

國立臺灣大學電機資訊學院資訊網路與多媒體研究所

博士論文

Graduate Institute of Networking and Multimedia

College of Electrical Engineering & Computer Science

National Taiwan University

Doctoral Dissertation

以物件與事件為基礎之視訊內容調適架構

A Semantic Framework for Object-Based and Event-Based  
Video Content Adaptation

鄭文皇

Wen-Huang Cheng

指導教授：吳家麟 博士

Advisor: Ja-Ling Wu, Ph.D.

中華民國 97 年 6 月

June, 2008



國立臺灣大學博士學位論文  
口試委員會審定書

以物件與事件為基礎之視訊內容調適架構  
A Semantic Framework for Object-Based and Event-Based  
Video Content Adaptation

本論文係鄭文皇君（學號 D93944001）在國立臺灣大學資訊網路與多媒體研究所完成之博士學位論文，於民國九十七年六月五日承下列考試委員審查通過及口試及格，特此證明

口試委員：

吳宗麟

（簽名）

（指導教授）

陳良弼

杭學鳴

李林山

李素瑛

許永英

李素瑛

所長：

洪一平

（簽名）





To my family and teachers



## Acknowledgements

I would like to thank all those who have contributed to my education as well as to my life.

First and foremost, I would like to express my sincere gratitude to my advisor, Prof. Ja-Ling Wu, for his great support and constant encouragement throughout my dissertation. One of the most valuable things I learned from him is “how to do research”, meanwhile “being open-minded to other’s work”. That, in a word, is to “stay hungry, stay foolish”. His insightful guidance and passion for knowledge really make me enjoy the fun of finding and solving scientific problems.

I would also like to thank the other members in my dissertation committee — Profs. Lin-Shan Lee, Arbee L.P. Chen, Hsueh-Ming Hang, Jane Yung-Jen Hsu, Suh-Yin Lee, and Mark Liao. I would like to extend my thanks as well to the other professors who were in the committee of my proposal defense — Profs. Yi-Ping Hung, Wen-Chin Chen, Yung-Yu Chuang, and Winston H. Hsu.

Many thanks to Dr. Chun-Hsiang Huang and Dr. Wei-Ta Chu being my good research consultants at lab. They always have some good ideas to enlighten me when I get stuck in my research.

I would also like to thank colleagues with whom I had the pleasure to work during my doctorate program — Sung-Wen Wang, Chia-Hu Chang, Min-Chun Tien, Jyh-Ren Shieh, Ching-Ju Lin, Junn-Yen Hu, Chih-Cheng Hsu, Yi-Hon Hsiao, Ming-Fang Weng, Tz-Huan Huang, Kuan-Ting Chen, Ken-Yi Li, Edward Shen, Yun Chung Shen, and Wan-Chun Ma.

Also, thanks to members of my own research group for their support — Chi-Chang Hsieh, Ping-Chieh Chang, Hong Ming Chen, Yang-Ting Yeh, Chih-Yu Yan, Yen-Lin Huang, Heng-Yi Lin, Yi-Tang Wang, Ping-Yen Hsieh, Chen-Wei Chou, Po-Wei Chen, Kuei-Yi Hsieh, and Ming-Hsiu Chang.

I would also like to thank my lab and school assistants —Ya-Ling Chen, Yin-Tzu Lin, Yi-Chia Lai, Jun-Cheng Chen, and I-Chun Lai, for their help to deal with all kinds of life and school stuff. Especially, I deeply appreciate Ya-Ling's always timely reminder for me so as not to miss any important deadline of which I should be aware.

I would like to give the special thanks to my dear father and mother, Jui-Yuan Cheng and Pi-Yu Chen, for their being always by my side. What my mom has told me becomes my maxim to keep in mind, "Life is just like an endless competition and the success belongs to who keeps the enthusiasm and fighting spirit from start to finish." Finally, Chun-Yen, thank you, my love. You're always lighting up my heart with the things you do and say. I feel so happy just being with you.





# Curriculum Vita

## Wen-Huang Cheng

### Education

- 2008 Doctor of Philosophy, Graduate Institute of Networking and Multimedia, National Taiwan University.
- 2004 Master of Science, Department of Computer Science and Information Engineering, National Taiwan University.
- 2002 Bachelor of Science, Department of Computer Science and Information Engineering, National Taiwan University.

### Experience

- 2007 Summer Visiting Student, University of Tokyo, Tokyo, Japan.
- 2007 Research Intern, IBM T.J. Watson Research Center, Hawthorne, NY, USA.

### Honors

- 2008 Member of the Phi Tau Phi Scholastic Honor Society, Taiwan, R.O.C.
- 2006 Excellent Work Award, work title: “Expand your vision: glasses-based multimedia information communication platform”, the 1st Acer Long-Term Smile Contest, Taiwan, R.O.C., 2006. (only 5 among 147 works are awarded)
- 2006 Excellent Work Award, work title: “Life Linker - smart digital photo frame”, the 1st Acer Long-Term Smile Contest, Taiwan, R.O.C., 2006. (only 5 among 147 works are awarded)
- 2005 Best Paper Award, IPPR Conference on Computer Vision, Graphics and Image Processing (CVGIP), Taipei, Taiwan, R.O.C.
- 2004 Best Student Paper Award, Workshop on Consumer Electronics and Signal Processing (WCEsp), Hsinchu, Taiwan, R.O.C.

### Selected Publications – Journal

- J1 Wen-Huang Cheng, Chia-Wei Wang, and Ja-Ling Wu, “Video adaptation for small display based on content recomposition,” *IEEE Trans. Circuits and Systems for Video Technology*, vol. 17, no. 1, pp. 43-58, January 2007.

- J2 Chia-Chiang Ho, Ja-Ling Wu, and Wen-Huang Cheng, "A practical foveation-based rate shaping mechanism for MPEG videos," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 15, no. 11, pp. 1365-1372, November 2005.
- J3 Wen-Huang Cheng, Wei-Ta Chu, and Ja-Ling Wu, "A visual attention based region-of-interest determination framework for video sequences" *IEICE Trans. Information and Systems Journal*, vol. E-88D, no. 7, pp. 1578-1586, July 2005.
- J4 Wei-Ta Chu, Wen-Huang Cheng, and Ja-Ling Wu, "Semantic context detection using audio event fusion," *EURASIP Journal on Applied Signal Processing*, 2005.
- J5 Wei-Ta Chu, Wen-Huang Cheng, Jane Yung-Jen Hsu, and Ja-Ling Wu, "Towards semantic indexing and retrieval using hierarchical audio models," *ACM Multimedia Systems*, vol. 10, no. 6, pp. 570-583, 2005.

#### Selected Publications – Conference

- C1 Wen-Huang Cheng and David Gotz, "Context-based page unit recommendation for web-based sensemaking tasks," *Proc. Intl. World Wide Web Conf. (WWW'08)*, pp. 1073-1074, 2008.
- C2 Chi-Chang Hsieh, Wen-Huang Cheng, Chia-Hu Chang, Yung-Yu Chuang, and Ja-Ling Wu, "Photo navigator," *Proc. ACM Intl. Conf. Multimedia (MM'08)*, 2008.
- C3 Wen-Huang Cheng *et al.*, "Semantic-event based analysis and segmentation of wedding ceremony videos" *Proc. ACM Intl. Workshop on Multimedia Information Retrieval (MIR'07)*, pp. 95-104, 2007.
- C4 Chia-Wei Wang, Wen-Huang Cheng, Jun-Cheng Chen, Shu-Sian Yang, and Ja-Ling Wu, "Film narrative exploration through analyzing aesthetic elements," *Proc. Intl. MultiMedia Modeling Conf. (MMM'07)*, 2007.
- C5 Wen-Huang Cheng, Chun-Wei Hsieh, Sheng-Kai Lin, Chia-Wei Wang, and Ja-Ling Wu, "Robust algorithm for exemplar-based image inpainting," *Proc. Intl. Conf. Computer Graphics, Imaging and Vision (CGIV'05)*, pp. 64-69, 2005.
- C6 Wen-Huang Cheng, Wei-Ta Chu, Jin-Hau Kuo, and Ja-Ling Wu, "Automatic video region-of-interest determination based on user attention model," *Proc. IEEE Intl. Symposium on Circuits and Systems (ISCAS'05)*, pp. 3219-3222, 2005.

- C7 Wei-Ta Chu, Wen-Huang Cheng, and Ja-Ling Wu, “Generative and discriminative modeling toward semantic context detection in audio tracks,” *Proc. Intl. MultiMedia Modeling Conf. (MMM’05)*, pp. 38-45, 2005.
- C8 Wei-Ta Chu, Wen-Huang Cheng, Ja-Ling Wu, and Jane Yung-Jen Hsu, “A study of semantic context detection by using SVM and GMM approaches,” *Proc. IEEE Intl. Conf. Multimedia and Expo (ICME’04)*, pp. 1591-1594, 2004.
- C9 Wen-Huang Cheng, Wei-Ta Chu, and Ja-Ling Wu, “Semantic context detection based on hierarchical audio models,” *Proc. ACM Intl. Workshop on Multimedia Information Retrieval (MIR’03)*, pp. 109-115, 2003.

#### Patent

- P1 Wen-Huang Cheng and David Gotz, “Context-based document unit recommendation for sensemaking tasks”, US Patent, Application Number YOR920080235US1, 2008.





## Abstract

# A Semantic Framework for Object-Based and Event-Based Video Content Adaptation

Wen-Huang Cheng

In pervasive media environments, adaptation is one key technology to support universal multimedia access by transforming multimedia contents to fit the usage environments. In terms of personalization, effective adaptation can greatly benefit from taking into account the semantics of multimedia contents. The goal of this dissertation is to be able to provide systematic approaches to improve automatic multimedia adaptation at the semantic level.

In this dissertation, a generic adaptation framework and the fundamental design principles are proposed. By exploiting specific domain knowledge, we bridge the gap between low-level computational features and high-level semantic concepts, whereby the associated adapting operations can be effectively designed to maximize the user's multimedia experience. Based on the proposed framework, our works focus on the semantic adaptation of video contents, where two alternative approaches for semantics modeling are investigated: the object-based and the event-based. In the object-based approach, a visual model is constructed for locating semantic video objects so as to improve the user's browsing experience of high-quality professional videos on the devices with small displays. In the event-based approach, both the visual and aural information are exploited to characterize semantic video events that can be used to benefit the user's navigation in hours-long home videos. The two systems can be viewed as the technical realization of the proposed adaptation framework and demonstrate the effectiveness of automatic high-level semantics analysis.

## 中文摘要

### 以物件與事件為基礎之視訊內容調適架構

鄭文皇

在普及媒體環境中，內容調適是用以實現普遍多媒體存取的一種關鍵技術。具體而言，其藉由多媒體內容的轉換以使轉換後之多媒體內容符合相對應的使用環境。從個人化應用的角度來看，有效的內容調適可得益於對多媒體內容語意的深刻理解。因此，本論文的目標即在於提供一套具系統化之研究方法以提昇自動化多媒體調適的語意層次。

在本論文中，我們提出一個通用型之調適架構以及相對應之基本設計原則。藉由導入特定領域知識，我們適度跨越存在於低階可計算特徵值與高階語意概念間之語意鴻溝，並藉此有效開發與其所屬之調適運算以求得使用者多媒體經驗之最佳化。在前述所提出的架構之上，我們的研究聚焦於視訊內容之語意調適，其中具體探討兩種用於語意模型化之方法，分別是以物件為基礎與以事件為基礎之方法。在以物件為基礎之方法中，我們建構一個可用於定位視訊中具語意性物件之視覺模型，以提昇使用者在小螢幕行動裝置上觀賞高畫質專業影片時之瀏覽經驗。另一方面，在以事件為基礎之方法中，我們同時利用視訊中之視覺與聽覺資訊，以描繪具語意性事件之多媒體特性，並應用於滿足使用者對於長時間家庭影片之實際瀏覽需要。此兩個系統可視為前述所提出調適架構之具體技術實現，並可藉此顯現自動化高階語意分析之可行性與有效性。

# Contents

<b>Acknowledgements</b>	<b>v</b>
<b>Curriculum Vita</b>	<b>vii</b>
<b>Abstract</b>	<b>xi</b>
<b>List of Figures</b>	<b>xvi</b>
<b>List of Tables</b>	<b>xix</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	1
1.2 Semantic Multimedia Content Adaptation . . . . .	4
1.2.1 A Generic Framework . . . . .	4
1.2.2 From Signal to Semantic Levels . . . . .	8
1.2.3 Adaptive Optimization . . . . .	12
1.3 Problem Statement . . . . .	14
1.4 Summary of Contributions . . . . .	15
1.4.1 Framework Development for Semantic Adaptation . . . . .	16
1.4.2 Video Adaptation Based on Semantic Objects . . . . .	16
1.4.3 Video Adaptation Based on Semantic Events . . . . .	17
1.5 Organization of the Dissertation . . . . .	17
<b>2 Basics and Literature Review</b>	<b>19</b>
2.1 Semantic Concept Ontology . . . . .	19
2.1.1 Typical Examples . . . . .	20
2.1.2 Relationship Building . . . . .	23
2.2 Semantic Concept Analysis . . . . .	24
2.3 Semantic Content Adaptation . . . . .	26
2.3.1 Adaptation Taxonomy . . . . .	26

2.3.2	Adaptation Strategy . . . . .	28
2.4	Framework Correspondence . . . . .	29
2.4.1	Semantic Object Based Video Adaptation . . . . .	29
2.4.2	Semantic Event Based Video Adaptation . . . . .	30
2.5	Summary . . . . .	31
<b>3</b>	<b>Semantic Object Based Video Adaptation</b>	<b>33</b>
3.1	Introduction . . . . .	34
3.2	Related Work . . . . .	37
3.3	User-Interest Finding . . . . .	41
3.3.1	Visual Attention Modeling . . . . .	41
3.3.2	Video ROIs Determination . . . . .	47
3.4	Content Recomposition . . . . .	51
3.4.1	UIOs Extraction . . . . .	53
3.4.2	Background Repairing . . . . .	54
3.4.3	Media Aesthetics Based Video Objects Reintegration . . . . .	55
3.5	Experimental Results . . . . .	60
3.5.1	Recomposition Results . . . . .	61
3.5.2	User Studies . . . . .	65
3.5.3	Time Efficiency Analysis . . . . .	74
3.6	Summary . . . . .	75
<b>4</b>	<b>Semantic Event Based Video Adaptation</b>	<b>83</b>
4.1	Introduction . . . . .	84
4.2	Related Work . . . . .	87
4.3	Wedding Event Taxonomy . . . . .	90
4.4	Event Features Development and Extraction . . . . .	92
4.4.1	Key Observations . . . . .	92
4.4.2	Selected Features for Event Modeling . . . . .	96
4.5	Wedding Modeling . . . . .	107
4.5.1	Wedding Event Modeling . . . . .	109
4.5.2	Event Transition Modeling . . . . .	112
4.5.3	Wedding Segmentation Using HMM . . . . .	113
4.6	Experimental Results . . . . .	115
4.6.1	Event Recognition Analysis . . . . .	116
4.6.2	Video Segmentation Analysis . . . . .	119
4.6.3	Performance Comparisons with LCRF Models . . . . .	122
4.6.4	Extension to the Scenario with Known Event Ordering . . . . .	124
4.7	Summary . . . . .	126



