\text{Length}(S_i)} \quad (5.5)$$

Then, given the total summary length  $T_s$  (e.g. 2% of the original video length), the duration of the segment from the  $i^{\text{th}}$  scene is calculated as the relative value of the normalized cumulative scene importance scaled down by the total summary length:

$$\text{dur}(S_i) = T_s \cdot \frac{I(S_i)}{\sum_j I(S_j)} \quad (5.6)$$

Recalling the rationale explained above, if the duration derived for a scene is less than 2 seconds, that scene is discarded from the summary. In addition to emphasizing the most relevant scenes for the summary, this mechanism also filters out the noise created by unintentionally created, but possibly dynamic short scenes (e.g. dropping the camera while it is still on).

Finally, in the last step of the summarization process, a segment of the duration  $\text{dur}(S_i)$  is created for each scene  $S_i$  around the time stamp  $t_i$  characterized by the maximum importance value and between the limits defined using the following expressions:

$$t_{i,\min} = t_i - \frac{\text{dur}(S_i)}{2} \quad \text{and} \quad t_{i,\max} = t_i + \frac{\text{dur}(S_i)}{2} \quad (5.7)$$

## 5.4 TRECVID Evaluation

In this section, we focus on a comparative evaluation involving a large number of summarization systems participating in the BBC Rushes task 2008, including the system we presented earlier in this chapter. While a detailed evaluation of the entire BBC Rushes round 2008 was presented in [Over08a], which included an extensive description of the assessment procedure and a discussion of the quality of the submitted summaries in general, we address in this section the evaluation aspects that particularly concern our proposed method relatively to other methods.

Before we discuss the relative performance of our summarization method, we briefly explain the summary assessment process applied in TRECVID. The summary videos were evaluated based on qualitative and quantitative measures. To obtain a qualitative evaluation, all the summaries for a given video were viewed by three human judges individually, using mplayer on Linux, in a window 125mm x 102mm, at 25 fps, and in a random order. In a timed process, each judge went through each summary by using the “Play” and “Pause” buttons with the objective to assess the general usability and perceptual quality of the summary. To do this the judge answered the following questions by choosing an option from the 5-point Likert scale<sup>†</sup>:

- “The summary contains many nearly identical segments”,
- “The summary contains color bars, clapboards, and/or totally black or totally white frames”, and
- “The summary is presented in a pleasant tempo and rhythm.”

In addition to the qualitative criteria described above, a number of quantitative measures were taken into account as well to assess the quality of the summaries:

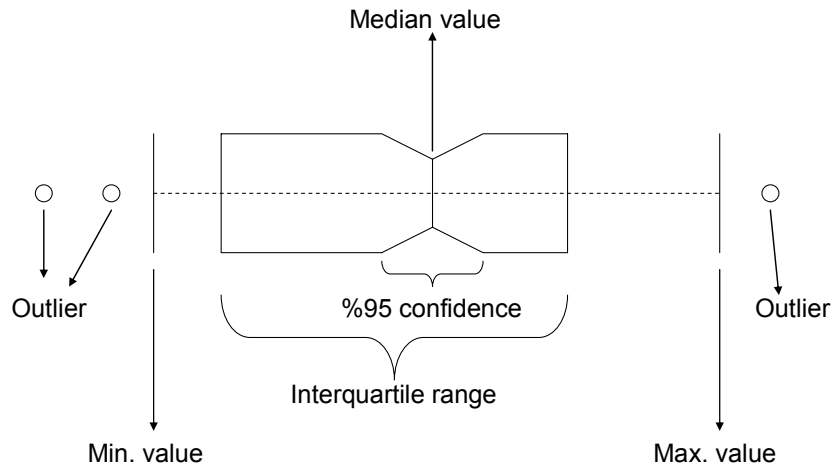
- Fraction of the ground truth objects/events found in the summary by the human judge during the inspection as described above (value range: [0, 1], step 0.08),
- Time needed by the human judge to compare a summary against the ground truth (expressed in seconds),
- Duration of the summary relative to the 2% duration target (computed by subtracting the actual length from the target 2% length), and
- System time required to generate the summary (expressed in seconds).

31 groups participated in the BBC Rushes summarization track of 2008. The systems developed for and evaluated in this track are a good coverage of the state-of-the-art approaches to summarization platform. Many different approaches have

---

<sup>†</sup> [http://en.wikipedia.org/wiki/Likert\\_scale](http://en.wikipedia.org/wiki/Likert_scale)

been implemented for the extraction of the most relevant parts in the video, for removal of the junk and redundant sections and also for formatting the created summary. The summaries obtained were evaluated for each criterion separately, and their scores with respect to each criterion were calculated by combining together the scores of three judges. The Tukey-style boxplot [Cleveland93] used to represent the performance over all summaries and per criterion is illustrated in Figure 5.4. Figures 5.5a-5.5c depict the Tukey-box plots for all systems and all evaluation criteria.



**Figure 5.4** Tukey style plot which is used to represent the performance of each summarization system for each evaluation criterion.

Since the summary format was left as a free parameter, many different formats were deployed for condensing as much information as possible within the 2% target summary duration. These techniques can be grouped into two main categories that were already discussed at a general level earlier in this chapter:

- Simple (e.g. fast-forward) methods, and
- Video skims based on various automated editing approaches, and also in combination with static summaries (e.g. storyboard).

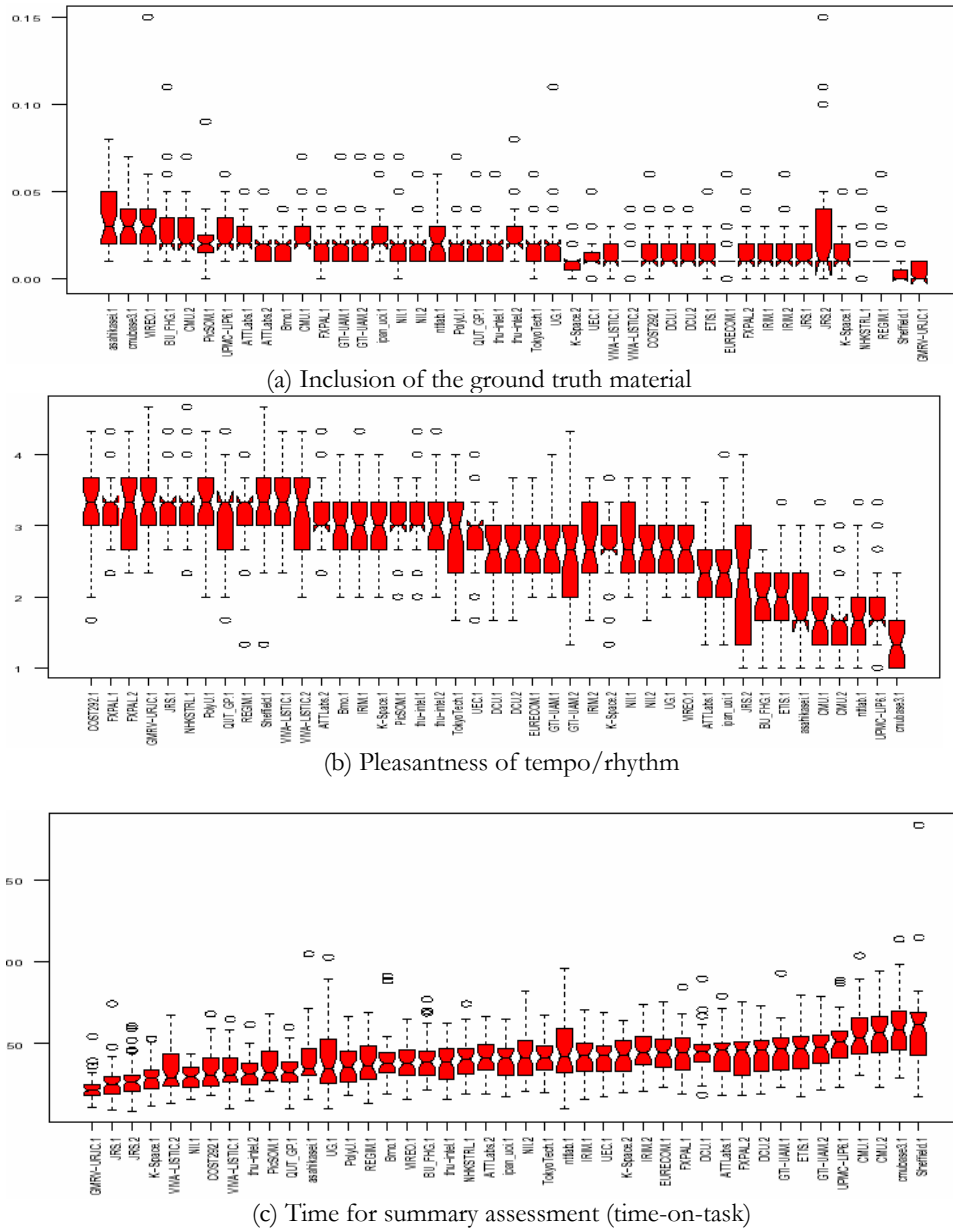
The best representative of the category of “simple” summarization techniques was the fast-forwarding ones that increased the frame rate of the video play uniformly, e.g. up to 50x normal speed [Christel08], to reach the target summary duration. Although being a trivial solution, a summary obtained through a fast-forward

action can be seen as a baseline and, as such, it can provide interesting insights into the relative performance of other, more complex summarization methods, not only in terms of the overall quality of the summary, but also considering the computational complexity of the summarization method. While these methods, not surprisingly, appear to perform best regarding the *inclusion of ground truth objects / events* criterion (Figure 5.5a), Figures 5.5b-c clearly show that the pleasantness and usability of such a summary in terms of the time-on-task criterion is extremely low. This extreme example suggests that the *inclusion of ground truth objects / events* criterion is not informative of the true quality of a summarization if considered in isolation from other qualitative and quantitative criteria.

A higher level of sophistication of the methods from the second category enables these methods to find a better trade-off among different criteria. We now analyze the methods belonging to this category and following the same general idea as in our approach. We choose to focus on those steps in the proposed approaches that take place following the preprocessing (“junk” frame removal) step, since these steps could not be designed based on the knowledge about a particular data set, but needed to be sufficiently generic to successfully handle any raw video footage.

While in our approach no video segmentation step was performed in the initial phase, many approaches rely on pre-defined video segments (shots or sub-shots) as the content segments that are either used for feature extraction in the subsequent summarization steps, or directly as the segments to be included in the summary (e.g. [Bredin08], [Liu08]). The fact that our system performed among the best related to the redundancy removal, pleasantness and time-on-task criteria, and also much better than [Bredin08], [Liu08] may indicate that avoiding the pre-segmentation step not only reduces the computational complexity of the summarization method, but also leads to better results.

The clustering step is performed either after the segmentation operation to find out related or similar segments (e.g. repetitions) or, if no segmentation is performed (like in our approach), on the video or audio data directly. While different alternative clustering techniques have been employed, like the K-means clustering on the sub-shots in [Liu08] or hierarchical clustering in [Dumont-Merialdo08] and [Chen08], we believe that the use of spectral clustering has had positive influence on the much better results obtained by our method with respect to these criteria than the methods mentioned above. It is, however, interesting to see that the method [Bredin08] that skipped the clustering step and performed summarization directly on pre-segmentation results reached a similar performance as [Liu08], [Dumont-Merialdo08] and [Chen08]. A possible conclusion here is that using a suboptimal clustering in the summarization process has the same effect as using no clustering.



**Figure 5.5** Ranked performances of different summarization systems with respect to three criteria: (a) fraction of ground truth material included, (b) pleasantness and usability, and (c) time for summary assessment (time-on-task). While the summary produced by our method (COST292) is only of average quality regarding the inclusiveness criterion (a), it is top ranked in terms of the pleasantness or among the best regarding the time-on-task criterion.

In the last step, the most relevant/informative clusters or segments are selected and their parts are extracted to be included in the summary. For this purpose, typically an importance value is calculated. [Liu08] calculates a frame importance value using temporal content variation and spatial image salience. [Dumont-Merialdo08] calculates the frame activity using motion intensity. Our method was one of the few that also relied on the information from the sound track for this purpose. As explained in Section 5.2.3, we utilized the audio channel as the most significant information source when assessing the importance of the video material at a given time stamp. Indeed, the observations made on the results and in view of the ground truth showed that, especially for the dialogue scenes, the usage of audio information in the form of arousal values significantly contributed to the discovery of the most information intensive parts and to the creation of a dynamic summary experience.

## 5.5 Discussion

In this chapter we introduced a system for creating summaries from unedited rushes videos. The system is designed to create an output in which the redundancy and noisy content is minimized, and, at the same time, all relevant parts of the story are included. Moreover, all this took place under a strict summary duration limit (2% of the original video length). Furthermore, the summary should be pleasant to watch, and the time needed to understand the content from the summary should be minimized.

Our general approach is relatively common in the sense that it considers usual system components, from redundancy removal, via content selection, to pruning the selected content down to the pre-specified limit of summary duration. Also the initial (preprocessing) step in our approach addressed a specific problem (“junk” material) that is characteristic for the raw rushes video footage. However, our goal in the development of the core system was to be as generic as possible and to base the summary generation on as few assumptions about the analyzed data as possible. In that sense, we have not explicitly studied the expected effect of individual system components on the global system performance in the specific rushes scenario. Instead, we based our algorithmic and design choices on the assumptions that are valid for general video material. In addition, our study also indicated the importance of the audio modality in video summarization systems. Our system was the only one in the BBC Rushes round 2008 that relied largely on the audio modality in modeling the importance of the video content that is to be considered for inclusion in the summary. The encouraging results of TRECVID evaluation, and in particular those related to the pleasantness of tempo and rhythm of the summary and the estimated time-on-task required for an expert to obtain an insight in the video content from the summary, revealed a high potential of the algorithmic and design choices in our approach.

---

Further improvements in the quality of rushes summaries and, in general, the summaries of general video documents should be searched through the development of mechanisms for automatically selecting the modalities that are relevant for summary creation in a given use case. In other words, a measure that could indicate based on a preliminary analysis of a part of the collection which modalities are most likely to point to the relevant content to be included in the summary would be helpful. For instance, although the audio-based arousal modeling proved to be a useful tool in selecting the important parts in the videos in the BBC Rushes 2008 use case, this might not be necessarily true for another video collection.



# Appendix A

## Applications

All the algorithms and methods explained in the previous chapters of this thesis are implemented in real life systems and were demonstrated in various events and conferences. Implementation of a method in a real life system introduces different problems and challenges than the standard training and test procedures for the performance evaluation of an algorithm. On the other hand, it is only after embedding the algorithm in a real world scenario and letting it work in an uncontrolled environment that it is possible to have an idea about its true practical applicability.

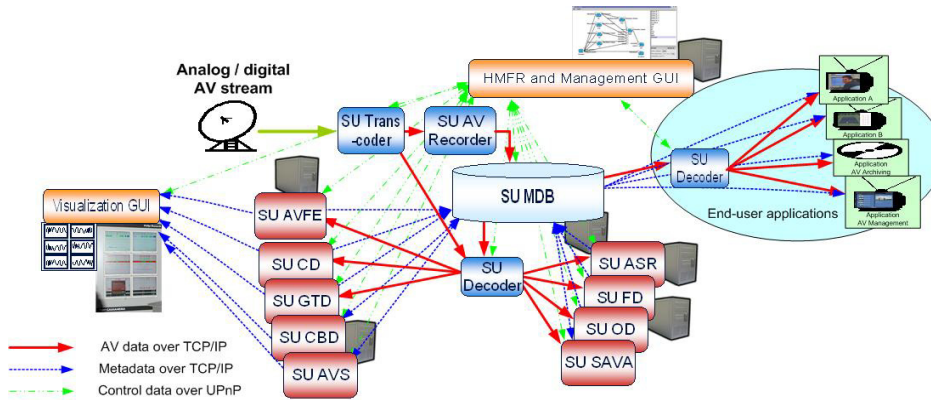
In this appendix we introduce the different systems that our algorithms are embedded in and explain their functionality. In this context, we refer to three systems that are related to the three methods developed in the context of this thesis:

- The Philips CASSANDRA system [Nesvadba05] that our shot boundary detection algorithm from Chapter 2 is developed for,
- The FabChannel concert video broadcasting website and Production Street [Houten05] that the automatic concert video indexing system described in Chapter 4 is embedded in, and
- The COST292 Rushes analysis and summarization system [Naci08a] that we developed based on the material presented in Chapter 5 to participate in TRECVID 2008.

## A.1 Shot Boundary Detection

The CASSANDRA Framework [Nesvadba05] at Philips Research is a prototype system developed for simulating the distributed computing scenarios in In-Home networks. Its architecture is designed to allow available computational resources to be shared for complex multimedia content analysis applications. Furthermore, the system has been used to evaluate subjectively different multimedia content analysis algorithms, both individually and in combinations, processing vast hours of real-life (broadcast) video content.

The CASSANDRA system has a modular structure that allows the development and insertion of *service units*. Each service unit is an independent process that communicates with other service units using the TCP/IP protocol. The outcomes of each service unit can be monitored on a Graphical User Interface. Although the target platform for the system is Linux on x86, its software architecture allows easy migration to other potential platforms in the future.



**Figure A. 1** The CASSANDRA Framework.

The shot boundary detection system introduced in Chapter 2 has been implemented as one of the service units in the CASSANDRA system. To qualify as a service unit, the algorithm needed to satisfy two main criteria.