<b>5</b>	<b>Conclusions and Future Work</b>	<b>133</b>
5.1	Conclusions . . . . .	133
5.2	Future Research . . . . .	135
	<b>Bibliography</b>	<b>137</b>



# List of Figures

1.1 Concept of Universal Multimedia Access (UMA): Enabling interoperable and transparent access to rich multimedia contents over various usage environments. . . . .	2
1.2 Two types of adaptation engines: (a) content blind and (b) content aware but without the ability of semantics extraction. . . . .	6
1.3 The proposed generic framework for multimedia content adaptation, which is characterized by the capability of active content analysis and the use of domain knowledge to enable automatic adaptation at the semantic level. . . . .	7
1.4 Overview of the abstraction levels of computational media descriptions. . . . .	8
1.5 The Chinese classical painting “Listening to the Qin” (partial), by Ji Zhao (Emperor Huizong of the Song Dynasty, China), 1082-1135. . .	10
1.6 Sample media descriptions of the painting in Figure 1.5, given at two abstraction levels of (a) audiovisual features and (b) perceptual arousals. . . . .	11
1.7 Interrelationship between the key elements (adaptation, resource, utility) of adaptive optimization in searching for the optimal adapting operation that maximizes the adaptation utility of a multimedia content. (Figure adapted from [Cha02]) . . . . .	12
2.1 Illustrations of 101 TRECVID concepts. (Figure excerpted from [SWvG <sup>+</sup> 06]) . . . . .	22
2.2 Example of a tennis match with (a) a video snapshot and (b) the types of tennis events. (Figure excerpted from [TWC <sup>+</sup> 08]) . . . . .	22
2.3 Adaptation taxonomy with examples of the corresponding adapting operations. . . . .	27
2.4 The work of semantic object based video adaptation is represented by the proposed generic framework given in Figure 1.3. . . . .	29

2.5 The work of semantic event based video adaptation is represented by the proposed generic framework given in Figure 1.3. . . . .	30
3.1 Flowchart of the proposed framework for conducting video adaptation. . . . .	36
3.2 Examples of semantic distortion in adapted videos: (a) and (c) are two original frames from the classical film “ <i>Lawrence of Arabia</i> ”, and (b) and (d) are the corresponding adapted results using [fil], respectively. With partial coverage, the two men of (a) no longer look at each other’s eyes when they are chatting in (b), and the man in (d) seems more like to burn himself with the burning match rather than just hold it in (c). (Courtesy of FlikFX Ltd.) . . . . .	39
3.3 Example of feature maps: (a) original video frame, (b) intensity, (c) red-green color, (d) blue-yellow color, (e) x-motion, and (f) y-motion feature maps. . . . .	43
3.4 Examples of a video frame with (a) one and (d) two ROIs (indicated by the white squares); (b) and (e) are the corresponding saliency maps, and (c) and (f) are the 3-D profiles of the saliency maps of (a) and (d), respectively. . . . .	45
3.5 Comparisons of the ROI and the UIO representations for user-interests. They are respectively indicated by the solid and dotted lines. In (a) and (b), the number of contained semantic objects (man together with a car versus one single car) is different. . . . .	49
3.6 Example of flooding operations with a $6 \times 5$ ROI. In (a), the number of each pixel indicates which border it belongs to. In (b), the left and the right pixels of a thick solid line are marked as the background and UIO, respectively. Their valid neighbors are connected with the arrows. (Let $C_i$ be a color in RGB space and $d_\theta(C_2, C_3) > T_d$ .) . . . . .	50
3.7 Examples of video objects separation. The columns from left to right are successively the original frames with ROIs, extracted UIOs, and repaired backgrounds. . . . .	52
3.8 The virtual 3-D scene model. All video objects of a frame are re-projected onto a target screen. An object is perceived larger (i.e., the star-shaped UIO) while it comes closer to the screen. . . . .	55
3.9 Comparison of our approach with the conventional approach (direct-resizing) for the clips of subgroup 1. . . . .	62
3.10 Comparison of our approach with the conventional approaches (direct-resizing and linear-resizing) for the clips of subgroup 2. . . . .	63

3.11 Example of the displayed web page for TRIAL-I. For reality, both the pair of testing clips are presented on a virtual cellular phone. (See Subsection 3.5.2 for details.) . . . . .	66
3.12 Comparison of the user study between our approach and the conventional approach (direct-resizing) for the clips of subgroup 1 at different resolution formats. . . . .	72
3.13 Comparison of the user study between our approach and the conventional approaches (direct-resizing and linear-resizing) for the clips of subgroup 2 at different resolution formats. . . . .	73
3.14 Failure examples of our approach. The columns from left to right are successively the original frames with ROIs, extracted UIOs, and recomposed frames. . . . .	78
4.1 Sample key-frames of the thirteen wedding events. . . . .	91
4.2 Example of a music signal with (a) its spectrogram using short-time Fourier transform and (b) its corresponding line map. . . . .	97
4.3 Classification results of the audio types of speech (the left subplot) and music (the right subplot) on three audio datasets of (a) Internet radio, (b) Internet radio with added white noises (5 dB), and (c) audio tracks from home videos, using a multi-class SVM classifier built upon the three audio features proposed in Section 4.4.2. . . . .	101
4.4 Examples of (a) two power spectrums of a wedding audio from consecutive time instances, one with applause (the top solid curve) and another without applause (the bottom dotted curve), and (b) a sigmoidal filter function. . . . .	102
4.5 Precision-recall curves of the applause detection results using two different thresholds. (See Section 4.4.2 for details.) . . . . .	103
4.6 Examples of (a) a video frame with (b) the thresholded image and (c) the bridal white map with projection histograms. . . . .	105
4.7 Precision-recall curves of the bride indication results. (See Section 4.4.2 for details.) . . . . .	108
4.8 Examples of wedding event models of (a) the <i>RE</i> event and (b) the <i>WK</i> event. . . . .	111
4.9 A simplified example of the HMM for wedding segmentation. (See Subsection 4.5.3 for details.) . . . . .	114
4.10 Edit operations for transforming (a) a reference event string to (b) the one for comparison. . . . .	120
4.11 (a) A sample wedding program accompanied with the transcribed event ordering, and (b) the state diagram in form of a Markov chain built according to the above event ordering. . . . .	125

# List of Tables

3.1	Weights for the feature maps under different camera motion types.	46
3.2	Screen sizes used in the experiments.	60
3.3	Source clips used in the experiments.	79
3.4	Test conditions of the user studies. (See Subsection 3.5.2 for details.)	79
3.5	User study of the relative preference (RP) of our approach with regard to the conventional approaches.	80
3.6	Time efficiency analysis of the proposed framework for recomposing a $320 \times 240$ video frame.	81
4.1	Taxonomy of wedding events	90
4.2	The tendency of wedding events in their behavior of speech/music types, applause activities, picture-taking activities, and leading roles (from the second to the fifth columns, respectively).*	93
4.3	Examples of flash distributions of four successive wedding events in a ceremony.*	94
4.4	The collection of six wedding videos used in our experiments.	104
4.5	An even transition model of the wedding events.	113
4.6	The statistics of means $\mu$ and variances $\sigma^2$ of event duration for each of the event categories in our video collection (unit: seconds).	117
4.7	The recognition results of all wedding events (unit: seconds).	128
4.8	The recognition results solely based on the feature similarity of wedding events without exploiting the event transition modeling.	129
4.9	The segmentation results without duration-based filtering (unit: event segments).	129
4.10	The segmentation results with duration-based filtering (unit: event segments).	129
4.11	The percentage of total event duration for each of the event categories in our video collection.	130
4.12	LCRF recognition results of all wedding events (unit: seconds).	131

4.13 LCRF segmentation results without duration-based filtering (unit: event segments). . . . .	132
4.14 LCRF segmentation results with duration-based filtering (unit: event segments). . . . .	132
4.15 Segmentation results in the case when event orderings are available (unit: event segments). . . . .	132
4.16 The second-based recognition rate of wedding events for all clips in our video collection. . . . .	132



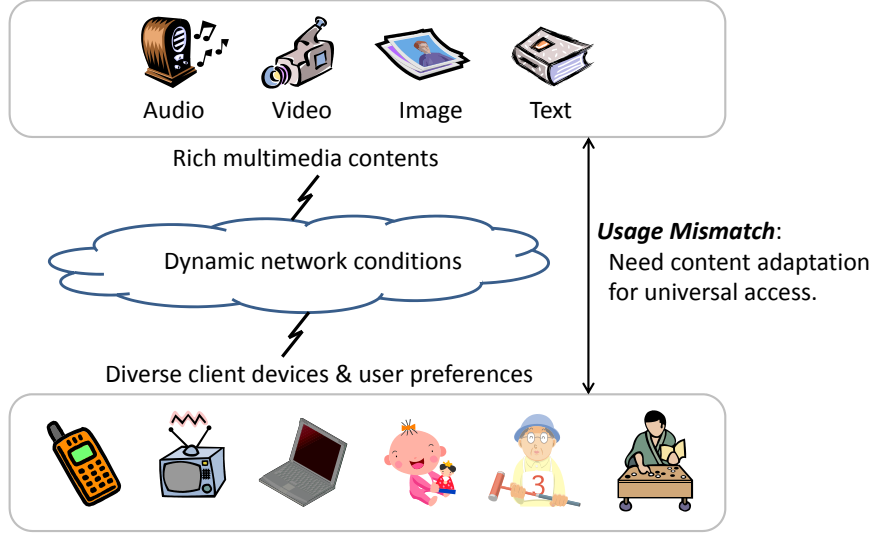
# Chapter 1

## Introduction

### 1.1 Motivation

In recent years, multimedia has brought people a new cultural revolution and became an indispensable part of our daily lives. Rapid advances in multimedia technology speed up the creation and the distribution of multimedia contents. One evidence is that people are now not only passive content consumers but also active content contributors. On the Internet, for example, they are able to acquire various kinds of multimedia information, such as music [yah], news [cnn], live sports [nba], and movie trailers [net]. Meanwhile, they can also create their own personal contents and effortlessly share with the public through social networking websites, such as YouTube [you], Flickr [fli], Facebook [fac], and MySpace [mys]. The explosive blossom of multimedia contents constructs a ubiquitous media environment and hastens the growth of innovative multimedia services and applications.

The multimedia development also drives people's desire to stay connected with the world, regardless from everywhere, at anytime, and along with any devices, networks, and preferences. For instance, regular commuters on public transporta-



**Figure 1.1:** Concept of Universal Multimedia Access (UMA): Enabling interoperable and transparent access to rich multimedia contents over various usage environments.

tion are pleased to spend time by watching broadcast TV programs through their mobile phones. In addition, homesick students may feel comfort by viewing stored photos, videos, or even voice of the loved family members from a remote home portal. However, many technical obstacles need to be overcome before we can make it all possible. As shown in Figure 1.1, the increasing variety of usage environments complicates the content delivery path and leads to a growing mismatch with the rich set of multimedia sources [BPdWK06, RJ05]. This poses great technical challenges for enabling the Universal Multimedia Access (UMA) [PB03, MSL99, BGP03], i.e., to achieve interoperable and transparent access to multimedia contents.

*Adaptation* generally refers to the technology that supports UMA by either adapting the multimedia content to fit a usage environment or adapting the usage environment to accommodate the content [BPdWK06, PB03, CV05]. Since the



usage environment is usually inflexible and hard to change, the research society mainly focuses on adapting the content. An intuitive solution is to have multiple versions of a content in advance (e.g. multi-version coding [CAL96]). The method is simple but its drawbacks are also obvious. It requires more storage space and is difficult (usually impossible) to offer an adapted version for every possible usage environment. Therefore, a better and widely-accepted idea is to adapt the content during delivery, depending on the user's actual requests and situation [BPdWK06, PB03, vBSE<sup>+</sup>03]. This is accomplished by appending adaptation hints to individual content. That is, descriptive metadata, namely media descriptions, are used to identify the content characteristics so as to aid in the process of making adaptation decisions under usage constraints [vBSE<sup>+</sup>03, CSP01, TLS04]. The media descriptions can be defined from different abstraction levels of low-level features to high-level semantic concepts, which relate directly to the semantic level of feasible adaptation choices associating with the content. For example, high-level descriptions (e.g. what subjects are present in a video scene) can help to semantically satisfy the user's personal interests [TLS04]. By contrast, low-level descriptions, such as the spatial resolution hint [CSP01], can only be employed to fit the user's physical constraints on the display size.

Therefore, in terms of personalization, effective adaptation can greatly benefit from taking into account the semantics of multimedia contents. Much research in the last few years has been conducted to reach this goal, but how to efficiently extract the semantics is still the hardest bottleneck [NH02, RH99, Chu06]. There exists a huge gap between the rich meaning and interpretation that humans could read in the content and the simplicity of low-level features that the current algorithms can actually compute [NH02, SWS<sup>+</sup>00, DV03, FLE04]. This makes the advance in multimedia adaptation be lagging far behind the user's expectations

and becomes an open problem. Our work, as motivated by the above observations, focuses on developing systematic approaches to improve automatic multimedia adaptation at the semantic level. By exploiting specific domain knowledge, we bridge the gap between low-level computational features and high-level semantic concepts, whereby the associated adapting operations can be effectively designed to maximize the user's multimedia experience. Our work attempts to enable one more step towards the development of truly semantic multimedia systems and expects to inspire more pioneering researches to march forward further.

## 1.2 Semantic Multimedia Content Adaptation

### 1.2.1 A Generic Framework

Adaptation engine is a technical realization of the adapting functionality that transforms multimedia contents in order to satisfy the usage constraints, such as device capabilities, network characteristics, and user preferences. Practical examples include the tools for format transcoding [AWSZ05], speech transcription [LL01], image mosaicing [RH99], and video summarization [BMM99]. In this section, a number of design requirements for an effective adaptation engine are first discussed. A generic framework for multimedia content adaptation is then proposed.