The first criterion is the real-time requirement. As the shot-boundary detection service unit is a fundamental one that operates continuously and feeds the extracted information to other service units, it is imperative that it doesn't flood the available system resources. The implementation of the proposed algorithm in the CASSANDRA platform performed up to 10 times faster than the real time video play. In addition, the algorithm ran on raw video data, as opposed to the majority of alternative efficient algorithms that work in the compressed domain. This allows our system the flexibility of being capable of extracting the shot boundary informa-

tion in (faster than) real-time from different sources of multimedia data, independent of the choice of the compression and transmission system.

Also related to the complexity requirements is the criterion that a shot boundary detector should be able to handle as many different transition types as possible, and especially various types of gradual transitions. Contrary to many existing algorithms that we explained in more detail in Chapter 3, which are basically limited to specific simple transition types like fade in/out or simple screen wipes, our method detects any type of shot transitions independent of the number and type of the graphical effects applied.

Long-term subjective evaluation of the proposed algorithm as a part of the CASSANDRA system has shown that the proposed algorithm is stable and has the potential to function successfully in a real-life use scenario.

## A.2 The MultimediaN Concert Video Browser System

The automatic event extraction system from concert videos explained in Chapter 4 is developed for and integrated in the MultimediaN\* Concert Video Browser framework. This system demonstrates a video interaction environment for efficiently browsing live video registrations of pop, rock and other music concerts. The system is based on the FabChannel's web based concert video browser (Fab-Player) which, in addition to the indexing module developed in this thesis, also contains an automatic multimedia content analysis module. This module analyzes the audiovisual content of live concert registrations at the affective level and estimates the level of excitement elicited in a user while watching the video. This way we aim at better managing the relatively unorganized nature of concert videos and improving the browsing experience of large concert video collections with minimum human intervention needed in the collection indexing and organization steps.

The MultimediaN Concert Video Browser framework is composed of three main components: The first component is the *video concert broadcasting system* developed by FabChannel [FabChannel]. The system consists of a concert database recorded in the biggest concert halls in Amsterdam (Melkweg and Paradiso), and of a web-based interface that lets the users to choose a concert and display it (Figure A.2). This system has been of the most successful web concepts in the past years and received many awards, including the renowned Webby Award in 2006. For developing the advanced browsing framework, the classical interface has been changed to the advanced browsing interface that lets the users to access the applause, instrument solos and the concert segments with varying excitement levels.

The second component of the framework is the *Production Street* developed by the Telematica Institute. Production Street is a general purpose video editing sys-

---

\* Dutch BSIK MultimediaN research program (2004-2009), in the context of which the research for this thesis was performed

tem and application programming unit, which aims at enriching the video watching experience using metadata. The output of the algorithms inserted into Production Street serves to provide such enrichment and is formatted using the MPEG-7 standard.

The last component of the system consists of various algorithms integrated in the Production Street. The research performed for this thesis contributed two such algorithms, namely the (affective) multimedia content analysis and indexing solutions mentioned above. The content analysis algorithm models the excitement elicited at a user in the form of an excitement time curve, the peaks of which indicate the potential highlights of a concert. The algorithm used for this purpose is an adaptation and implementation of the method proposed in [Hanjalic05]. The second algorithm we implemented for the MultimediaN Concert Video Browser System is the automatic event detection algorithm whose details we gave in Chapter 4. This algorithm, which is implemented as a DLL library, is computationally efficient and compatible with the requirements of the Production Street. The algorithm processed the whole concert database of FabChannel and produced the output metadata in MPEG-7 format.

This framework shown in Figure A.2 has been demonstrated in many events and conferences [Houten05, Naci07a] and received a very good feedback from the users. In addition to the quantitative performance figures reported in previous chapters, these demonstrations provided the situation for subjective evaluation for the framework and for the assessment of its practical applicability.

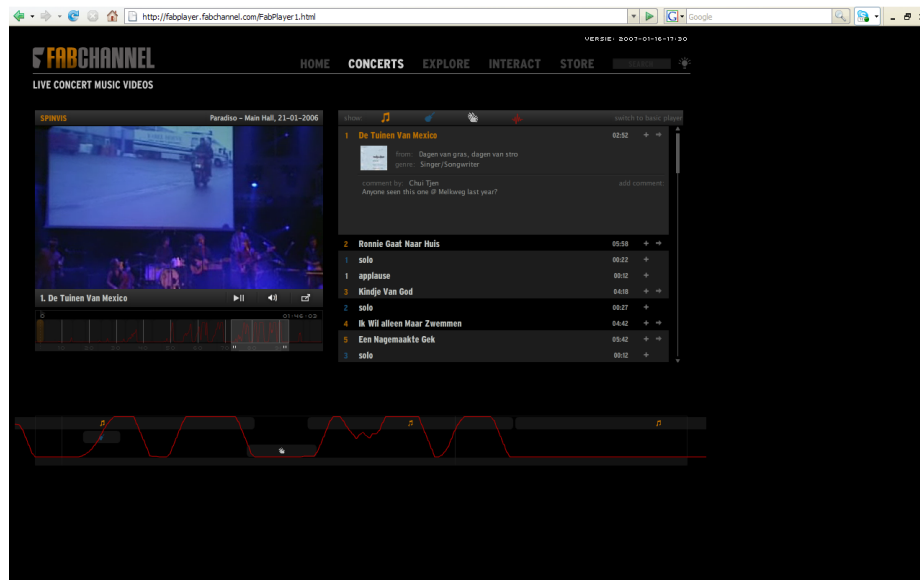
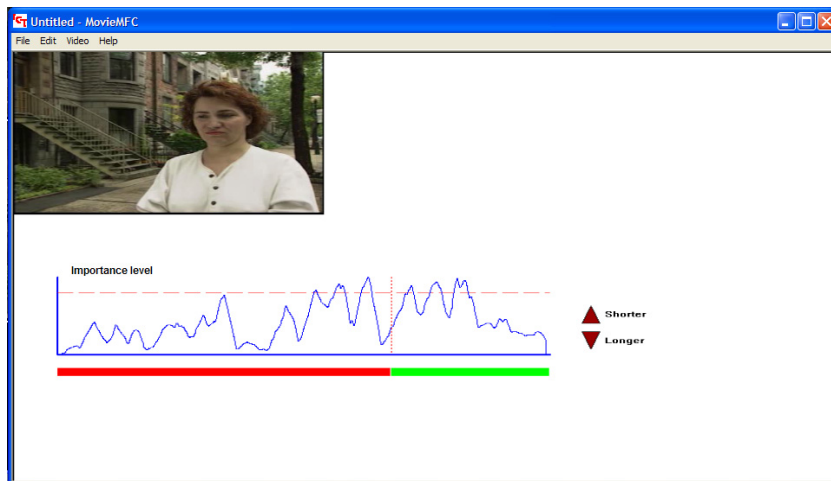


Figure A. 2 Advanced FabChannel Interface.

### A.3 Ruses Summarization

Based on the ideas and methods described in Chapter 5, we implemented a platform and interface for rushes summarization. The platform enables displaying of the outcomes of different algorithmic patches in the system (e.g. the clustering unit, junk frame extraction unit, audio excitement level unit) as shown in Figure A.3 and assessing their effect on the final summary extracted. In this way, influence of different algorithmic and design choices on the summary generation can be investigated. In addition, the platform can accept a new rushes video as input and automatically generate a summary in a pre-specified length. The platform was demonstrated in TRECVID 2007 and TRECVID 2008, and performed as one of the best systems in the Ruses task in 2008 [Naci08a].



**Figure A. 3** The rushes summary extraction interface. One can observe the content importance function extracted automatically by the system using the methodology that is similar to the one for the affective multimedia content analysis for live concert video registrations.