- **Requirement 1: Application Awareness.**

The basic requirement for an adaptation engine is *application awareness*. The adaptation engine should be first aware of specified usage constraints associated with the target applications and then be able to properly adapt the contents. For example, when delivering high-quality videos onto the

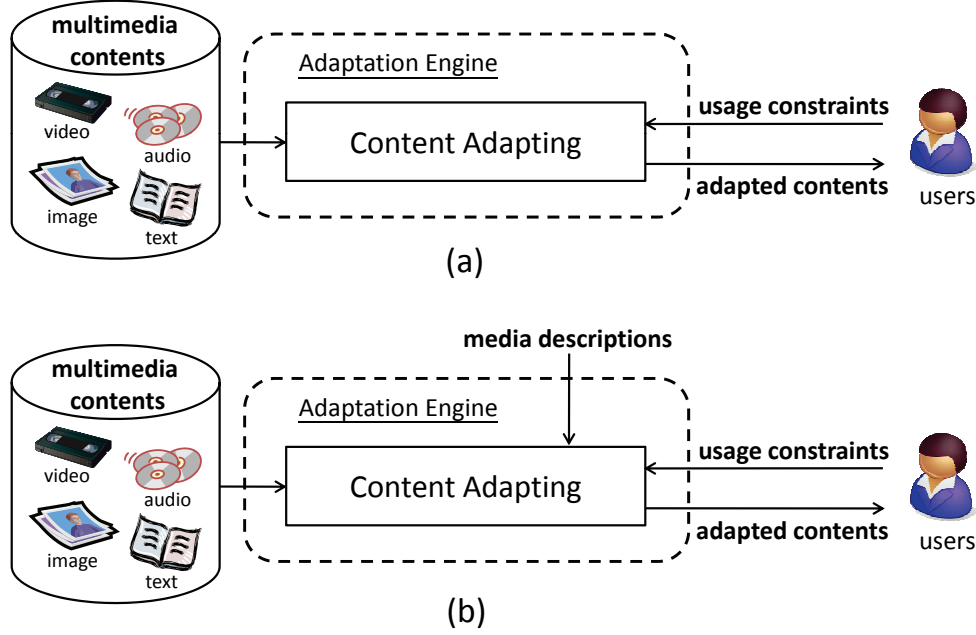
user's mobile phone, the information about the device capabilities has to be specified for converting the videos into affordable coding formats, such as the spatial resolution, bitrate, and frame rate [AWSZ05]. To facilitate the information exchange in between, several international bodies have recently developed a set of standardized description tools to detail the characteristics of networks, users, and terminals, e.g. the Usage Environment Description (UED) tool in MPEG-21 [BPdWK06].

- **Requirement 2: Content Awareness.**

The second and arguably the most important requirement is *content awareness*. Users are the final content consumers and what they are really interested in is the appearances presented in the contents. An adaptation engine can well satisfy the user preferences only if it is aware of the contents to some extent. The use of media descriptions is a common technique to describe information about or present in the content [vBSE<sup>+</sup>03, CSP01]. The information can be obtained either from automatic content understanding or previously computed metadata. The media descriptions are then employed to guide the adaptation process. For example, instead of constant temporal subsampling, the descriptions of highlight index help to summarize sports videos in a more meaningful way [BKOK04]. Some standardized media descriptions can be found in the MPEG-7 standard [CSP01], such as the Description Schemes (DSs).

- **Requirement 3: Semantics Extraction.**

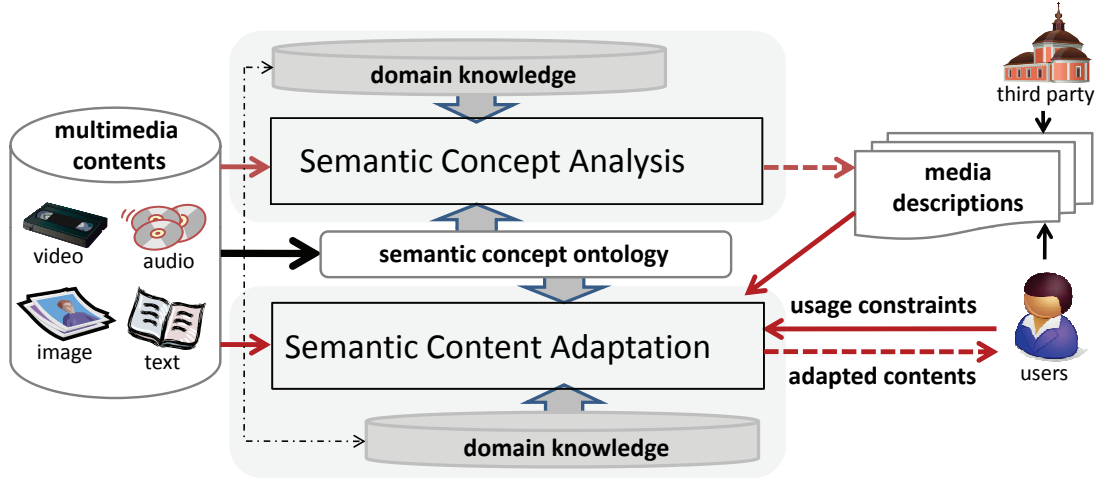
The last desirable requirement is *semantics extraction*. It means the capability of being able to actively extracting the semantic-level information from the contents. That is, pre-computed media descriptions are not always



**Figure 1.2:** Two types of adaptation engines: (a) content blind and (b) content aware but without the ability of semantics extraction.

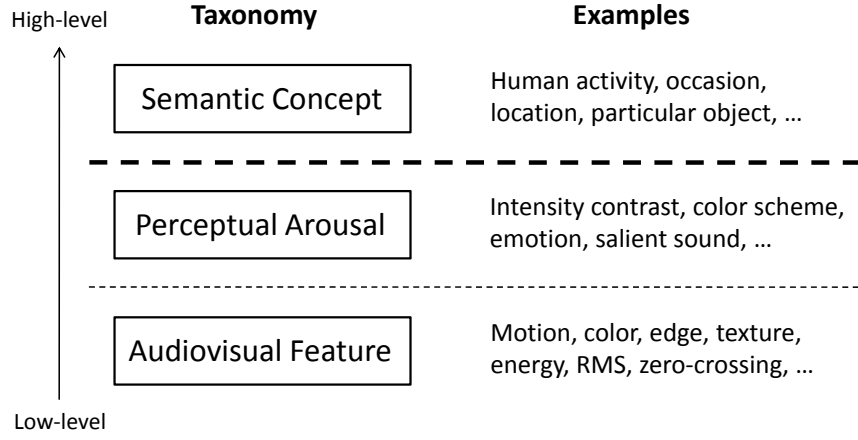
available and multimedia contents often can not be accessed until the time to be adapted. An adaptation engine should possess the capability to dynamically analyze and in a sense to understand the contents [CSE05, Her07]. Since in practice, it is difficult to know what information should be discovered and how it can be used in the decision-making of adaptation, certain domain knowledge is generally required to benefit the analysis process. For example, by exploiting the theories of media aesthetics, the basic story units can be extracted from a movie [WCC<sup>+</sup>07].

Overall, the above criteria determine the essential capabilities of a general adaptation engine. Failing to satisfy any of the design requirements will make it suffer from a loss of generality. Two types of examples are illustrated in Fig-



**Figure 1.3:** The proposed generic framework for multimedia content adaptation, which is characterized by the capability of active content analysis and the use of domain knowledge to enable automatic adaptation at the semantic level.

Figure 1.2. The first type, in Figure 1.2(a), is content blind such that it is aware of only the user's physical constraints, e.g. the format transcoder [AWSZ05] and the coding rate shaper [Ho03]. Another type, in Figure 1.2(b), is content aware but the content information depends on the provision of external algorithms. It would sacrifice some adapting functionality if the extra information is unavailable. Therefore, based on the previous discussions, the proposed generic framework for multimedia content adaptation is illustrated in Figure 1.3, which is composed of two functional modules including: the semantic concept analysis module and the semantic content adaptation module. The semantic concept analysis module analyzes multimedia contents using specific domain knowledge and generates the corresponding media descriptions to identify the content characteristics. Based on both the media descriptions and the user's usage constraints, the semantic content adaptation module then makes adaptation decisions and performs the adaptation



**Figure 1.4:** Overview of the abstraction levels of computational media descriptions.

on the multimedia contents. The construction details of the proposed framework will be discussed in Chapter 2.

### 1.2.2 From Signal to Semantic Levels

Media descriptions are descriptive metadata used to annotate multimedia contents, which can be encoded with various attributes of contents, such as the spectrogram of an audio, the emotion of an image, and the captured events of a video [vBSE<sup>+</sup>03, CSP01, TLS04]. The supporting level of an adaptation engine in terms of the user preference is determined by the quality of obtainable media descriptions. To clarify the position of our work, this section first gives definitions of the computational media descriptions from different abstraction levels of low-level features to high-level concepts, and then introduces the proposed notion of semantic adaptation.

- **Audiovisual Features:** Audiovisual features are the measurable physical properties of the multimedia signals being observed [Chu06, Dje02]. They are directly derivable from the multimedia contents and do not need to be interpreted with any human meaning [Dje02]. Some commonly used features include motion (for video), color, edge, texture (for both image and video), energy, root-mean-square, and zero-crossing (for audio) [WLH00]. For the simplicity of extraction, most of the presented adaptation engines in the literature are built on the feature domain, involving feature matching, clustering, and modeling. Typical examples include the systems of content-based image retrieval [SWS<sup>+</sup>00] and scene-based video summarization [CV05].
- **Perceptual Arousals:** Perceptual arousals are the multimedia patterns that might lack common or objective definitions in human meaning but tend to arouse the user's attention, feeling, or emotion [Dje02, HX05]. The arousing patterns can be viewed as the ones that are either intended to be formed by the authors or naturally perceived by the majority of users. Some examples include the speech sound in audio [MLZL02], the color arrangement of an artistic image [LG04], and the affective plays of a movie [WC06]. Instead of truly content understanding, the extraction of perceptual arousals is a feasible compromise in developing personalized multimedia applications, such as the attention-based detection of sports highlights [BKOK04] and the emotion-based movie indexing and summarization [HX05, WC06].
- **Semantic Concepts:** Semantic concepts are entities that take place or exist in time and space in the world, including activities (e.g. skiing, dancing), occasions (e.g. wedding, birthday), locations (e.g. beach, park), and particular objects in the scene (e.g. actor, tree) [Chu06, CEJ<sup>+</sup>07]. Several entities are able to jointly constitute a composite entity. For instance, a

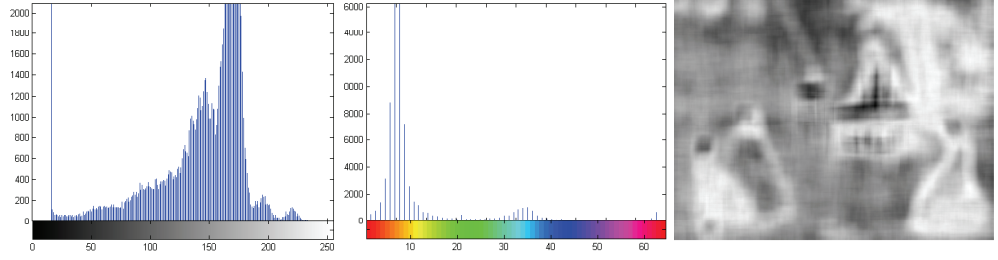


**Figure 1.5:** The Chinese classical painting “Listening to the Qin” (partial), by Ji Zhao (Emperor Huizong of the Song Dynasty, China), 1082-1135.

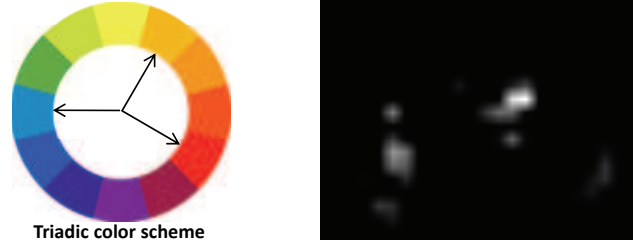
“wedding” entity is formed by the “groom”, “bride”, “officiant”, “guests”, and “church” entities. Another characteristic of the semantic concepts is context dependent. The same semantic concept can convey different human meaning if the background contexts are different. For example, the concept of “people together singing a birthday song” means the wishes from one’s friends in the case of a birthday party, but it may simply indicate one of the many performances within a group singing contest. In dealing with the analysis of semantic concepts, the associated context can provide the knowledge base to infer and identify the actual semantic meanings.

Overall, the three abstraction levels of computational media descriptions are illustrated in Figure 1.4. The abstraction levels from low to high correspond to the expressiveness of media descriptions about the “fact” contained in multimedia contents. Consider the painting shown in Figure 1.5. At the lowest level of audiovisual features, the painting can be described as an M-by-N digitized im-





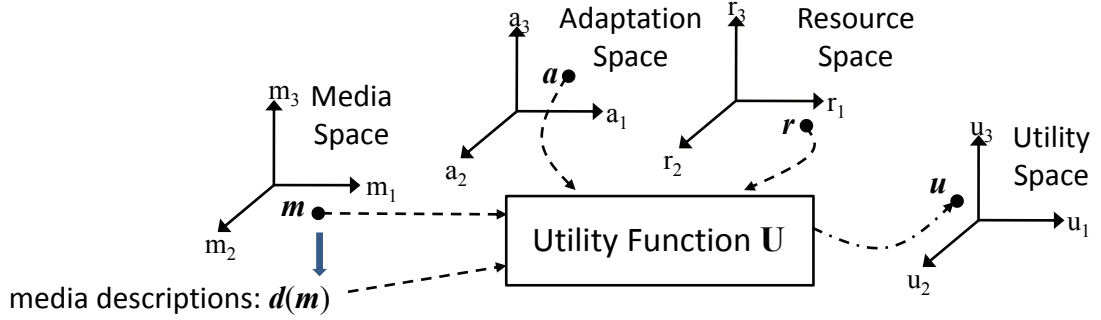
(a) Audiovisual features: intensity histogram, color histogram, and texture map [GW01].



(b) Perceptual arousals: color scheme [LG04] and intensity map [IKN98].

**Figure 1.6:** Sample media descriptions of the painting in Figure 1.5, given at two abstraction levels of (a) audiovisual features and (b) perceptual arousals.

age with medium intensity, dominant brown color, and less texture, as shown in Figure 1.6(a). Alternatively, it can be presented with the perceptual arousals as containing a color harmonious layout and several intensity-attractive regions (the faces), as shown in Figure 1.6(b). Obviously, the descriptions of either way are not in normal human forms to communicate the message about the painting. By contrast, the semantic concepts, such as outdoor, concert, performer, and audiences, would more faithfully reflect what users perceived from the painting. Therefore, in our work, we attempt to go beyond the scope of conventional feature or arousal based content analysis and propose a systematic adaptation framework at the semantic level. By bridging the semantic gap between systems and users, the developed adaptation engines are able to really satisfy the human's needs.



**Figure 1.7:** Interrelationship between the key elements (adaptation, resource, utility) of adaptive optimization in searching for the optimal adapting operation that maximizes the adaptation utility of a multimedia content. (Figure adapted from [Cha02])

### 1.2.3 Adaptive Optimization

An effective adaptation engine is able to make dynamic decisions in response to various usage environments. Meanwhile, the decisions are often required to be “optimal” so as to maximize the user’s multimedia experience. In the adaptation terminology, the optimality is represented with the *adaptation utility* [Cha02, Sun02, WKCK07]. Adaptation utility is defined as the quality of adapted contents with respect to some specific attributes that can be either at objective or subjective levels, e.g. the strength of audio signals [KMS05] versus the attractiveness of video highlights [WCC<sup>+</sup>07]. A metric used to measure the adaptation utility is then called a utility function, where a commonly used one is the peak signal-to-noise ratio (PSNR) [WOZ01]. In searching for optimal solutions to the selected utility functions, the adaptation problem can be conceptually transformed into the form of constrained optimization, namely adaptive optimization [Cha02, WKCK07, RL02].