## Bibliography

- [Agostini03] G. Agostini, M. Longari and E. Pollastri, “Musical Instrument Timbres Classification with Spectral Features”, *Eurasip Journal on Applied Signal Processing*, vol.1, pp. 5-14, 2003.
- [Alattar98] A.M. Alattar, “Wipe Scene Change Detector for Segmenting Uncompressed Video Sequences”, *IEEE International Symposium on Circuits and Systems*, pp. 249-252, 1998.
- [Amir03] A. Amir, S. Srinivasan and D. Ponceleon, “Efficient video browsing using multiple synchronized views,” in *Video Mining*, Boston, MA: Kluwer, August, 2003.
- [Aner02] A. Aner, and J. R. Kender, “Video summaries through mosaic-based shot and scene clustering”, in *Proceedings of the European Conference on Computer Vision*, Denmark, 2002
- [Ankush02] M. Ankush, L. F. Cheong and T. S. Leung, “Robust identification of gradual shot-transition types”, in *Proceedings of IEEE International Conference on Image Processing*, pp. 413-416, 2002.
- [Ariki03] Y. Ariki, M. Kumano, and K. Tsukada, “Highlight scene extraction in real time from baseball live video”, in *Proceedings of the 5th International Workshop on Multimedia Information Retrieval (ACM SIGMM)*, pp. 209–214, 2003.
- [Ashley95] J. Ashley, M. Flickner, J.L. Hafner, D. Lee, W. Niblack and D. Petkovic, “The Query By Image Content (QBIC) System”, *SIGMOD Conference*, 1995.
- [Babaguchi00] N. Babaguchi, “Towards abstracting sports video by highlights”, in *Proceedings of the ICME Conference*, New York, 2000.
- [Backstrom06] L. Backstrom, D. Huttenlocher, J. Kleinberg and X. Lan, “Group formation in large social networks: Membership, growth, and evolution”, *Proceedings of 12th International Conference on Knowledge Discovery in Data Mining*, pp. 44-54, New York: ACM Press, 2006.
- [Ballard82] D.H. Ballard and C.M. Brown, *Computer Vision*, Prentice Hall, New Jersey, USA, 1982.
- [Barbieri07] M. Barbieri, *Automatic Summarization of Narrative Video*, PhD Thesis, Technische Universiteit Eindhoven, 2007.

- [Berry99] M. Berry, Z. Drmac, and E. Jessup, "Matrices, Vector Spaces and Information Retrieval", *SIAM Review*, vol. 41, no. 2, pp. 335-362, 1999.
- [Bescos04] J. Bescos, "Real-Time Shot Change Detection over Online MPEG-2 Video", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 14 No. 4, April, 2004.
- [Bouthemy99] P. Bouthemy, M. Gelgon, and F. Ganansia, "A Unified Approach to Shot Change Detection and Camera Motion Characterization", *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 9(7), pp. 1030-1044, October, 1999.
- [Bowyer06] K.W. Bowyer, K. Chang and P. Flynn, "A Survey of Approaches and Challenges in 3D and Multimodal 3D+ 2D Face Recognition", *Computer Vision and Image Understanding*, 101:1–15, 2006.
- [Burred08] J.J. Burred, M. Haller, S. Jin, A. Samour and T. Sikora, "Audio Content Analysis", *Semantic Multimedia and Ontologies: Theory and Applications*, pp. 123-162, Springer, London, 2008.
- [Cai05] R. Cai, L. Lu, and A. Hanjalic, "Unsupervised content discovery in composite audio", *Proc. ACM Multimedia*, pp. 628-637, 2005.
- [Campbell06] M. Campbell, S. Ebadollahi, D. Joshi, M. Naphade, A. Natsev, J. Seidl, J. R. Smith, K. Scheinberg, J. Tešić, L. Xie and A. Haubold, "IBM Research TRECVID-2006 Video Retrieval System", in *Proceedings of TRECVID 2006 Workshop*, Gaithersburg, USA, 2006.
- [Cao06] J. Cao, et al., "Intelligent Multimedia Group of Tsinghua University at TRECVID 2006", in *Proceedings of TRECVID 2006 Workshop*, Gaithersburg, USA, 2006.
- [Carmichael08] J. Carmichael, M. Larson, J. Marlow, E. Newman, P. Clough, J. Oomen and S. Sav, "Multimodal Indexing of Electronic Audio-Visual Documents: A Case Study for Cultural Heritage Data", in *Proc. 6th International Workshop on Content-Based Multimedia Indexing (CBMI)*, London, June, 2008.
- [Cheong00] L.F. Cheong, "Scene-Based Shot Change Detection and Comparative Evaluation", *Computer Vision and Image Understanding*, 79, 2, pp. 224-235, August 2000.
- [Chan05] C.H. Chan and G.J.F. Jones, "Affect-based Indexing and Retrieval of Films", *Proceedings of the 13th annual ACM international conference on Multimedia*, pp. 427-430, 2005.
- [Chung07] Y.Y. Chung, E.H.C. Choi, Z. Zhao, M.A.M. Shukran, D.Y. Shi and F. Chen, "An Efficient Tree-based Quantization for Content Based Music Retrieval System", *Proceedings of the 2007 WSEAS International Conference on*

- Computer Engineering and Applications*, Gold Coast, Australia, January 17-19, 2007.
- [Cleveland93] W.S. Cleveland, *Visualizing Data*, Summit, NJ, Hobart Press, 1993.
- [Cooper04] M. Cooper, "Video Segmentation Combining Similarity Analysis and Classification", *Proc. ACM Multimedia*, 2004.
- [DelBimbo99] A. Del Bimbo, *Visual Information Retrieval*, Morgan Kaufmann Publishers, Inc., 1999.
- [Deng08] J.D. Deng, C. Simmermacher and S. Cranefield, "A Study on Feature Analysis for Musical Instrument Classification", in *IEEE Transactions on Systems, Man, and Cybernetics*, Part B, 2008.
- [Don05] A. Don, L. Carminati, and J. Benois-Pineau, "Detection of Visual Dialog Scenes in Video Content based on Structural and Semantic Features", in *Proc. CBMI'05*, Letonie, 2005.
- [Duda00] R.O. Duda, P.E. Hart and D.G. Stork, *Pattern Classification*, Wiley, NY, 2000.
- [Dufaux00] F. Dufaux, "Key Frame Selection to Represent a Video", in *Proceedings of the ICIP Conference*, vol. 2, pp. 275–278, 2000.
- [Ellis07] D.P.W. Ellis and G.E. Poliner. "Identifying 'Cover Songs' with Chroma Features and Dynamic Programming Beat Tracking" in *Proc. of the Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, volume IV, pp. 1429–1432, Honolulu, USA, April 2007.
- [Eronen01] A. Eronen, "Comparison of Features for Musical Instrument Recognition," *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, pp. 19–22, 2001.
- [FabChannel] <http://www.fabchannel.com/>
- [Faloutsos95] C. Faloutsos and K.-I. Lin. "Fastmap: A Fast Method for Indexing, Data-mining and Visualization of Traditional and Multimedia Datasets", in *Proc. of the ACM SIGMOD Conference*, pp. 163-174, San Jose, CA, 1995.
- [Ferguson03] R. Ferguson, *Automatic Segmentation in Concert Recordings*, M.A. Thesis, McGill University, 2003.
- [Ferret02] O. Ferret, "Using Collocations for Topic Segmentation and Link Detection", *Proc. of the 19th International Conference on Computational linguistics*, Taipei, Taiwan, 2002.
- [Freeman06] L. Freeman, *The Development of Social Network Analysis*, Vancouver, Canada, Empirical Pres, 2006.

- [Gantz08] J. F. Gantz, D. Reinsel, C. Chute, W. Schlichting, S. Minton, A. Toncheva and A. Manfrediz, "The Expanding Digital Universe: An Updated Forecast of Worldwide Information Growth Through 2011", International Data Corporation, sponsored by EMC Corporation, March, 2008.
- [Gillet05] O. Gillet and G. Richard, "Automatic Transcription Of Drum Sequences Using Audiovisual Features", in *Proc. of ICASSP*, Philadelphia, USA, 2005.
- [Gillet07] O. Gillet, S. Essid, and G. Richard, "On the Correlation of Automatic Audio and Visual Segmentations of Music Videos", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 17, No. 3, March 2007.
- [Girgensohn99] A. Girgensohn, and J. Boreczky, "Time-Constrained Keyframe Selection Technique", in *Proceedings of the IEEE International Conference on Multimedia and Systems*, vol. 1 (Florence, Italy), pp. 756-761, 1999.
- [Gong03] Y. Gong and X. Liu, "Video Summarization and Retrieval using Singular Value Decomposition", *ACM Multimedia Syst. Journal*, vol. 9, pp. 157-168, 2003.
- [Grauman06] K. Grauman and T. Darrell, "Unsupervised Learning of Categories from Sets of Partially Matching Image Features", in: *Proc. IEEE Computer vision and Pattern Recognition*, 2006.
- [Han05] S.-H. Han, and I.-S. Kweon, "Scalable Temporal Interest Points for Abstraction and Classification of Video Events", in *International Conference on Multimedia and Expo.*, 2005.
- [Hanjalic02] A. Hanjalic, "Shot-Boundary Detection: Unraveled and Resolved?", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 2, February, 2002.
- [Hanjalic03] A. Hanjalic, "Generic Approach to Highlights Extraction from a Sport Video" in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, vol. 1, (Barcelona, Spain), 1-4, 2003.
- [Hanjalic04] A. Hanjalic, *Content-Based Analysis of Digital Video*, Kluwer Academic Publishers, 2004.
- [Hanjalic05] A. Hanjalic and L-Q. Xu, "Affective Video Content Representation and Modeling", in *IEEE Transactions on Multimedia*, vol. 7(1), pp. 171-180, 2005.
- [Hanjalic97] A. Hanjalic, R.L. Legendijk, and J. Biemond, "A New Method for Key Frame based Video Content Representation", in *Image Databases and Multimedia Search*, World Scientific, pp. 97-107, 1997.