In the state-of-the-art methodology [Cha02, WKCK07], three key elements are included in the construction of the adaptive optimization problems, i.e. adaptation, resource, and utility. Each individual aspect is defined using a parameter space. As illustrated in Figure 1.7, given a multimedia content  $\mathbf{m}$ , the adaptation space is the space of feasible adapting operations  $\mathbf{a}$ , the resource space is the affordable resources  $\mathbf{r}$  under the usage constraints, and the utility space is the value range of utility values  $\mathbf{u}$  obtained from the corresponding utility function  $\mathbf{U}$  with  $\mathbf{m}$ ,  $\mathbf{a}$ , and  $\mathbf{r}$  as arguments. The formulation is as follows.

$$\hat{\mathbf{a}} = \underset{\mathbf{a}}{\operatorname{argmax}} \mathbf{U}(\mathbf{m}, \mathbf{a}, \mathbf{r}), \quad (1.1)$$

where the optimal solution  $\hat{\mathbf{a}}$  is adopted as the final adapting operation to be performed on  $\mathbf{m}$ . However, it can be found that  $\mathbf{m}$ 's media descriptions  $\mathbf{d}(\mathbf{m})$  are not taken into account in the above equation. In practice,  $\mathbf{d}(\mathbf{m})$  could play a more important role than  $\mathbf{m}$  itself in the process of making adaption decisions. That is, as shown in the proposed adaptation framework (referring to the semantic content adaptation module in Figure 1.3), the knowledge about how the current adapting operation can be performed on a multimedia content is obtained from its media descriptions and their quality level has direct impacts on the estimated adaptation utility. For example, cropping images with manually-labeled regions-of-interest (ROIs) naturally receives higher subjective scores than those with automatically-detected ones [CWW07]. Similarly, the capability difference between detectors of the semantic concepts is another influential factor, such as the diverse tradeoffs between the precision and the recall performances.

Therefore, for practical applications, the preliminary formulation developed in Equation 1.1 should also include a content's media descriptions into the computation of its utility values. One possibility is to extend the  $\mathbf{U}$ 's definition to take media descriptions as additional arguments. This extension is able to pro-

vide more complete characterization of the proposed adaptation framework and allows the underlying components to be conceptually integrated in a more tightly coupled manner. In addition, some other observations can be made from the above formulation as well. First, it was originally introduced from a macro view to depict a general relationship between the key elements involved in adaptation [Cha02]. There exists much freedom to specify a workable instance for each of the elements. For example, candidates of the adapting operation for image resizing can be of any related technologies, such as the nearest-neighbor, bilinear, or bicubic interpolation [GW01]. By contrast, we put more emphasis on applying the formulation to identify the operational behaviors of the proposed adaptation framework. Possible combinations of the elements are then limited to the functionality that a realized adaptation engine actually supports. Second, the utility function  $U$  is not universal but changed according to the target applications. In real use, if it has certain time constraints (e.g. responding to users in real time) or there is a large search space for finding the optimal solution, the computations are often forced to terminate early and simply return a sub-optimal solution from the ones examined.

### 1.3 Problem Statement

In this dissertation, we address the problem of semantic adaptation for multimedia contents. The problem can be described as:

*Given multimedia contents and the background contexts, develop systematic techniques to identify the semantic concepts, whereby the associated adapting operations can be chosen to maximize the adaptation utility while satisfying the objective usage constraints.*

To facilitate the achievement of our objective, the task involves the research effort from several aspects:

- The construction of semantics ontology: The subject of semantics ontology is the study of the categories of semantic concepts that exist in the context of the examined domain [DMK<sup>+</sup>05]. The product of the study provides a fundamental basis for defining the user's semantic interests about the contents.
- The extraction of multimedia semantics: The subject of semantics extraction is the study of efficient algorithms for analyzing and extracting the semantic concepts embedded in multimedia contents. The modality integration of multimodal inputs should be investigated to offer reliable performance.
- The determination of adaptation strategy: The subject of adaptation strategy is the study of flexible mechanisms for making adaptation decisions in response to dynamic usage environments. The tradeoff among different constraints should be balanced to maximize the utility of adapted contents.

## 1.4 Summary of Contributions

This dissertation is devoted to the study of semantic adaptation for multimedia contents, where both the theoretical foundation and the practical realization are investigated in support of real-world adaptation scenarios. The main contributions of our work in solving the problems are threefold, as summarized as follows.

### 1.4.1 Framework Development for Semantic Adaptation

A generic framework is developed for the semantic multimedia content adaptation, in which the function of active content analysis is integrated for better management over the contents and the use of domain knowledge effectively enhances the semantic level of computational media descriptions. In terms of adaptation utility, an optimization formulation is derived to give theoretical support of the framework, which identifies the uniqueness of our work and makes the connection to general ideas in the research fields. Meanwhile, the design principles behind the framework are explicitly specified and can be served as development guidelines for practical adaptation engines.

### 1.4.2 Video Adaptation Based on Semantic Objects

A methodology built upon the proposed framework is developed for enabling semantic object based video adaptation. In this work, the knowledge of media aesthetics [Zet98, BT03] is employed to define the semantic concepts in professional videos, including the user-interest and the background objects. According to the content characteristics of the semantic objects, a unimodal approach using only the visual information is developed for the object modeling and detection. The analysis outputs are then applied to an adaptation scenario of delivering high-quality videos onto small devices. By recomposing the video content to visually emphasize the user's semantic interests in the video, the browsing experience can be effectively improved.

### 1.4.3 Video Adaptation Based on Semantic Events

A methodology built upon the proposed framework is developed for enabling semantic event based video adaptation. In this work, the knowledge of wedding customs [Spa01, War06] is employed to define the semantic concepts in home videos of the western wedding ceremonies, including thirteen kinds of semantic events like the ring exchange and the wedding kiss. According to the content characteristics of the semantic events, a multimodal approach using both the visual and aural information is developed for the event modeling and detection. The analysis outputs are then applied to an adaptation scenario of partitioning hours-long videos into semantically meaningful segments. Through the explicit recognition of semantic events, personalized applications can be built to satisfy the individual's preferences.

## 1.5 Organization of the Dissertation

The rest of this dissertation is organized as follows. In the next chapter, we review the literature on studies of the semantic adaptation from aspects relating to the main building blocks of the proposed generic framework, i.e. semantic concept ontology, semantic concept analysis, and semantic content adaptation. The correspondence between the proposed framework and the other works we have done in this dissertation is also described. In Chapter 3 and Chapter 4, two systematic approaches for modeling semantics in video contents are first investigated, respectively on object-based and event-based semantic concepts. Chapter 3 then presents an object-based mechanism for recomposing a video scene to improve the user's browsing experience of high-quality videos on the devices with small displays. Chapter 4 exploits the computable semantic events to benefit the user's

navigation in hours-long videos by providing a structured video index that allows directing access to the requested contents. Finally, Chapter 5 presents the conclusions of our work and possible directions of our future research.





## Chapter 2

# Basics and Literature Review

This chapter reviews relevant literature on studies of the semantic adaptation. It makes comparisons and gives discussions on the traditional work to further clarify the basics of our research philosophy. In particular, we conduct the review along the line of the main building blocks of the proposed adaptation framework. Section 2.1 presents the topic of semantic concept ontology. Section 2.2 and Section 2.3 are relating to the semantic concept analysis and the semantic content adaptation, respectively. In Section 2.4, the correspondence between the proposed framework and the other works we have done in this dissertation is described. Section 2.5 concludes this chapter.

### 2.1 Semantic Concept Ontology

*Ontology* is a form of knowledge representation about the world and usually represented as nodes and links between the nodes [Jim05, NST<sup>+</sup>06, GLF06]. The nodes represent facts within a domain (e.g. “red” and “color”) and the links represent relationships between those facts (e.g. red “is a” color). According to applications, the facts of a node can be stored either by name (e.g. the literal text of “red”) or by actual contents (e.g. a 32-bit RGB value of “red”), and allow

to be defined at different conceptual levels [CV07]. For example, by common sense, “cat” and “dog” are generally considered as facts of the same level and can be further assigned to more general notions, such as “animal”, at a higher level of the ontology. Therefore, a semantic concept ontology then refers to the ontology constructed for representing knowledge about specific semantic concepts as described in Section 1.2.2.

In the rest of this section, Section 2.1.1 first reviews several typical constructions of the ontology in the literature. We then discuss the relationship building between ontological nodes in Section 2.1.2.

### 2.1.1 Typical Examples

The use of ontologies to incorporate domain knowledge has long been studied. A number of typical examples are described in the following.

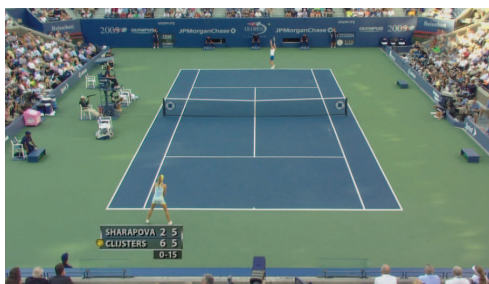
- **WordNet:** WordNet is arguably the most popular and widely used lexical database of English, in which the words are manually organized into sets of synonyms (e.g. “car” and “automobile”) and related by the semantic relationships such as antonymy (an opposite to), hypernymy/hyponymy (a specialization of/a generalization of), and so forth [Fel98]. It is optimized for general-purpose use and has been extensively applied to a variety of textual analysis (e.g. keyword expansions).
- **Cyc:** Cyc aims to formalize facts about everyday life into a logical framework, in which the facts are manually translated into assertions based on the first-order logic [Len95]. For example, the fact “red is a color” can be written as an assertion in the form of a predicate-arguments tuple, i.e. “(is-a red color)”. Cyc is also equipped with inference engines to deduce new facts

from the ones already stored. As compared to Cyc, it is also for general-purpose use but could make more practical reasoning over real-world texts. ConceptNet is a similar work inspired by Cyc [LS04].

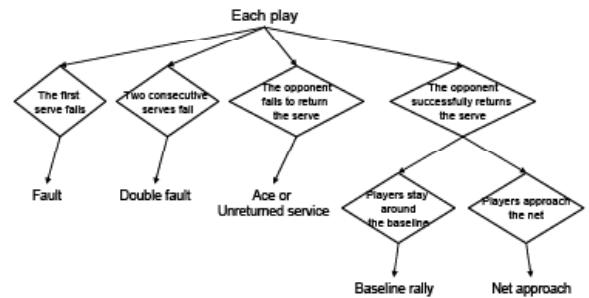
- **LSCOM (Large-Scale Concept Ontology for Multimedia):** LSCOM is a collaborative research effort in creating a large set of vocabularies for describing the semantics in multimedia contents, especially for broadcast news videos, such as face, people, flag, animal, and vehicle [NST<sup>+</sup>06]. The vocabularies are handcrafted by experts from various research communities (e.g. information retrieval and computational linguistics) and required to meet certain criteria, such as utility (high relevance to realistic video retrieval problems), feasibility (high likelihood of automated extraction), and observability (high frequency of occurrence in actual video data sets). LSCOM has been shown a useful tool to benefit the video research in its target domains.
- **TRECVID (TREC Video Retrieval Evaluation):** TRECVID is an international competition sponsored by the National Institute of Standards and Technology (NIST) of USA for encouraging the research in content-based video retrieval [tre]. For this purpose, TRECVID standardizes a large set of high-level concepts as the benchmark for participants in comparing their results. The TRECVID concepts are mainly extended from the LSCOM corpora [NST<sup>+</sup>06, HCY08], and they are also with an emphasis on broadcast news videos. Some examples are illustrated in Figure 2.1. Today, TRECVID has become one of the most important activities in video research communities, and the number of standardized concepts is still growing to extend the coverage on various content domains.



**Figure 2.1:** Illustrations of 101 TRECVID concepts. (Figure excerpted from [SWvG<sup>+</sup>06])



(a)



(b)

**Figure 2.2:** Example of a tennis match with (a) a video snapshot and (b) the types of tennis events. (Figure excerpted from [TWC<sup>+</sup>08])

Overall, some observations can be made from the above discussions. First, in multimedia research fields, the development of ontologies has gradually shifted its focus from general-purpose to domain-specific ones [WLC06]. That does not mean the generalization is a wrong direction. However, when the background context is unknown, it is naturally more difficult to identify which facts should be taken into account and how to appropriately define their relationships, as described in Section 1.2.2. Second, for practical use, the effectiveness of a large scale ontology is not necessarily better than that of a smaller one [HCY08]. For example, a high coverage of the overall semantic space within a target domain can be achieved by adding more facts, but it may concurrently cause the problem of semantic ambiguity between the whole ones. As reported in the previous work [HCY08], the performance of interactive video retrieval could inversely go down at some point as the increase of facts. Therefore, for the ontology design, the ontological fitness to actual applications is still one of the most important considerations.

### 2.1.2 Relationship Building

Relationship building refers to the task of linking ontological nodes according to certain relational rules defined in the target domain [WLC06, HCY08]. For example, under the lexical domain of English, the two nodes “rock” and “stone” can be connected by the relationship of synonym. As mentioned in Section 2.1.1, most of the relationships heavily rely on handcrafting by the experts with prior domain knowledge. For the experts, it is not a difficult task but could be labor-intensive and extremely time-consuming. As described in the previous work [WLC06], for example, it took about three hours for one person to complete an animal ontology with 200 nodes. However, automating the process seems almost impossible with today’s technology [HCY08].

In the literature, a compromising approach to reduce the user's burden is allowing the data to connect related nodes automatically and the actual relational meanings are then suggested by human knowledge [HCY08, ZSX07, Fre04]. Techniques of the social network are possible solutions [Fre04]. We consider, for example, the problem of constructing relationships between English words. Given  $N$  text documents, the degree of relevance between two words,  $w_1$  and  $w_2$ , can be measured by their co-occurrence using a simple similarity metric, say Jaccard coefficient  $JC(w_1, w_2)$  [Sal83], as defined below:

$$JC(w_1, w_2) = \frac{\Theta(w_1 \wedge w_2)}{\Theta(w_1) + \Theta(w_2) - \Theta(w_1 \wedge w_2)}, \quad (2.1)$$

where  $\wedge$  denotes the AND operator between words, and  $\Theta$  is a function that returns an estimate of the frequency of occurrence of either a single or pair of words with respect to the  $N$  documents. The relationship between any two words of high relevance are then judged by humans. In this way, it is helpful to reveal some implicit but insightful relations. For example, the word pair of “black” and “white” is often used as representatives of “evil” and “virtue”. However, some actual relations may be also filtered out at the same time simply due to their low relevance values.