- [Hanjalic99] A. Hanjalic, R.L. Lagendijk and J. Biemond, "Automated High-level Movie Segmentation for Advanced Video-retrieval Systems", *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 9, no. 4, pp. 580--588, 1999.
- [Haque06] S. Haque, R. Tongeri and A. Zaknich, "Zero Crossings with Adaptation for Automatic Speech Recognition", *Proc. of 11th Australian International Conference on Speech Science and Technology*, Auckland, New Zealand, December, 2006.
- [Hauptmann08] A.G. Hauptmann, M.G. Christel and R. Yan, "Video Retrieval based on Semantic Concepts", *Proceedings of the IEEE*, vol.96, no.4, pp. 602-622, April 2008.
- [Higgins80] R.C. Higgins, "The Utilization of Zero Crossing Statistics for Signal Detection", *J. of the Acoustical Society of America*, vol. 67, pp. 1818-1820, May, 1980.
- [Houten05] Y. van Houten, U. Naci, B. Freiburg, R. Eggermont, S. Schuurman, D. Hollander, J. Reitsma, M. Markslag, J. Kniest, M. Veenstra and A. Hanjalic, "The MultimediaN Concert Video Browser", *IEEE International Conference on Multimedia and EXPO*, Amsterdam, 2005.
- [Hsu06] W.H. Hsu, L.S. Kennedy and S.-F. Chang, "Video Search Reranking via Information Bottleneck Principle", *Proceedings of the 14th annual ACM international conference on Multimedia*, 2006.
- [Huang08] T.S. Huang, C.K. Dagi, S. Rajaram, E.Y. Chang, M.I. Mandel, G.E. Poliner and D.P.W. Ellis, "Active Learning for Interactive Multimedia Retrieval", *Proceedings of the IEEE*, vol. 96(4), pp. 648-667, April, 2008.
- [Ilow07] J. Ilow and P. Venkatasubramanian, "Applications of Level Crossing Theory to Clipping Noise Characterization in Filtered OFDM Signals", *IEEE CCNC'07*, pp. 470-473, 2007.
- [Irani98] M. Irani, P. Anandan and S. Andhsu, "Video Indexing based on Mosaic Representations", in *Proceedings of the IEEE*, vol. 86, pp. 905-921, 1998.
- [Jaimes05] A. Jaimes, T. Nagamine, J. Liu, K. Omura and N. Sebe, "Affective Meeting Video Analysis", *IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1412-1415, 2005.
- [Jaimes08] A. Jaimes and N. Sebe, "Multimodal human-computer interaction: A survey", *Computer Vision and Image Understanding, Special Issue on Vision for Human-Computer Interaction*, vol. 108, issues 1-2, October, 2008.
- [Jain00] A. K. Jain, R.P.W. Duin, and J. Mao, "Statistical Pattern Recognition: A review," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, pp. 4-37, Jan. 2000.

- [Kac43] M. Kac, "On the Average Number of Real Roots of a Random Algebraic Equation", *Bull. Amer. Math. Soc.*, vol. 49, pp. 314-322, 1943.
- [Kakadiaris05] I. Kakadiaris, G. Passalis, T. Theoharis, G. Toderici, I. Konstantinidis and N. Murtuza, "Multimodal Face Recognition: Combination of Geometry with Physiological Information", in *Proc. Computer Vision and Pattern Recognition (CVPR)*, San Diego, CA, June 2005.
- [Kang99] E.K. Kang, S.J. Kim, and J.S. Choi, "Video Retrieval based on Scene Change Detection in Compressed Domain", *IEEE Trans. Consum. Electron.*, vol. 45(3), pp. 932-936, 1999.
- [Kasutani01] E. Kasutani and A. Yamada, "The MPEG-7 Color Layout Descriptor: A Compact Image Feature Description of High-speed Image/Video Segment Retrieval", in *Proc. ICIP'01*, Greece, 2001.
- [Kedem80] B. Kedem, *Binary Time Series*, New York, NY, Dekker, 1980.
- [Kedem81] B. Kedem and E. Slud, "On Goodness of Fit of Time Series Models: An Application of Higher Order Crossings", *Biometrika*, vol. 68, pp. 551-556, 1981.
- [Kedem86] B. Kedem, "Spectral Analysis and Discrimination by Zero Crossings", *Proc. of the IEEE*, vol. 74, no.11, pp. 1477-1493, November, 1986.
- [Kedem87] B. Kedem, "Higher Order Crossings in Time Series Model Identification", *Technometrics*, vol. 29, no. 2, pp. 193-204, May, 1987.
- [Kender98] J.R. Kender and B.-L. Yeo, "Video Scene Segmentation via Continuous Video Coherence", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1998.
- [Kim02] C. Kim, and J.-N. Hwang, "Object-Based Video Abstraction for Video Surveillance Systems", *IEEE Trans. Circ. Syst. Video Technol.*, vol. 12, pp. 1128-1138, 2002.
- [Kitahara07] T. Kitahara, *Computational Musical Instrument Recognition and its Application to Content-based Music Information Retrieval*, PhD Thesis, Kyoto University, Kyoto, Japan, 2007.
- [Kittler98] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas, "On Combining Classifiers", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, March, 1998.
- [Knees07] P. Knees, T. Pohle, M. Schedl and G. Widmer, "A Music Search Engine Built upon Audio-based and Web-based Similarity Measures", *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, Amsterdam, 2007.

- [Kobla99] Kobla, V., D. DeMenthon, and D. Doermann, "Special Effect Edit Detection using Video Trails: A Comparison with Existing Techniques," *SPIE Storage and Retrieval for Image and Video Databases VII*, pp. 302–313, 1999.
- [Kokaram06] A. Kokaram, N. Rea, R. Dahyot, M. Tekalp, P. Bouthemy, P. Gros and I. Sezan, "Browsing Sports Video: Trends in Sports-Related Indexing and Retrieval Work," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 47–58, 2006.
- [Komlodi98] A. Komlodi and G. Marchionini, "Key Frame Preview Techniques for Video Browsing", in *DL: Proceedings of the 3rd ACM Conference on Digital Libraries*, ACM Press, New York, pp. 118–125, 1998.
- [Kontostathis06] A. Kontostathis and W. Pottenger, "A Framework for Understanding Latent Semantic Indexing (LSI) Performance", *Information Processing and Management*, 2006
- [Kraaij06] W. Kraaij, P. Over, T. Ianeva and A.F. Smeaton, "TRECVID 2006 - An Introduction", in *Proc. TREC Video Retrieval Evaluation (TRECVID)*, Gaithersburg, MD, November 2006.
- [Kraemer06] P. Kraemer, J. Benois-Pineau, and M. Gracia Pla, "Indexing Camera Motion Integrating Knowledge of Quality of the Encoded Video", in *Proc. SAMT'06*, 2006.
- [Kratz06] M. F. Kratz, "Level Crossings and Other Level Functionals of Stationary Gaussian Processes", *Probability Surveys*, vol.3, pp. 230-288, 2006.
- [Langville06] A.N. Langville and C.D.Meyer, *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press. ISBN 0-691-12202-4, 2006.
- [Lee03] J.-H. Lee, G.-G. Lee and W.-Y. Kim, "Automatic Video Summarizing Tool Using MPEG-7 Descriptors for Personal Video Recorder", *IEEE Transactions on Consumer Electronics*, vol. 49, issue 3, pp. 742 – 749, August 2003.
- [Leferve03] S. Leferve, J. Holler and N. Vincent, "A Review of Real-Time Segmentation of Uncompressed Video Sequences for Content-Based Search and Retrieval", *Real Time Imaging*, vol. 9, pp. 73-98, 2003.
- [Levine85] M. Levine, *Vision in Man and Machine*, Mcgraw Hill, Columbus, 1985.
- [Lienhart01a] R. Lienhart, "Reliable dissolve detection," in *Proc. SPIE*, vol. 4315, pp. 219–230, 2001.