## 2.2 Semantic Concept Analysis

Semantic concept analysis refers to the detection of targeted semantic concepts in multimedia contents, which is often achieved by means of automatic mechanisms [Chu06, CEJ<sup>+</sup>07, WLC06, HCY08]. In the multimedia research communities, we can observe two mainstreams on developing the detectors. One is to generalize the capability of existing ones to various content types [CEJ<sup>+</sup>07, WLC06],

and another is to focus on specific application domains by exploiting the corresponding domain knowledge [BKOK04, Chu06], as detailed below.

Along the first direction, TRECVID [tre, HCY08] is no doubt in recent years the most influential activity that gathers major research efforts around the world for working toward effective detectors of semantic concepts, cf. Figure 2.1. As mentioned in Section 2.1.1, although the selected concepts are slightly biased to broadcast news videos, the detector design is required for general use in video contents [HCY08]. For each TRECVID participant, a standardized evaluation method is used to evaluate the performance of their detectors. Specifically, the accuracy of detection is measured by mean average precision (MAP) in a video-shot basis [tre]:

$$MAP = \frac{1}{R} \sum_{r=1}^R P(r), \quad (2.2)$$

where  $r$  is the participant's rank of a detected shot for a specified semantic concept,  $R$  is the number of ranked shots in total, and  $P(r)$  is the precision in the ranking at a given cut-off rank  $r$ . Recent reports [tre, HCY08] show that there exist large variations between the attainable MAP performances of different semantic concepts, ranging from less than 0.1 (e.g. "corporate leader") to above 0.6 (e.g. "weather"), with an average of close to 0.2. Few specialized ones (e.g. "face") can achieve even higher scores, but at present the overall performance seems far from practical.

On the other hand, some research effort concentrates on limited application scopes as a tradeoff to sacrifice the partial generality of detectors for acceptable accuracy [BKOK04, Chu06]. It also opens the way for taking advantage of the domain knowledge to better satisfy the application's needs by making specific designs on the detectors. Since the study of sporting videos mostly belongs to this category, we then explain by taking the event analysis of broadcast tennis

videos as an example [TWC<sup>+</sup>08]. As illustrated in Figure 2.2, a tennis match is composed of certain events, including the fault, double fault, ace, baseline rally, and net approach. They are observed to be distinguishable by several audiovisual hints from the tennis knowledge, such as the player’s relative position in the court, the moving distance of players, the sound effects of applause or cheer, and the number of racquet hits [TWC<sup>+</sup>08]. Specifically, the recognition accuracy of the tennis events can be as high as 0.8 to 0.9 for those detectors that utilize the above information in their development.

Based on the above discussions, we believe that moderate user intervention can help to take advantage of the both methods. That is, the detectors of semantic concepts can be generalized to some extent between a single domain and the fully general ones, such as the scope of ball sports with respect to that of soccer or baseball. The user then only has to specify the target scope by some ways (e.g. a list) to benefit the selection of appropriate detectors.

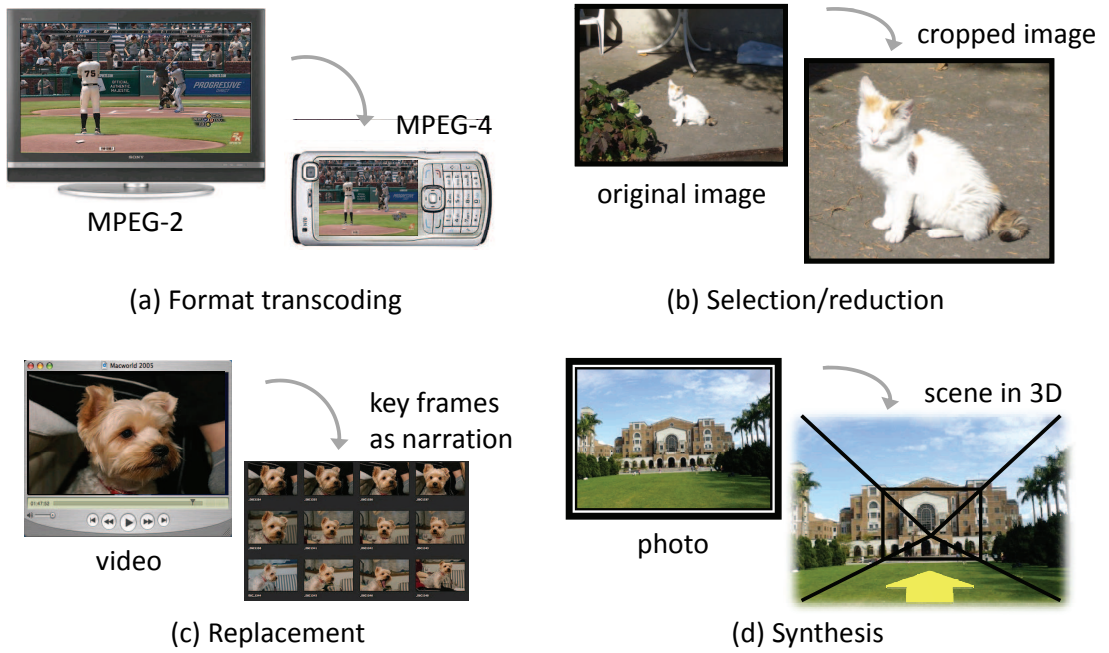
## 2.3 Semantic Content Adaptation

In this section, we first present a brief review on the conventional categorization of adapting operations in Section 2.3.1. Section 2.3.2 then describes the determination of effective adaptation strategies for the adapting operations.

### 2.3.1 Adaptation Taxonomy

As defined in Section 1.1, the scope of adaptation technologies covers a broad class of adapting operations that either adapting the multimedia content to fit a usage environment or adapting the usage environment to accommodate the content [BPdWK06, PB03, CV05]. Conventionally, by Chang’s definitions [CV05],





**Figure 2.3:** Adaptation taxonomy with examples of the corresponding adapting operations.

the adapting operations can be classified into four main categories: format transcoding, selection/reduction, replacement, and synthesis, as described below. Note that the taxonomy is originally proposed for video operations, but it is generic enough to be applicable for most of the multimedia operations as well.

- **Format Transcoding:** It refers to the conversion of multimedia contents from one form of coded representation to another. For example, MPEG-2 videos are transcoded to MPEG-4 formats for Internet streaming [XLS05].
- **Selection/Reduction:** It refers to the elimination/degradation of some components of the multimedia contents. For example, images are cropped to preserve only the ROI regions [HWG04].

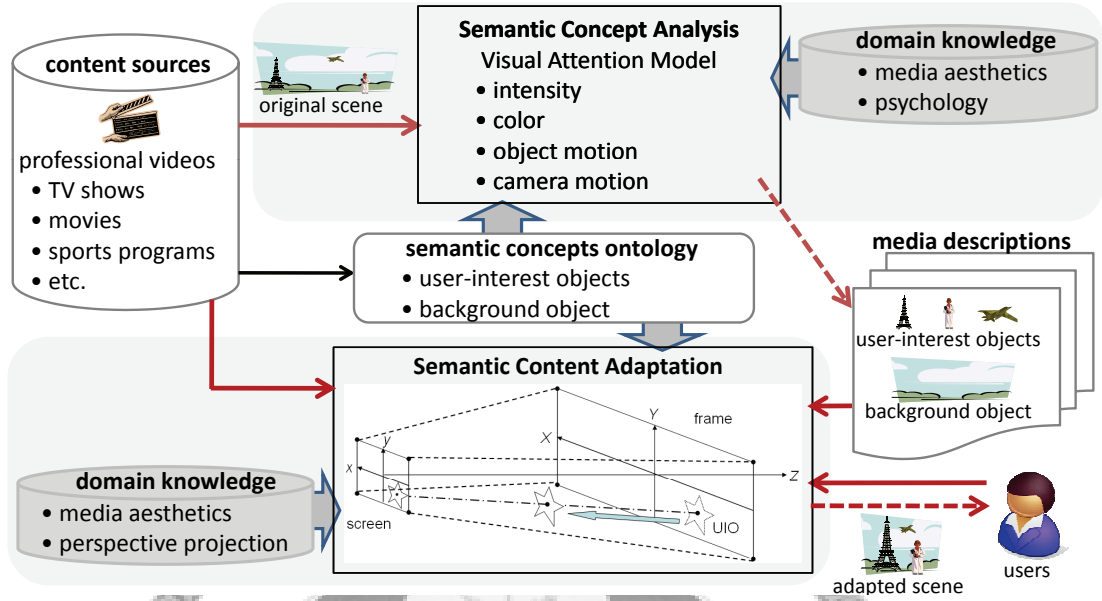
- **Replacement:** It refers to the substitution of selected elements in a multimedia content with more efficient counterparts. For example, videos are represented with a set of key frames as the narrative summary [WCC<sup>+</sup>07].
- **Synthesis:** It refers to the generation of new presentations from original multimedia contents. For example, photos are to be presented in 3D for enhancing the user's browsing enjoyment [HAA97].

Examples associated with each of the categories are illustrated in Figure 2.3. Obviously, the taxonomy is developed based on the basic types of adapting functionality [CV05, TV07, AWSZ05]. In practical applications, they can also be combined to analyze more complex adapting operations.

### 2.3.2 Adaptation Strategy

Adaptation strategy is the study of feasible mechanisms for making adaptation decisions of multimedia contents, as described in Section 1.2. The strategy determination can largely benefit from the awareness of multimedia semantics and the knowledge in relation to usage environments [ABBH08, KMS05, JP08].

For example, the previous research in mobile TV shows that the user's minimal perceptible size of videos is not a constant but changes according to the type of video contents [ABBH08]. Later, the influence of other factors, such as the picture ratio and the audio bitrate, is also reported to be content-dependent on the user's video experience [JP08]. Clearly, the interrelationship between multimedia contents and the usage environments plays an important role in the development of effective adapting operations. Therefore, in a sense, the proposed notion of semantic adaptation not only refers to the semantic level of concept analysis but also



**Figure 2.4:** The work of semantic object based video adaptation is represented by the proposed generic framework given in Figure 1.3.

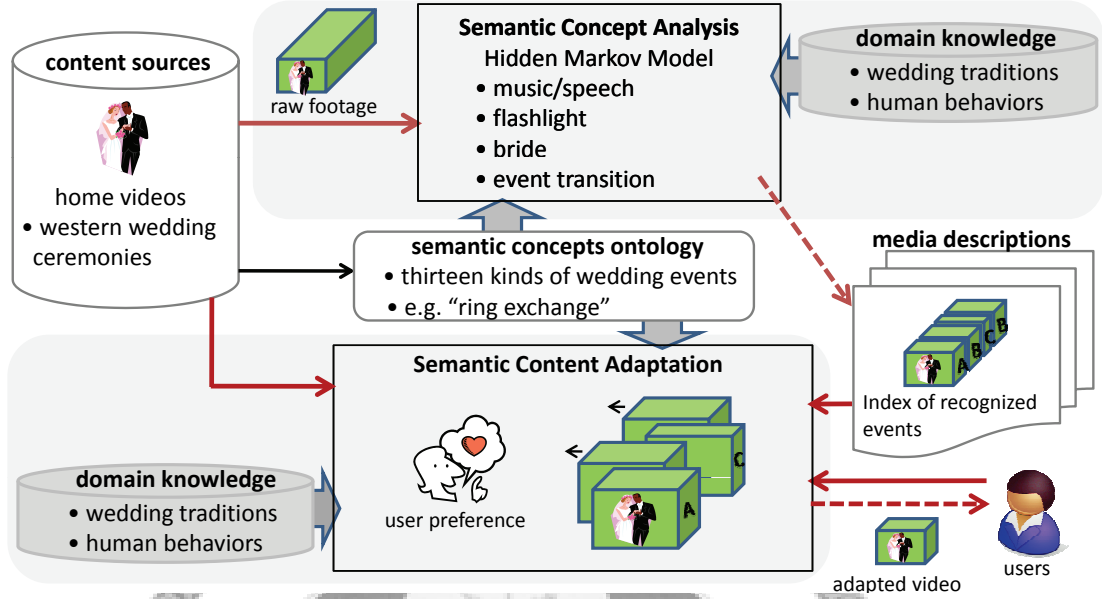
the efficient utilization of content semantics in the determination of adaptation strategies.

## 2.4 Framework Correspondence

In this section, we illustrate the correspondence between the proposed framework and the other works we have done in this dissertation, while the work details are referred to the rest of this dissertation.

### 2.4.1 Semantic Object Based Video Adaptation

The work of semantic object based video adaptation presented using the proposed framework is illustrated in Figure 2.4. This work focuses on the adaptation



**Figure 2.5:** The work of semantic event based video adaptation is represented by the proposed generic framework given in Figure 1.3.

scenario of delivering high-resolution videos onto small-display devices, in dealing with the content sources of professional videos, such as TV shows, movies, sports programs, and so forth. The targeted semantic concepts include the user-interest and the background objects, along with the domain knowledge of media aesthetics, psychology, and perspective projection.

## 2.4.2 Semantic Event Based Video Adaptation

The work of semantic event based video adaptation presented using the proposed framework is illustrated in Figure 2.5. This work focuses on the adaptation scenario of partitioning hours-long home videos into semantically meaningful segments, in dealing with the content sources of home videos, i.e. western wedding ceremonies. The targeted semantic concepts include thirteen kinds of wedding

events, such as the bride entering and the ring exchange, along with the domain knowledge of wedding traditions and human behaviors.

## **2.5 Summary**

This chapter reviews relevant literature on the semantic adaptation, from several perspectives of the semantic concept ontology, the semantic concept analysis, and the semantic content adaptation. A new frontier of the research is to integrate the analysis of content semantics and the development of adapting operations for efficient adaptation, which is recognized as a promising but challenging direction. That motivates our work of this dissertation.





## Chapter 3

# Semantic Object Based Video Adaptation

The browsing of quality videos on small hand-held devices is a common scenario in pervasive media environments. In this chapter, we propose a novel framework for video adaptation based on content recomposition. Our objective is to provide effective small size videos which emphasize the important aspects of a scene while faithfully retaining the background context. That is achieved by explicitly separating the manipulation of different video objects. A generic video attention model is developed to extract user-interest objects (UIOs), in which a high-level combination strategy is proposed for fusing the adopted three types of visual attention features: intensity, color, and motion. Based on the knowledge of media aesthetics, a set of aesthetic criteria is presented. Accordingly, these objects are well reintegrated with the direct-resized background to optimally match the specific screen sizes. Experimental results demonstrate the efficiency and effectiveness of our approach.