- [Lienhart01b] R. Lienhart, "Reliable Transition Detection in Videos: A Survey and Practitioner's Guide", *International Journal of Image and Graphics (IJIG)*, vol.1, no.3, pp.469-486, 2001.
- [Lienhart99] R. Lienhart, "Abstracting Home Video Automatically", *Proceedings of the 7th ACM international conference on Multimedia*, Orlando, Florida, pp. 37–40, 1999.
- [Liu03] T. Liu, H.-J. Zhang, and F. Qi, "A Novel Video Key-frame Extraction Algorithm based on Perceived Motion Energy Model", *IEEE Trans. Circ. Syst. Video Technol.*, vol. 13, no. 10 (Oct.), pp. 1006–1013, 2003.
- [Lowe04] D.G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints", *International Journal of Computer Vision*, vol. 60(2), pp.91-110, November, 2004.
- [Lu02] L. Lu, H.-J. Zhang and H. Jiang, "Content Analysis for Audio Classification and Segmentation", *IEEE Trans. Speech and Audio Proc.*, vol. 10, no. 7, pp. 504-516, October 2002.
- [Lu05] S. Lu, I. King and M. Lyu, "Video Summarization using Greedy Method in a Constraint Satisfaction Framework", in *Proceedings of the 9th International Conference on Distributed Multimedia Systems (DMS)*, (Miami, FL), pp. 456–461, 2005.
- [Lu06] L. Lu and A. Hanjalic. "Towards Optimal Audio Keywords Detection for Audio Content Analysis and Discovery", *Proc. ACM Multimedia*, pp. 825-834, Santa Barbara, CA, Oct.23-27, 2006.
- [Lupatini98] G. Lupatini, C. Saraceno and R. Leonardi, "Scene Break Detection: A Comparison", in *Proceedings of 8th Workshop on Continuous-Media Databases and Applications*, pp. 34-41, 1998.
- [Magedanz08] T. Magedanz, "MS vs. P2P and Web 2.0 - Understanding the Role of the IP Multimedia System (IMS) in Face of a Converging Telco and Internet Service World", *Applications and Services in Wireless Networks*, Kassel, 2008.
- [Meila00] M. Meila and J. Shi, "Learning Segmentation by Random Walks", in *Proc. NIPS*, 2000.
- [Meila01] M. Meila and J. Shi, "A Random Walks View of Spectral Segmentation", *International Conference on AI and Statistics (AISTAT)*, Key West, FL, January, 2001.
- [Meng95] J. Meng, Y. Juan and S. F. Chang, "Scene Change Detection in a MPEG Compressed Video Sequence," in *Proc. SPIE 2419: Digital Video Compression*, pp. 267-272, 1995.

- [Mermelstein76] P. Mermelstein, "Distance Measures for Speech Recognition, Psychological and Instrumental", in *Pattern Recognition and Artificial Intelligence*, C. H. Chen, Ed., pp. 374–388. Academic, New York, 1976.
- [Miura03] K. Miura, R. Hamada, I. Ide, S. Sakai and H. Tanaka, "Motion based Automatic Abstraction of Cooking Videos", in *IPSJ Trans. Comput. Vis. Image Media*, vol. 44, 2003.
- [Mobasher07] B. Mobasher, "Data Mining for Personalization", in *the Adaptive Web: Methods and Strategies of Web Personalization*, pp. 90-135, Berlin, Heilderberg, Springer, 2007.
- [Mooers59] C.N. Mooers, Information Retrieval Selection Study. Part II: Seven System Models. Cambridge, MA: Zator Company, 1959.
- [Morchen06] F. Morchen, A. Ultsch, M. Thies and I. Lohken, "Modeling Timbre Distance with Temporal Statistics from Polyphonic Music", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14(1), pp. 81- 90, January, 2006.
- [Naci03] S.U. Naci. "Self Score Normalization and Frame Pruning Techniques for Speaker Verification Systems", M.S. Thesis, Bogazici University, 2003.
- [Naci04] S.U. Naci and A. Hanjalic, "A Survey on Detection and Identification of Gradual Video Shot Transitions", *Technical Report ICT-2004-07*, 2004.
- [Naci05a] S.U. Naci and A. Hanjalic, "TU DELFT at TRECVID 2005: Shot Boundary Detection", *Proceedings of TRECVID 2005*, November 2005.
- [Naci05b] S.U. Naci and A. Hanjalic, "A Unified Framework for Fast and Effective Shot Transition Detection based on Analysis of Spatiotemporal Video Data Blocks", *4th International Workshop on Content-Based Multimedia Indexing (CBMI 2005)*, Riga, Latvia, 2005.
- [Naci06] S.U. Naci and A. Hanjalic, "Low Level Analysis of Video using Spatiotemporal Pixel Blocks", *Lecture Notes in Computer Science*, vol. 4105, pp. 777-784, Springer Berlin / Heidelberg, 2006.
- [Naci07a] S.U. Naci and A. Hanjalic, "Intelligent Browsing of Concert Videos", *Proceedings of the 15th international conference on Multimedia (ACM MM '07)*, ACM Press, Augsburg, Germany, 2007.
- [Naci07b] S.U. Naci and A. Hanjalic, "An Audio Feature Based System for Concert Video Parsing", *Proc. of the 13th Conference of the Advanced School of Computing and Imaging (ASCI 2007)*, pp. 401-404, 2007.
- [Naci08a] S.U. Naci, U. Damjanovic, B. Mansencal, J. Benois-Pineau, C. Kaes, M. Corvaglia, E. Rossi and N. Aginako, "The COST292 Experimental Framework for RUSHES Task in TRECVID 2008", in *TVS'08: Proceedings of*

- the International Workshop on TRECVID Video Summarization*, pages 1–20, ACM, 2008.
- [Naci08b] S.U. Naci and A. Hanjalic, "Content-Based Indexing Of Music Concert Recordings Based On Crossing-Rate Features", *6th International Workshop on Content-Based Multimedia Indexing (CBMI 2008)*, London, UK, June 2008.
- [Nesvadba05] J. Nesvadba, P. Fonseca, A. Sinitsyn, H. Broers, A. Korostelev, J. Ypma, B. Kroon, H. Celik, J. Lukkien, A. Hanjalic, S.U. Naci, J. Benois-Pineau, P. de With and J. Han, "Real-Time and Distributed AV Content Analysis System for Consumer Electronics Networks", *IEEE International Conference on Multimedia and EXPO*, Amsterdam, 2005.
- [Ng01] A. Ng, M.I. Jordan, and Y. Weiss, "On Spectral Clustering: Analysis and an Algorithm," in *Proc. NIPS*, Dec 2001.
- [Ngo03] C.-W. Ngo, Y.-F. MA and H.-J. Zhang, "Automatic Video Summarization by Graph Modeling", in *Proceedings of the International Conference on Computer Vision (ICCV)*, vol. 1, Nice, France, 2003.
- [Nielsen07] A. B. Nielsen, S. Sigurdsson, L. K. Hansen and J. Arenas-García, "On the Relevance of Spectral Features for Instrument Classification", *Proceedings of the IEEE International Conference on Aoustics, Speech and Signal Processing (ICASSP'07)*, vol. 2, pp. 485-488, April, 2007.
- [Odobez03] J. M. Odobez, D. Gatica-Perez and M. Guillemot, "Video Shot Clustering using Spectral Methods", in *3rd Workshop on Content-Based Multimedia Indexing (CBMI)*, Rennes, France, September, 2003.
- [OPENCV] Opencv. <http://opencvlibrary.sourceforge.net>, 2007.
- [Ouyang03] J.-Q. Ouyang, J.-T. Li, and Y.-D. Zhang, "Replay Boundary Detection in MPEG Compressed Video", in *Proceedings of the Machine Learning and Cybernetics International Conference*, vol. 5, 2003.
- [Over08a] P. Over, A. F. Smeaton, and G. Awad, "The TRECVID 2008 BBC Rushes Summarization Evaluation", in *TVS'08: Proceedings of the International Workshop on TRECVID Video Summarization*, pages 1–20, ACM, 2008.
- [Over08b] P. Over, A. F. Smeaton, and G. Awad, "TRECVID 2008 - Goals, Tasks, Data, Evaluation Mechanisms and Metrics", in *Proceedings of the TRECVID'08*, 2008.
- [Paice96] C.D. Paice, "Method for Evaluation of Stemming Algorithms based on Error Counting", *JASIS*, vol. 47(8), pp. 632-649, 1996.
- [Papandreou07] G. Papandreou, A. Katsamanis, V.Pitsikalis and P.Mara, "Multimodal Fusion and Learning with Uncertain Features Applied to