## 3.1 Introduction

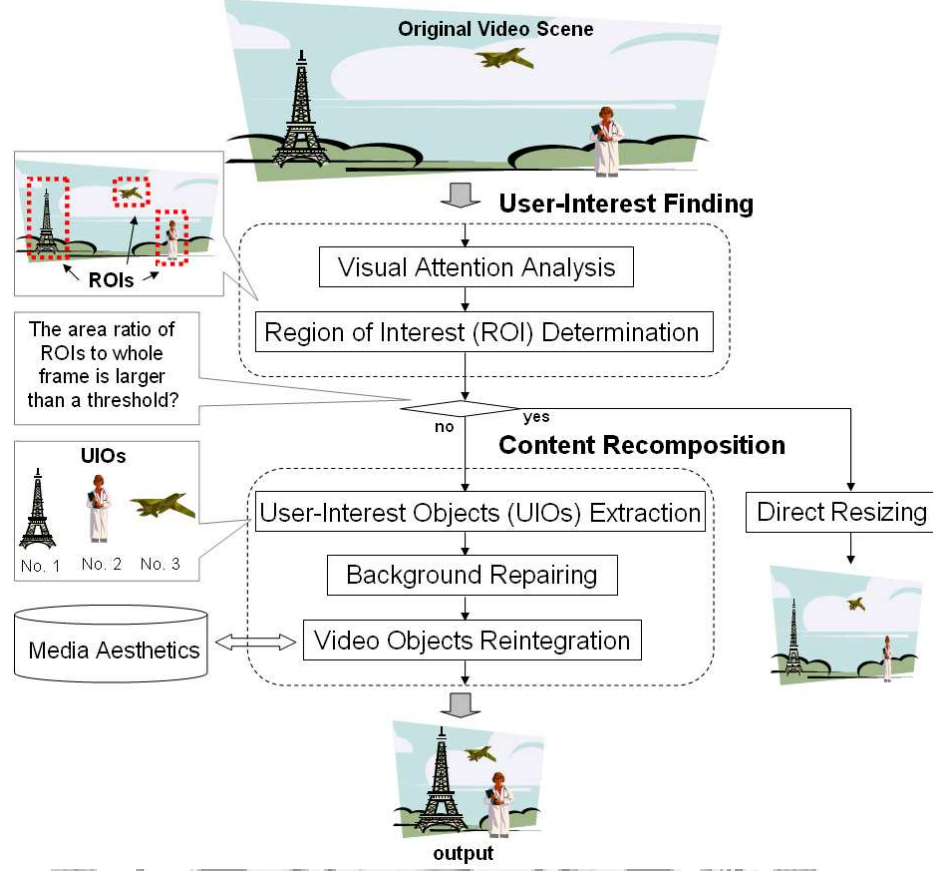
On the Internet, multimedia content has been widely used for sharing information among users. Their transparent access from almost everywhere at anytime through all kinds of devices is desired and often required. To enable such universal multimedia access (UMA), one key technology is *video adaptation* [PB03, MSL99, BGP03, CV05]. In general, it is defined as the mechanism of transforming a video stream with one or more operations to meet specific application needs, such as device capabilities, network characteristics, and user preferences. At the user's end, hand-held devices including cellular phones, Smartphones, PDAs, and Pocket PCs are now in widespread use for their mobility and portability. In order to compete with desktop computers for practical computing tasks, they are not only developed for more powerful functionality but also equipped with more storage capacity. However, one exception is the display. For the portable requirement of hand-held devices, the screen size is kept permanently unchanged and even as small as possible. With the rapid growth of quality video sources (e.g., mobile TV, VCD/DVD on demand), the physical limitation would seriously disrupt user's viewing experience [MSL99, LCC<sup>+</sup>01, CXF<sup>+</sup>03]. Thus, it is crucial to develop an efficient tool for facilitating video presentation on devices with limited display.

The conventional schemes that have been proposed for adapting videos on a small display can be divided into two categories: *spatial transcoding* and *frame cropping* [AWSZ05, XLS05]. The former subsamples each frame to preserve intact video contexts and the latter discards partial surroundings to highlight specific user interests. Due to the bias of their design purposes, an adaptation engine has to make visual trade-offs between the subject readability and content completeness [KMS05, STG<sup>+</sup>04, LG05]. However, sacrificing either aspect is usually intolerable



because they are both important in our viewing experience. For example, when watching sports programs, player recognition and full-court variation are both important visual concerns [KMS05, LG05]. The difficulties with the conventional schemes arise because they both passively attempt to adapt the plain frame but not the actual content it contains. Consequently, the adapting process is forced to specify a desired area of the source frame (maximally itself) and uniformly stuff it onto the target screen. Until we move away from that paradigm, the obtained performance will fall short of our expectations [STG<sup>+</sup>04, LG05].

In this work, we propose a novel framework for video adaptation based on content recomposition. Our objective is to provide *effective* small size videos that emphasize important aspects of the scene while retaining the background context for adaptive delivery. We focus on non-uniform processing of different video regions by giving more display resource (i.e., space) to the important ones and less to the other parts. Specifically, we use visual attention analysis to extract user-interest objects (UIOs) of a scene. With regard to the background, these objects are downsized at a light level and with constant aspect ratio (AR). Then, according to the principles of media aesthetics [BT03, Zet98, DV01], they are well reintegrated with the direct-resized background to optimally match various screen sizes of client devices (cf. Figure 3.1). Note that in this chapter the term *video objects* will be used interchangeably to indicate the collection of UIOs and the background. The recomposing-based framework provides a number of advantages over the conventional schemes. First, it improves the visibility of user’s interests as well as retains faithful context information, e.g., the viewer can see not only who but also where a person is in the video. Second, it allows multiple key objects to be emphasized at the same time and we can easily control the visual importance by adjusting their relative sizes. Third, it is robust to the shape distortion of



**Figure 3.1:** Flowchart of the proposed framework for conducting video adaptation.

objects caused by changes in video aspect ratio, which gives consistent content experience to different viewers.

The main contributions of our work are twofold. First, a generic visual attention model is developed for video user-interest finding. The model is universal for its adequate utilization of inherent video characteristics, such as object and camera motions. Specifically, a high-level feature combination strategy based on the camera motion information is proposed. In addition, the motion feature model is integrated with confidence measures to improve its robustness and reliability.

Second, based on the knowledge of media aesthetics, a set of aesthetic criteria is presented for guiding relevant decisions-making during video objects reintegration, such as the background positions to place UIOs. Without requiring user intervention, video content is automatically recomposed with satisfactory resultant visual rationality. We have conducted many experiments on various kinds of video data to demonstrate the efficiency and effectiveness of our approach.

The rest of this chapter is organized as follows. After a discussion of related work, Section 3.3 presents a video attention model and associated algorithms for user-interest analysis and determination. The media aesthetics based content recomposition are described in Section 3.4. Section 3.5 shows experimental results, and Section 3.6 presents our concluding remarks and the directions of our future work.

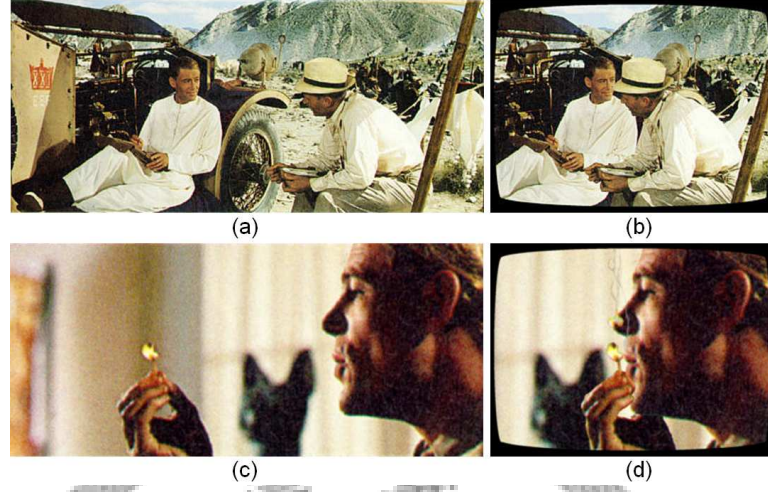
## 3.2 Related Work

In this section, we review previous studies on visual content adaptation. According to the design purposes, they are classified into three major categories, including *transcoding*, *cropping*, and *hybrid*. Meanwhile, their advantages and disadvantages to small displays will be briefly described as well.

In earlier works [CV05, AWSZ05, XLS05], the techniques of video transcoding have been extensively explored. The basic transcoding process is to convert a coded video signal from its original format into another one. An output format is determined entirely based on network and device constraints. Well-known transcoding methods include spatial resolution adjustment, temporal resolution adjustment, bit-rate adjustment, and coding syntax conversion [AWSZ05, XLS05]. Scalable video coding is considered a special kind of transcoding techniques [WOZ01, Tun02]. The scalability is accomplished by providing multiple

versions of a video stream so that the same contents of lower qualities are obtainable in different clients, e.g., Tung [Tun02] developed a unified MPEG-4 video codec that supports universal scalability. For clients with small displays, spatial transcoding is always required but it causes excessive spatial resolution reduction or visual quality degradation. Once a visual content is scaled down more than its *minimal perceptible size*, the quality of service (QoS) or quality of experience (QoE) is usually far from acceptable [CXF<sup>+</sup>03, KMS05]. For example, some important details, such as the gesture of a drama actor or the ball location of a sport game, are not easy or even impossible to be recognized. Another difficulty with the spatial transcoding occurs when the aspect ratio of a target screen is inconsistent with the source video. If we linearly reduce both dimensions of the video to fit into the screen, it leaves black borders (sometimes known as the letterbox) and wastes valuable display resource. On the other hand, if the video is non-linearly resized to occupy the whole screen, the resulting shape distortion of objects will annoy the viewer [Zet98].

Much attention is then put on cropping-based approaches [NYHK05, CSE05]. First, Mohan *et al* [MSL99] proposed a general framework for adapting multimedia web documents, in which each media item (e.g., a video clip) is described with a multimodal and multiresolution representation hierarchy called the InfoPyramid. An importance value is subjectively assigned to each of the item combinations as the transcoding hint for content servers to dynamically select the best output. Similar ideas are also applied in Lee's work [LCC<sup>+</sup>01]. Instead of treating one video frame as a whole, selective presentation (or frame cropping) is allowed to improve the visibility of user's regions of interest (ROIs). Following their work, Chen *et al* [CXF<sup>+</sup>03] developed an image adaptation system based on visual attention model. Using a simulated cognitive mechanism of human visual system



**Figure 3.2:** Examples of semantic distortion in adapted videos: (a) and (c) are two original frames from the classical film “*Lawrence of Arabia*”, and (b) and (d) are the corresponding adapted results using [fil], respectively. With partial coverage, the two men of (a) no longer look at each other’s eyes when they are chatting in (b), and the man in (d) seems more like to burn himself with the burning match rather than just hold it in (c). (Courtesy of FlikFX Ltd.)

(HVS) [IKN98, PS00, CCW05], the most important region is automatically determined. Better perceptual results have also been reported in other ROI-based video applications [LCS03, HWC05, HWG04]. However, from the viewpoint of content authors, it not only destroys the carefully worked-out compositions but also distorts the overall conveyed messages. For example, if a visual scene is composed of multiple key objects, some of them are necessarily thrown away and a single ROI would fail to show the overview of their interrelationship. Moreover, significant information loss leads to viewer’s misinterpretation about the original meaning that the authors want to communicate.

To preserve a complete video context or to clarify the specific user interest is not an either-or problem. Some hybrid approaches that lie between the two opposite extremes have been proposed. Liu *et al* [LXMZ03] presented a novel solution

for browsing large pictures on mobile devices. All of the important regions are serially displayed and an optimal browsing path is calculated according to predicted shifting of visual attention. *Pan and Scan* [Zet98] addressed an analogous technique for high-resolution video sources, but its discontinuous nature severely annoys the audience [Zet98, fil]. Besides, the requirement of additional temporal resolutions conflicts with the primary video structures. The FilkFX corporation developed an awarded commercial system for transferring wide-screen films to the 4:3 aspect ratio of TV screens [fil]. The intention is to generate a visually approximate replacement without object distortions. Therefore, each of the film frames is condensed by eliminating the background portions of little significance and the main actors are artificially brought together to concentrate viewer's attention. However, without considering the original spatial interrelationship of video objects, semantic distortions are often generated, cf. Figure 3.2. Recent work introduces non-uniform manipulations of the background and foreground information. Setlut *et al* [STG<sup>+</sup>04] decomposed an image into separate objects and unequally shrank them according to their relative visual importance. A side effect is that the relative size of different objects may be altered. Liu *et al* [LG05] exploited a non-linear warping transformation to emphasize the attractive foreground image regions but severe visual distortions are inevitable. Overall, the maintenance issue of user-perceived visual rationality in adapted results is not well addressed in the adaptation literature. Furthermore, although experiments show that non-uniform processing is more flexible to achieve superior performance, most discussions are confined to still images. These observations motivate our approach for motion pictures.

### 3.3 User-Interest Finding

UIOs are the semantic objects that catch part of the viewer's attention in videos, such as a walking person, a flower, an automobile, etc. Accordingly, an ROI is defined as the rectangular frame portion that contains some UIOs. Since the actual UIO shapes can be arbitrary, the ROIs serve as the tight bounding boxes of them. Their correct identification is the first key step for successful content recomposition. In intelligent image applications, a powerful mechanism for identifying the ROIs is visual attention modeling [IKN98, PS00]. Without truly understanding an image's content, several attentive features are extracted and combined into a single saliency map for representing local conspicuity. The computational attention methodology simplifies the problem of complex semantic analysis into a series of low-cost heuristic decisions. Several researches extend its capability for motion pictures by utilizing high-level video characteristics, such as speech, video genre, and lexical information [MLZL02, HCPW03]. Unfortunately, most designs are too domain-specific to be applied to general purpose applications [CCW05]. In the rest of this section, we will explain how to model generic visual attention in video clips. Rather than blindly adding semantic features (e.g., human face and text), a novel strategy for feature combination is presented to take the author's intentions into account. In addition, methods for dynamically determining the number of ROIs and their attributes (i.e., position and size) are also introduced.

#### 3.3.1 Visual Attention Modeling

Visual attention refers to the ability of a viewer concentrating his attention on some visual objects or regions. Previous research showed that this physiological

process could be modeled by a saliency-based attention model [IKN98, HCPW03], i.e., a saliency map computes an attractive value for each pixel or image block. Based on our previous studies [CCW05], three types of video-oriented visual features (intensity, color, and motion) are adopted to model the visual attraction by using the same idea.

### Contrast Based Intensity and Color Feature Model

One of the most important ingredients of a visual attention model is the contrast [EZW97]. In psychology, perceptual experiments have shown that the intensity and some color pairs possess high spatial and chromatic opposition. Accordingly, we include three contrast based feature models: intensity, red-green color contrast and blue-yellow color contrast, into our visual attention analysis module. The contrast maps are respectively defined as follows.