- Audiovisual Speech Recognition”, *IEEE 9th Workshop on Multimedia Signal Processing*, 2007.
- [Park08] H. Park, J. Yang, J. Park, S.G. Kang and J.K. Choi, “A Survey on Peer-to-Peer Overlay Network Schemes”, *Advanced Communication Technology*, vol. 2, pp. 986-988, 2008.
- [Peyrard03] N. Peyrard and P. Bouthemy, “Motion based Selection of Relevant Video Segments for Video Summarization”, in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, Baltimore, MD, 2003.
- [Picard00] R. Picard, *Affective Computing*, The MIT Press Cambridge, 2000.
- [Porter03] S.V. Porter, M. Mirmehdi and B.T. Thomas, “Temporal Video Segmentation and Classification of Edit Effects”, *Image and Vision Computing*, vol. 21(13-14), pp. 1097--1106, December, 2003.
- [Potamianos03] G. Potamianos, C. Neti, G. Gravier, A. Garg and A.W. Senior, “Recent Advances in the Automatic Recognition of Audio-Visual Speech”, *Proceedings of the IEEE*, vol. 91, no. 9, September, 2003.
- [Qian06] X. Qian, G. Liu and R. Su, “Effective Fades and Flashlight Detection based on Accumulating Histogram Difference”, *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16(10), pp. 1245-1258, October, 2006.
- [Rice45] S.O. Rice, "Mathematical Analysis of Random Noise", *Bell System Tech. Journal*, pp. 23-24, 1945.
- [Robertson97] S.E. Robertson and K. Sparck Jones, “Simple Proven Approaches to Text Retrieval”, *Technical report TR356*, Cambridge University, Computer Laboratory, 1997.
- [Rui98] Y. Rui, T.S., Huang, M. Ortega and S. Mehrotra, “Relevance Feedback: a Power Tool for Interactive Content-based Image Retrieval”, *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 8, issue 5, pp. 644-655, September, 1998.
- [Salton88] G. Salton and C. Buckley, "Term-weighting Approaches in Automatic Text Retrieval", *Information Processing & Management*, vol. 24(5), 1988.
- [Schroeder68] M.R. Schroeder, “Period Histogram and Product Spectrum: New Methods for Fundamental Frequency Measurement”, *J. Acoust. Soc. Am.*, vol. 43, pp. 829–834, 1968.
- [Sebe02] N. Sebe, I. Cohen, A. Garg, M.S. Lew and T.S. Huang, “Emotion Recognition using a Cauchy Naive Bayes Classifier”, in *Proceedings of International Conference on Pattern Recognition*, Quebec, August, 17-20, 2002.

- [Serra08] J. Serra, E. Gomez, P. Herrera, and X. Serra, "Chroma Binary Similarity and Local Alignment Applied to Cover Song Identification" *IEEE Trans. on Audio, Speech and Language Proc.*, vol. 16, pp. 1138–1151, August, 2008.
- [Shadbolt07] N. Shadbolt, W. Hall and T. Berners-Lee, "The Semantic Web Revisited", *IEEE Intelligent Systems*, vol. 21(3), pp. 96-101, 2007.
- [Shi00] J. Shi and J. Malik, "Normalized Cuts and Image Segmentation", in *Proc. IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22(8), pp. 888–905, 2000.
- [Shi98] J. Shi, S. Belongie, T. Leung and J. Malik, "Image and Video Segmentation: The Normalized Cut Framework", in *IEEE Int'l Conf on Image Processing*, pp. 943-947, 1998.
- [Shih05] H.-C. Shih and C.-L. Huang, "Content-Based Scalable Sports Video Retrieval System", *IEEE International Symposium on Circuits and Systems*, Kobe, Japan, vol. 2, pp. 1553-1556, May 2005.
- [Sivic08] J. Sivic and A. Zisserman, "Efficient Visual Search for Objects in Video", *Proceedings of the IEEE*, vol.96, no.4, pp.548-566, April, 2008.
- [Smeaton06] A.F. Smeaton, P. Over, and W. Kraaij, "Evaluation Campaigns and TRECVID", in *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, Santa Barbara, California, USA, October 26 - 27, 2006.
- [Smeulders00] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta and R. Jain "Content-Based Image Retrieval at the End of the Early Years", in *Proc. IEEE Trans Pattern Analysis and Mach Intelligence*, vol. 22(12), pp. 1349-80, 2000.
- [Smith05] J.R. Smith, M. Campbell, M. Naphade, A. Natsev, and J. Tesic, "Learning and Classification of Semantic Concepts in Broadcast Video", in *Online Proceedings of the First International Conference on Intelligence Analysis*, McLean, VA, USA, 2005.
- [Snoek05] C.G.M. Snoek and M. Worring, "Multimodal Video Indexing: A Review of the State-of-the-Art", *Multimedia Tools and Applications*, vol. 25(1), pp. 5–35, 2005.
- [Snoek07] C.G.M. Snoek, M. Worring, A.W.M. Smeulders and B. Freiburg, "The Role of Visual Content and Style for Concert Video Indexing", *IEEE International Conference on Multimedia and Expo (CBMI'07)*, Beijing, China, 2007.
- [Song98] S. M.-H. Song, W.M. Kim, H. Kim and B.-D. Rhee, "On Detection of Gradual Scene Changes for Parsing of Video Data", in *Proceedings of Storage and Retrieval for Image and Video Databases VI*, San Jose, California, SPIE vol. 3312, pp. 404-413, January, 1998.

- [Sounders97] J. Saunders, “Real-Time Discrimination of Broadcast Speech / Music”, in *Proc. ICASSP’97*, Atlanta, GA, vol. 2, pp. 1331-1334, April 1997.
- [Srivastava07] S. Srivastava, M.R. Gupta and B.A. Frigyik, “Bayesian Quadratic Discriminant Analysis”, *Journal of Machine Learning Research*, vol. 8, pp.1277-1305, 2007.
- [Starzacher08] A. Starzacher and B. Rinner, “Evaluating KNN, LDA and QDA Classification for Embedded Online Feature Fusion”, in *Proc. of IEEE Int. Conf. on Intelligent Sensors, Sensor Networks and Information Processing*, pp.85-90, December, 2008.
- [Sull01] S. Sull, J.-R. Kim, Y. Kim, H. S. Chang, and S. U. Lee, “Scalable Hierarchical Video Summary and Search”, in *Proceedings of the SPIE Conference on Storage and Retrieval for Media Databases*, vol. 4315, pp. 553–561, 2001.
- [Sundaram02] H. Sundaram, L. Xie and S.-F. Chang, “A Utility Framework for the Automatic Generation of Audio-visual Skims”, *ACM Multimedia*, pp. 189-198, 2002.
- [Tao09] D. Tao, D. Xu and X. Li, *Semantic Mining Technologies for Multimedia Databases*, Information Science Reference, ISBN: 978-1605661889, 2009.
- [Taskiran01] C.M. Taskiran, A. Amir, D.B. Ponceleon, and E.J. Delp, “Automated Video Summarization using Speech Transcripts” in *Proceedings of the SPIE Conference on Storage and Retrieval for Media Databases*, vol. 4676, pp. 371–382, 2002.
- [Taskiran06] C.M. Taskiran, Z. Pizlo, A. Amir, D. Ponceleon and E.J. Delp, “Automated Video Program Summarization using Speech Transcripts”, *IEEE Transactions on Multimedia*, vol. 8(4), pp. 775-.791, 2006.
- [Tian08] X. Tian, L. Yang, J. Wang, Y. Yang, X. Wu and X.-S. Hua, “Bayesian Video Search Reranking”, *Proceedings of the 16th ACM international conference on Multimedia*, 2008.
- [TREC] <http://trec.nist.gov/overview.html>
- [TREC01] A.F. Smeaton, P. Over and R. Taban, “The TREC-2001 Video Track Report”, *Proc. of TREC Conference*, pp. 56, 2001.
- [TRECVID] <http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html>
- [Truong07] B.T. Truong and S. Venkatesh, “Video Abstraction: A Systematic Review and Classification”, *ACM Transactions on Multimedia, Computing, and Applications*, vol. 3, no. 1, Article 3, February, 2007.
- [Vempala00] S. Vempala, R. Kannan and A. Vetta, “On Clusterings - Good, Bad and Spectral,” in *Proc. 41st Symposium on the Foundation of Computer Science, FOCS*, 2000.