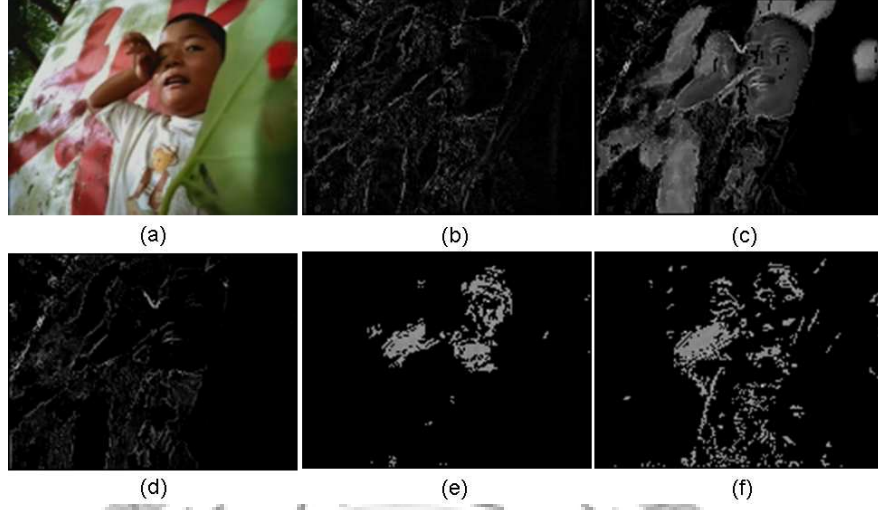
$$\mathcal{M}_I(p) = \max_{p' \in w_p} |\mathcal{I}(p) - \mathcal{I}(p')|, \quad (3.1)$$

$$\mathcal{M}_{RG}(p) = \max_{p' \in w_p} |(\mathcal{R}(p) - \mathcal{G}(p)) + (\mathcal{G}(p') - \mathcal{R}(p'))|, \quad (3.2)$$

$$\mathcal{M}_{BY}(p) = \max_{p' \in w_p} |(\mathcal{B}(p) - \mathcal{Y}(p)) + (\mathcal{Y}(p') - \mathcal{B}(p'))|, \quad (3.3)$$

where  $p = [x, y]^T$  is a position vector,  $w_p$  is a  $3 \times 3$  window centered at  $p$ , and  $\mathcal{I}$ ,  $\mathcal{R}$ ,  $\mathcal{G}$ ,  $\mathcal{B}$ ,  $\mathcal{Y}$  denote the intensity, red, green, blue, and yellow component value functions, respectively. That is, for each frame, the intensity and color feature values of a pixel  $p$  are computed from its local region  $w_p$ . For example, the intensity value of  $p$  is set to the maximum of the absolute intensity differences with its neighboring pixels  $p'$ .





**Figure 3.3:** Example of feature maps: (a) original video frame, (b) intensity, (c) red-green color, (d) blue-yellow color, (e) x-motion, and (f) y-motion feature maps.

### Motion Feature Model

Object motion plays an essential role to direct an audience's attention across the scene space of a video [BT03]. Two feature models: x-motion and y-motion, are respectively used to represent the horizontal and the vertical motion information in a scene. To find the motion activity of a specific direction, the two-dimensional (2-D) [NPZ03, J91] structure tensor ( $ST$ ) is evaluated for each frame pixel. Compared with other motion descriptors, the 2-D  $ST$  is adopted for its confidence measure can also be estimated. The 2-D  $ST$ ,  $J_x$ , for computing x-motion features is expressed as

$$J_x = \begin{bmatrix} J_{xx} & J_{xt} \\ J_{xt} & J_{tt} \end{bmatrix} = \begin{bmatrix} \sum_w H_x^2 & \sum_w H_x H_t \\ \sum_w H_x H_t & \sum_w H_t^2 \end{bmatrix}, \quad (3.4)$$

where  $w$  is the  $3 \times 3$  support window.  $H_x$  and  $H_t$  are respectively the partial derivatives of a horizontal slice along the spatial and the temporal dimensions as

defined in [NPZ03]. Consequently, the local motion angle  $\theta_x$  and its corresponding confidence measure ( $cm_x$ ) are computed as

$$\theta_x = \frac{1}{2} \tan^{-1} \frac{2J_{xt}}{J_{xx} - J_{tt}}, \quad (3.5)$$

and

$$cm_x = \frac{(J_{xx} - J_{tt})^2 + 4J_{xt}^2}{J_{xx} + J_{tt})^2}, \quad 0 \leq cm_x \leq 1. \quad (3.6)$$

The corresponding y-motion features,  $\theta_y$  and  $cm_y$ , can be obtained in the same way. Finally, the x-motion and the y-motion maps are individually calculated as:

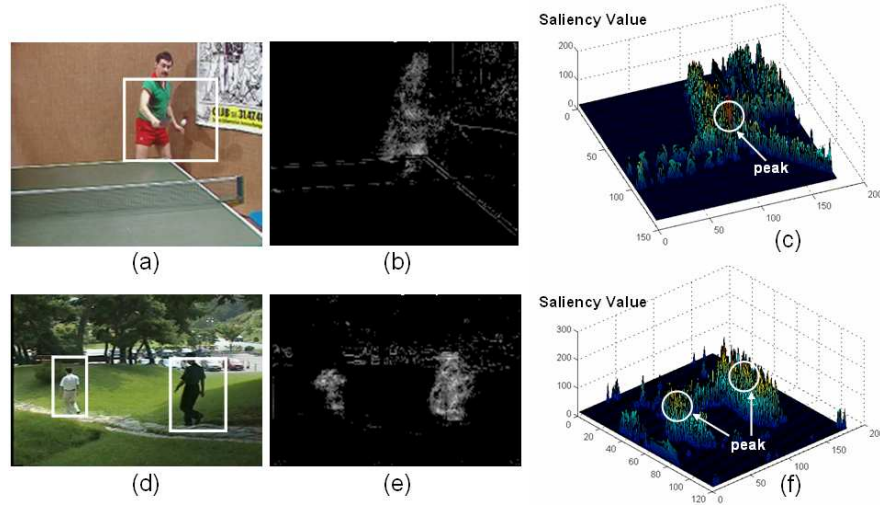
$$\mathcal{M}_X(p) = \theta_x \times cm_x, \quad (3.7)$$

$$\mathcal{M}_Y(p) = \theta_y \times cm_y, \quad (3.8)$$

where  $p = [x, y]^T$  is, again, a position vector. That is, for each frame, the motion feature values of a pixel  $p$  are its local motion angle  $\theta_x$  ( $\theta_y$ ) multiplied by the corresponding confidence measure  $cm_x$  ( $cm_y$ ). The confidence measure is used to suppress uncertain motions and to improve the reliability of obtained motion feature maps.

### Camera Motion Based Saliency Map Generation

For each video frame, the distributions of individual visual features are calculated and constructed as five feature maps, as shown in Figure 3.3. Then, according to the theory of visual attention model [IKN98, IK99], one saliency map is generated by their linear combinations, as described in the rest of this subsection. In the combining process, the selection of relative feature weights is important, which influences the accuracy of obtained saliency maps [IK99]. In the literature, the typical solution is assigning a single set of fixed feature weights



**Figure 3.4:** Examples of a video frame with (a) one and (d) two ROIs (indicated by the white squares); (b) and (e) are the corresponding saliency maps, and (c) and (f) are the 3-D profiles of the saliency maps of (a) and (d), respectively.

for a whole video, e.g., the equal-weights [IKN98, MLZL02] and the video-genre-based schemes [HCPW03]. However, they are inflexible to adapt themselves to content variations of a video. Later, some dynamic fusion schemes were proposed, e.g., [IK99, HZ04]. Although better results are obtained, the blindness to content semantics limits their extension for high-level applications.

From the viewpoint of media aesthetics [BT03, Zet98], different camera motions have different impacts on the audience's reception. They influence the relative importance of each visual feature and reveal what and where the author wants viewers to see. For example, a pan camera usually implies a tracking intention of some fast moving objects, e.g., a car in racing [BT03]. At this time, the horizontal motion feature would be more attractive to viewers and should have a larger weight than the other visual features. The study of task-oriented gaze control confirms the phenomenon. Under the same camera motion type, users' perceptual responses to the visual stimuli are generally consistent regardless of

**Table 3.1:** Weights for the feature maps under different camera motion types.

	$\mathcal{M}_I$	$\mathcal{M}_{RG}$	$\mathcal{M}_{BY}$	$\mathcal{M}_X$	$\mathcal{M}_Y$
zoom	0.2	0.2	0.2	0.2	0.2
pan	0.05	0.05	0.05	0.75	0.1
tilt	0.05	0.05	0.05	0.1	0.75
static	0.15	0.075	0.075	0.35	0.35
motion	0.05	0.05	0.05	0.425	0.425

the video genres [CCW05]. In videography, the techniques have been widely used in video productions especially in expert-produced videos [Zet98]. Accordingly, the fact that directors purposely move their camera to control the audience’s fixations appropriately serves as a high-level hint for integrating visual features [BT03, Zet98]. Therefore, we propose a camera motion based feature combination strategy for saliency map generation. Conceptually, this strategy can be viewed as one kind of the dynamic fusion schemes as prescribed. However, the fusion weights are determined by available high-level information (i.e., camera motion type) rather than the to-be-fused data itself.

In our work, five camera motion types are labeled: zoom, horizontal-pan, vertical-tilt, static-with-no-motion, and static-with-object-motion. Using an algorithm based on structure tensor histograms [NPZ03], one camera motion type is registered for every video frame. The saliency map  $S$  is generated according to the following equation:

$$S = \alpha_{c,1} \times FM_1 + \cdots + \alpha_{c,n} \times FM_n, \quad (3.9)$$

where  $FM_i$  is the  $i$ -th feature map of that frame, and  $\alpha_{c,i}$  is the weight of corresponding  $FM_i$  under a given camera motion type  $c$ . Table 3.1 lists the feature weights for the adopted camera motion types. Instead of manual assignment, these parameters are defined via a supervised training process for feature weights selection (please refer to Section 3.4.4 of [CCW05] for the details). Note that the

physical camera arguments (e.g., focal length) are not involved in the training process. The camera motion types are only used to classify the training data, and the same training process is separately conducted. The training data includes fifty video segments for each camera motion type, all of them are carefully chosen from various expert-produced films and TV programs. Each segment is 0.5 seconds long (roughly 15 frames) and contains a single camera motion type that is determined using the algorithm of [NPZ03].

The proposed feature combination strategy offers a number of advantages over the conventional ones [MLZL02, HCPW03, IK99]. First, it provides the capability of instant reaction to content variations. That is, the feature weights are dynamically selected according to the high-level hint of registered camera motion types. Next, it is generic because the camera operations are always available in videos and their classifications have been well-defined [BT03, Zet98, NPZ03]. Finally, it adds only moderate computational overhead since the required information (i.e., ST values) had been collected in the motion feature model.

### 3.3.2 Video ROIs Determination

In our work, an ROI is defined with two attributes: centroid position and region size (as described in the following). In this subsection, we describe how to compute the attributes for each ROI from a saliency map. Since there may be multiple key objects in a video frame, a method for dynamically determining the number of ROIs is also presented.

#### ROI Attributes Calculation

Saliency weighted regular moments [PRO98] are effective to calculate the center coordinate of a set of weighted data points. They are adopted in our work to

determine the centroid of each ROI. Let

$$m_{pq} = \sum_{x=1}^M \sum_{y=1}^N x^p y^q s(x, y), \quad p, q = 0, 1, 2, \dots, \quad (3.10)$$

where  $M, N$  are the dimensions of a saliency map and  $s(x, y)$  is the saliency value function corresponding to the pixel  $(x, y)$ . In the saliency map of the  $k$ -th frame, the centroid position of an ROI is given by  $(x_k, y_k) = (m_{10}/m_{00}, m_{01}/m_{00})$ . Further, based on our observations, the region size of each ROI is proportional to the spatial distribution (area) of saliency values on a saliency map. A saliency weighted invariant [PRO98] is defined to measure the variation of a computed centroid as follows:

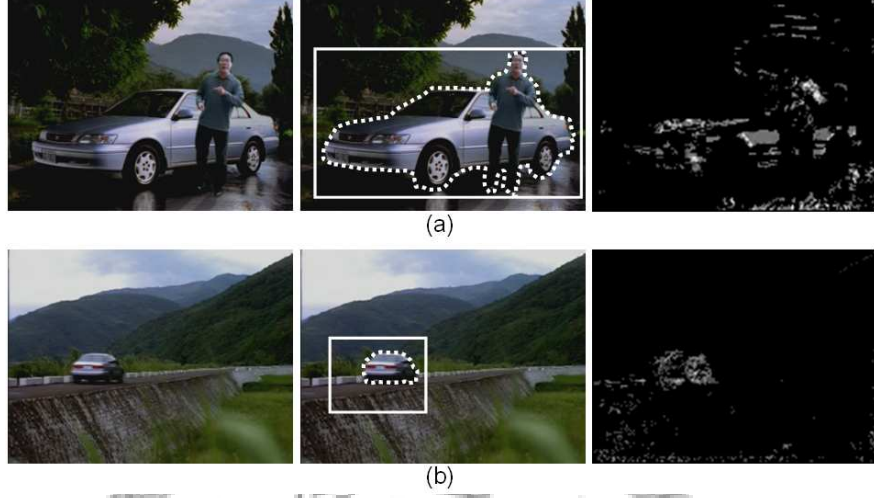
$$\eta_{pq} = \frac{\sum_{x=1}^M \sum_{y=1}^N (x - \bar{x})^p (y - \bar{y})^q s(x, y)}{m_{00}}. \quad (3.11)$$

Consequently, the region size is set as  $(e\sqrt{\eta_{20}}) \times (e\sqrt{\eta_{02}})$ , where  $e = 2$  is the expansion factor.