- [Viola04] P. Viola and M. Jones, "Robust Real-time Face Detection", *International Journal of Computer Vision*, vol. 57(2), pp. 137–154, 2002.
- [Wang00] Y. Wang, Z. Liu and J.-C. Huang, "Multimedia Content Analysis: Using Both Audio and Visual Clues", *IEEE Signal Processing Magazine*, vol. 17, no. 6, pp. 12-36, November, 2000.
- [Weihs07] C. Weihs, U. Ligges, F. Mörchen and D. Müllensiefen, "Classification in Music Research", *Advances in Data Analysis and Classification*, vol. 1(3), December, 2007.
- [Wu00] J. K. Wu, M. S. Kankanhalli, J.-H. Lim, and D. Hong, Perspective on Content-Based Multimedia Systems, Kluwer Academic, Hingham, MA, 2000.
- [Wu93] Z. Wu and R. Leahy, "An Optimal Graph Theoretic Approach to Data Clustering: Theory and Its Application to Image Segmentation", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 15, no. 11, pp. 1101-1113, Nov. 1993.
- [Wyse98] L. Wyse and S. Smoliar, "Toward Content Based Audio Indexing and Retrieval and a New Speaker Discrimination Technique", *Computational Auditory Scene Analysis*, Lawrence Erlbaum Associates, Inc., Mahwah, NJ, 1998.
- [Xiao06] J. Xiao, Y. Zhuang, T. Yang and F. Wu, "An Efficient Keyframe Extraction from Motion Capture Data", *Lecture Notes in Computer Science*, vol. 4035, pp.494-501, 2006.
- [Xu05] C. Xu, N.C. Maddage and X. Shao, "Automatic Music Classification and Summarization", *IEEE Trans. Speech and Audio Proc.*, vol. 13, no. 3, pp. 441-450, May 2005.
- [Yeo95] Yeo, B. and B. Liu, "Rapid Scene Change Analysis on Compressed Video", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 5, no. 6, pp. 533-544, 1995.
- [Yeung95] M. Yeung and B.-L. Leo, "Efficient Matching and Clustering of Shots", *in Proceedings of the Conference ICIP*, 338–341, 1995.
- [Yeung97] M.M. Yeung and Y. Boon-Lock, "Video Visualization for Compact Presentation and Fast Browsing of Pictorial Content", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 7, issue 5, pp. 771-785, October, 1997.
- [Yu04] X.-D. Yu, L. Wang, Q. Tian, and P. Xue, "Multi-level Video Representation with Application to Keyframe Extraction", *in Proceedings of the International Conference on Multimedia Modeling (MMM)*, pp. 117–121, 2004.

- [Yu97] J. Yu, G. Bozdagi and S. Harrington, "Feature Based Hierarchical Video Segmentation", *International Conference on Image Processing*, vol. 2, pp. 498-501, 1997.
- [Yuan07] J. Yuan, H. Wang, L. Xiao, W. Zheng, J. Li, F. Lin and B. Zhang, "A Formal Study of Shot Boundary Detection", *IEEE Trans. Circuits Syst. Video Techn.*, vol. 17(2), pp. 168-186, 2007.
- [Zabih95] R. Zabih, J. Miller and K. Mai, "A Feature-based Method for Detecting and Classifying Scene Breaks", in *Proc. ACM Multimedia '95*, San Fransisco, CA, November, pp. 189-200, 1995.
- [Zakhor90] A. Zakhor and A.V. Oppenheim, "Reconstruction of Two-Dimensional Signals from Level Crossings", *Proceedings of the IEEE*, vol. 78, no. 1, January, 1990.
- [Zang01] T. Zang and C.-C.J. Kuo "Audio Content Analysis for Online Audiovisual Data Segmentation and Classification", *IEEE Trans. Speech and Audio Proc.*, vol. 9, no. 4, pp. 441-457, May 2001.
- [Zhang03] X.-D. Zhang, T.-Y. Liu, K.-T. Lo, and J. Feng, "Dynamic Selection and Effective Compression of Keyframes for Video Abstraction" *Pattern Recogn. Lett.*, vol. 24(9-10), pp. 1523-1532, 2003.
- [Zhang93] H.J. Zhang, A. Kankanhalli and S.W. Smoliar, "Automatic Partitioning of Full Motion Video", *Multimedia Systems*, vol. 1, pp. 10-28, Jan, 1993.
- [Zhang97] H. Zhang, J. Wu, D. Zhong, and S. Smollar, "An Integrated System for Content-based Video Retrieval and Browsing", *Pattern Recogn.*, vol. 30(4), pp. 643-658, 1997.
- [Zhao03] R. Zhao and W. I. Grosky "A Novel Video Shot Detection Technique Using Color Anglogram and Latent Semantic Indexing", *ICDCS Workshops'03*, pp. 550-555, 2003.
- [Zheng08] Y. Zheng, S.Y. Neo, Y. Zhang, S. Lin and T.S. Chua, "Adaptive Multiple Feedback Strategies for Interactive Video Search", *Proceedings of the International Conference on Content-based Image and Video Retrieval*, pp. 457-464, 2008.
- [Zhenghua05] J.Y. Zhenghua, S.V.N Vishwanathan and A. Smola, "NICTA at TRECVID 2005 Shot Boundary Detection Task", in *Proceedings of TRECVID 2006 Workshop*, Gaithersburg, USA, 2005.
- [Zhuang98] Y. Zhuang, Y. Rui, T. Huang, and S. Mehrotra, "Adaptive Keyframe Extraction using Unsupervised Clustering", in *Proceedings of the International Conference on Image Processing (ICIP)*, Chicago, IL, pp. 866-870, 1998.



## Acknowledgements

Finally comes the part of the thesis that I enjoy writing the most...

Firstly, I would like to express my most sincere gratitude to Jan Biemond, both for accepting me to the ICT group and for providing guidance for my thesis all through four years. It is an honor for me to be one of Jan's last PhD students.

And Alan... It was not easy to manage a project with so many deadlines, demos and reports together with three PhD students whose main interests are having their degrees. He did it very well and always motivated us for yet another project demo. He taught me, with patience, the importance of reasoning in every step of a scientific work. His ideas, proposals and criticisms were always well-founded and to the point, which makes a discussion with him fruitful.

Of course it was not only Jan and Alan who astonished me with their intellectual level, creativity and liveliness on the 10<sup>th</sup> and 11<sup>th</sup> floors of our EWI building. Hasan comes first to mention as my roommate, my project-mate, my eternal colleague and more importantly, as my friend. Bart used to share the stress of preparing a demo or a presentation or, sometimes, an excuse(!) for the next day's project meeting with us in the same office until Philips took him from us forever. It was also nice to share the office with Gjenna for a short period of time.

Going all the way to the end of the corridor of the 10<sup>th</sup> floor, you may see our hard-working pattern recognition gurus: Bob Duin, David, Pavel, Wan-Jui. In each other's rooms, next to the coffee machine, in the elevator or in the cafeteria, you hardly ever see these guys discussing on anything but a classifier! They were one of the most important information sources for me and for the whole group. It was a great pleasure to have a coffee on the corridor or, even better, an ice-cream from the next-door gas station when the weather is good with Jacco, Jeneke, Jun Wang, Jeroen, Gineke, Marteen, Martin, Stefan, Feifei, Rogier and Yunlei.

PhD is not always hard-working, deadline stress and sleepless nights. The dart parties with Bartek and Ronald were among the most relaxing moments. It has always been fun to drop by the office of Zeki and Alper and have a long chat in the mother tongue about 'memleket meseleleri' (state of the homeland).

Also it is not possible for me to forget Peter, Omar, Richard, Theo, Jan, Mark, Ana-Iona and all other great figures in ICT group.

A special thanks to Emile Hendriks for always participating in our project meetings and sharing his ideas with us. And also to two other great brains of the ICT group, Inald Lagendijk and Marcel Reinders.

Hardly any thesis from the ICT group omits acknowledging Saskia, Anja, Robert, Ben and Hans without whose help the machine wouldn't run and this thesis is no exception.

While mentioning ‘the project’ so much, I would also like to thank to Jan Nesvadba from Philips Research, Bauke Freiburg, Ynze van Houten and all other members of the big MultimediaN project that I am proud of working for.

Another important contribution to my PhD research came from the COST292 project which provided me to collaborate with a network of researchers from all around Europe, especially Jenny Benois-Pineau and Boris Mansencal from LABRI, Université Bordeaux I; Ebroul Izquierdo, Qianni Zhang and Uros Damnjanovic from Imperial College, Selim Aksoy and Pinar Duygulu from Bilkent University and Aydın Alatan from ODTU.

For introducing me to the scientific research world and TU Delft, I would like to express my gratitude to Bülent Sankur.

Uğur, whenever I need any help you are always there, even here in the Netherlands. Also Turgay, although several thousand kilometers away, provided me with support and friendship. I finally thank to the last member of the gang, Hakan.

A considerable part of this thesis was completed after I started to work at the European Patent Office. Therefore, many thanks to my colleagues at the EPO for providing a relaxed atmosphere and especially to Claudia Mayer and Robert van der Zaal for their assistance in the translation to Dutch.

My dear Mum and Dad, I sincerely appreciate your enormous love and support which brought me to this day. And my dear sister Nida, finally I am also a doctor as you are...

Last but not least, actually the most, I thank Yasemin, my dear wife. You have contributed to this thesis at every stage, from cover design to text revision, didn’t complain when I was working in front of the computer many weekends, contrary, you always gave me your support and encouragement. More importantly, you turned my life into a wonderful adventure.

## **Curriculum Vitae**

Suphi Umut Naci was born on 5 July 1978, in Istanbul, Turkey. He completed his B.S. and M.Sc. degrees in Electrical and Electronic Engineering in Boğaziçi University in 2001 and 2004, respectively. He worked as a research engineer at GVZ Speech Technologies Company between 2001 and 2004. He pursued his PhD in Information and Communication Theory Group of Delft University of Technology between 2004 and 2008. Currently, he is working as a patent examiner at the European Patent Office. He is married and has one son.