Meanwhile, we propose using a tracking technique of the discrete Kalman filter [May79, WB04] to estimate and correct the computed ROI attributes. Generally, an ROI centroid with its corresponding positions in the previous frames constitute a smoothly continuous trajectory on the screen. Therefore, a predicted centroid  $(x_k^-, y_k^-)$  of the ROI can be obtained with the prior information as follows [WB04]:

$$\begin{bmatrix} x_k^- \\ y_k^- \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_{k-1} \\ y_{k-1} \\ \Delta_{x_{k-1}, k-2} \\ \Delta_{y_{k-1}, k-2} \end{bmatrix} + \begin{bmatrix} w_{x_{k-1}} \\ w_{y_{k-1}} \end{bmatrix}, \quad (3.12)$$

where  $(\Delta_{x_{k-1}, k-2}, \Delta_{y_{k-1}, k-2})$  is the centroid difference of the ROI between the  $(k-1)$ -th and the  $(k-2)$ -th frames, and  $w_{x_{k-1}}$  and  $w_{y_{k-1}}$  are two independent white noises with normal probability distribution  $\mathbf{N}(0, 0.5)$ . If the Euclidean distance of the

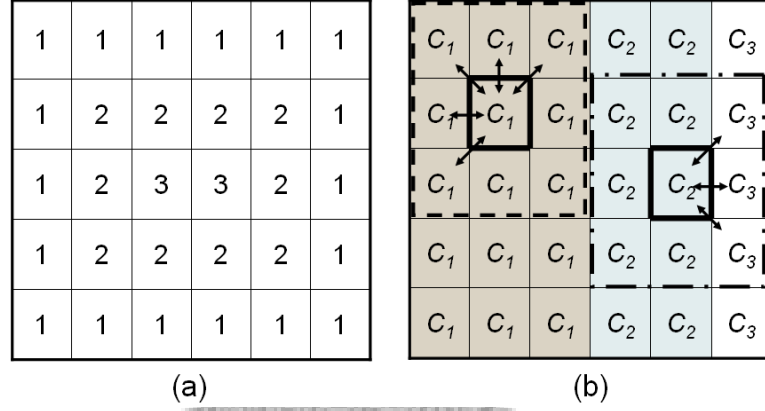


**Figure 3.5:** Comparisons of the ROI and the UIO representations for user-interests. They are respectively indicated by the solid and dotted lines. In (a) and (b), the number of contained semantic objects (man together with a car versus one single car) is different.

computed  $(x_k, y_k)$  and the predicted  $(x_k^-, y_k^-)$  positions is larger than a dynamic threshold  $\tau_k$ , the computed centroid  $(x_k, y_k)$  will be treated as unreliable. For example, a flash event often alters the spatial distribution of a saliency map and sharply shifts away the computed ROI centroid from where it should be. In this case, we would “propagate” the ROI from the previous frame instead. That is, the ROI centroid is set to the predicted position  $(x_k^-, y_k^-)$ , and the region size would be the same as that in the  $(k-1)$ -th frame. Finally, the dynamic threshold  $\tau_k$  of the ROI in the  $k$ -th frame is defined as

$$\tau_k = \gamma \|(\Delta_{x_{k-1, k-2}}, \Delta_{y_{k-1, k-2}})\|_2, \quad (3.13)$$

where  $\|\cdot\|_2$  denotes the two-norm operation of vectors and  $\gamma = 5$  is the tolerance factor.



**Figure 3.6:** Example of flooding operations with a  $6 \times 5$  ROI. In (a), the number of each pixel indicates which border it belongs to. In (b), the left and the right pixels of a thick solid line are marked as the background and UIO, respectively. Their valid neighbors are connected with the arrows. (Let  $C_i$  be a color in RGB space and  $d_\theta(C_2, C_3) > T_d$ .)

### Dynamic Determination of ROIs

Sometimes, there are more than one ROI in a video frame. For example, in one view of a tennis game, two players may form two different ROIs. This scenario has to be explicitly addressed. In a saliency map, each ROI usually consists of a set of saliency values peaked at the center of its 3-D profiles. For example, if a video frame has two ROIs (e.g. there are two separate moving persons in Figure 3.4(d)), its saliency map usually has two separate peaked sets, as shown in Figure 3.4(f). We assume that the saliency value ranges from 0 to  $R$  (in this work,  $R = 255$ ). If a pixel's saliency value is greater than a predefined threshold, it is added to the peak set ( $PS$ ). All pixels in the  $PS$  are further grouped via an unsupervised clustering algorithm called the adaptive sample set construction [Bow02]. Euclidean distance is chosen as the similarity measure because it works well when a data set has compact or isolated clusters [JMF99]. Then, the peak



set is divided into several disjoint subsets. That is,

$$PS = \bigcup_{i=1}^n PS_i, \text{ where } PS_i \cap PS_j = \emptyset \text{ if } i \neq j. \quad (3.14)$$

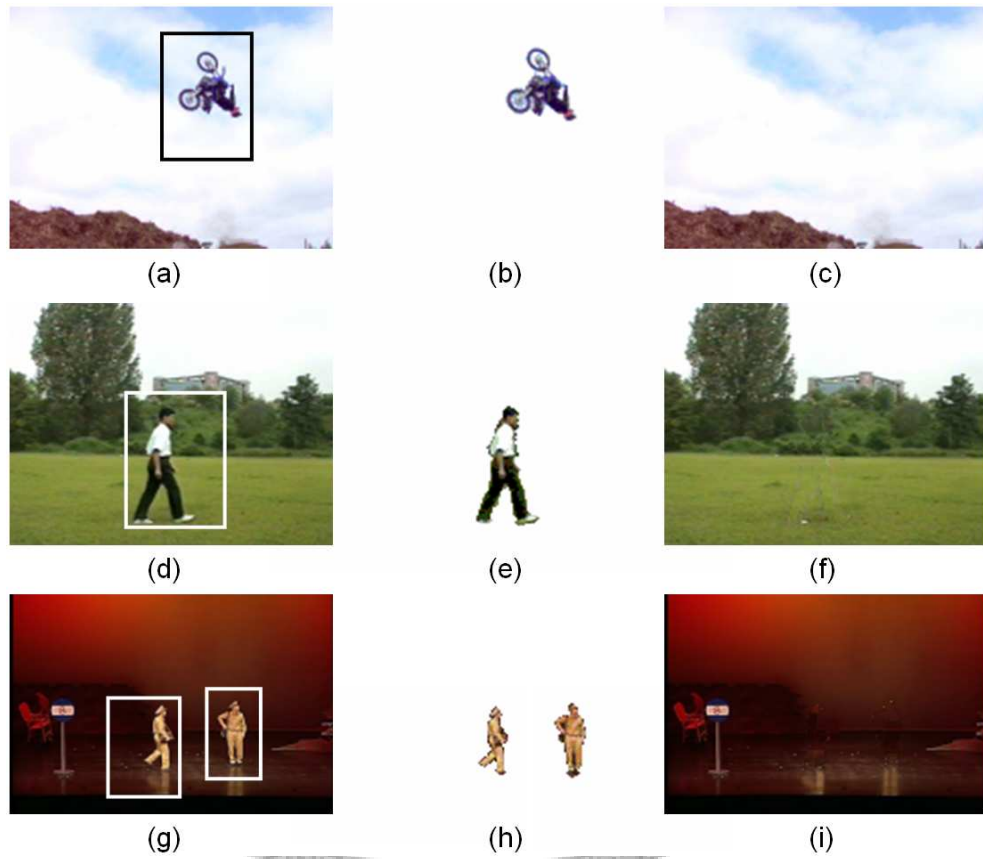
In this way, a saliency map is partitioned into  $n$  regions, and each region corresponds to a peak subset  $PS_i$ . One ROI is declared for each region. With this scheme, the number of ROIs can be automatically and dynamically determined for each video frame.

Note that in practice the number of ROIs is purposely kept no more than three [BT03, Zet98]. Too many attractive objects in a frame will distract the attention of viewers. In such a case, like the scene of a busy street, the global view is preferred over individual objects. Therefore, we terminate the on-going clustering process and determine a single ROI based on the whole saliency map.

### 3.4 Content Recomposition

In this section, we explain in detail the process of recomposing video content to fit within a target screen. After obtaining the ROI information, exact UIOs are separated from the background. Since the removal of UIOs leaves some scene holes on the background, an inpainting algorithm is applied to refill them. To emphasize the UIOs, we increase their relative scales with regard to the scene; meanwhile, to retain the video context, we resize the background to match the screen dimensions. The adapted result is then obtained by reintegrating all of the modified video objects. To ensure the resultant visual rationality, a set of criteria based on media aesthetics are developed for guiding the recomposition.

Before we proceed, a natural question may be raised here is whether all video scenes need to be recomposed. Obviously, if the UIOs have been appropriately emphasized by the author (i.e., large enough) as shown in Figure 3.5(a), other



**Figure 3.7:** Examples of video objects separation. The columns from left to right are successively the original frames with ROIs, extracted UIOs, and repaired backgrounds.

lightweight options such as direct resizing seem to be sufficient. Therefore, we compute an area ratio of ROIs to the whole frame as a simple condition for thresholding (cf. Figure 3.1). The threshold is empirically set to 0.65. For example, if the total area ratio of all ROIs in a frame is greater than the threshold, the direct resizing is applied instead.

### 3.4.1 UIOs Extraction

For simplicity, we assume that each ROI contains one single UIO. Aforementioned, the only difference between ROI and UIO is in the inclusion of partial background or not, cf. Figure 3.5. In this definition, a UIO can possess one to several semantic objects. For example, in Figure 3.5(b) the car itself constitutes a UIO, and in Figure 3.5(a) the car together with a man form another one. Since the ROI information has to be available for all video frames, the extraction task is transformed to explicitly segment the UIO mask from its corresponding regions. For video presentation, the appearance of each frame is very short to the viewer so that the segmented results need not to be perfect, and the efficiency (i.e., the processing speed) seems to play a more important role. Therefore, we apply a real-time flooding procedure to mark the redundant background parts of an ROI [CV05], in this work.

Conceptually, an ROI is composed of a set of non-overlapped rectangular borders with one-pixel width. For explanation, these borders are successively numbered as 1 to  $n$  from the most exterior one, e.g., the case of  $n = 3$  is shown in Figure 3.6(a). Initially, all pixels of the first border are marked as the background. Next, every pixel of the second border is compared with its neighbors that belong to the previous adjacent border. The valid neighbors are those not 2 pixels away from the target pixel, i.e., within a  $3 \times 3$  support window (cf. Figure 3.6(b)). If

the difference from any of its neighbor pixels is less than a fixed threshold  $T_d$ , it is marked as the background, otherwise the UIO. Figure 3.6(b) gives an example. The same process continues through out the following borders. Meanwhile, two stop conditions are set and either one will end the flooding. The first condition is when it reaches the  $n$ -th border, i.e., all pixels of the ROI have been scanned. The second one is when all pixels of the checked border at hand are marked as the UIO, i.e., it has got into the UIO interior and no more background pixels are left. Finally, the desired mask is obtained and used for extracting the UIO. Some examples of UIO extraction are shown in Figure 3.7.

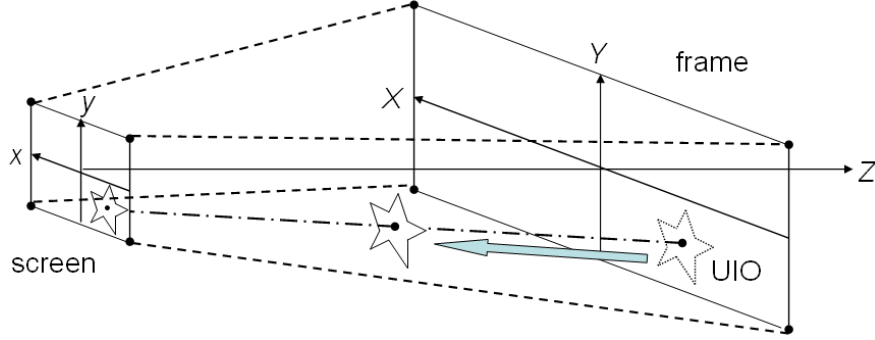
In the above, the difference of any neighboring pair  $(p_1, p_2)$  is defined by the *color vector angle* [LLH03], whose value is computed as

$$d_\theta(C_1, C_2) = \left( 1 - \frac{(C_1^T C_2)^2}{C_1^T C_1 C_2^T C_2} \right)^{1/2}, \quad (3.15)$$

where  $C_1$  and  $C_2$  are the RGB color vectors of  $p_1$  and  $p_2$ , respectively. In addition, the threshold  $T_d$  is set to 0.09 as suggested in [LLH03]. The similarity measure  $d_\theta$  is adopted for its insensitive to variations in intensity, yet sensitive to differences in hue and saturation. This property is useful for identifying meaningful object edges.

### 3.4.2 Background Repairing

To repair unfilled scene holes left by UIOs extraction, we develop an exemplar-based inpainting algorithm based on the work [CPT04], in which the visible parts of a frame serve as a source set of exemplars (i.e., image patches) to infer the target region. Unlike other kinds of inpainting schemes, the missing data is replicated rather than synthesized from available information. It is superior in reducing blurring artifacts. More importantly, it has better speed efficiency [CPT04]. Some



**Figure 3.8:** The virtual 3-D scene model. All video objects of a frame are re-projected onto a target screen. An object is perceived larger (i.e., the star-shaped UIO) while it comes closer to the screen.

examples are shown in Figure 3.7. The detailed algorithm and more corresponding results are reported in our previous work [CHL<sup>+</sup>05].

Based on the experiments, it is worthy to notice that most parts of a scene hole will be covered with the re-pasted UIOs. Few inpainted pixels, especially those along the scene hole boundary, are visible to viewers. Therefore, it is generally safe to moderate the computational overhead by merely repairing a partial region, e.g., we empirically suggest about 50% of the whole. For the same reason, we adopt image-based rather than video-based inpainting algorithm. The later often requires complex spatial-temporal analysis of videos, such as exact camera registration [PSB05, CFJ05]. For our applications, there is not much gain in doing so.

### 3.4.3 Media Aesthetics Based Video Objects Reintegration

As the final step, we reintegrate all of the separated video objects for content recomposition. Since the background is directly resized to match the target

screen size, the reintegration task becomes making a proper arrangement of enlarged UIOs on the resized background. The relevant issues of a UIO include the following two points: one is where to paste it and the other is how large it should be. The discussion is difficult for its subjective nature [CV05]. Careless handling often incurs negative effects on the visual rationality. That is, the visual structure of a content would be altered to distort the conveyed message, e.g., Figure 3.2. Fortunately, media aesthetics is an efficient process of examining media elements for identifying their roles in manipulating human perceptual reactions and synthesizing effective media productions [BT03, Zet98, DV01]. It provides us a reliable basis for the automatic decisions-making in an objective and rational way.

### Determination of UIO Positions

From the viewpoint of media aesthetics, the idea of increasing the relative scales of UIOs acts as enhancing the *depth cue* of a video scene [Zet98]. If an object is larger in the scene, it seems closer to the viewer. The relation can be described with a virtual 3-D scene model [WOZ01] (cf. Figure 3.8). It is obtained by using the perspective projection camera model. The major advantage of this model is its ability to enable transformations between different aspect ratios, as illustrated in Figure 3.8. In this way, the corresponding coordinates of each pixel in a frame can uniquely be determined on the screen. Accordingly, we take the centroid of a UIO, say  $(X_c, Y_c)$ , as its control point to obtain the target position  $(x_c, y_c)$ .

$$(x_c, y_c) = \left( \frac{Width_S}{Width_F} X_c, \frac{Height_S}{Height_F} Y_c \right), \quad (3.16)$$

where  $Width_F$ ,  $Width_S$ ,  $Height_F$ , and  $Height_S$  are widths and heights of the source frame and the target screen, respectively.

### Determination of UIO Size

For emphasizing the perceptual feeling, we would like to make the perceived UIOs as large as possible. However, to ensure the recomposed visual rationality, we introduce some aesthetic criteria to upper-bound the limits. For explanation, video objects of the  $f$ -th frame are represented as follows.

$$VO_f = \{BG_f, UIO_{f,1}, \dots, UIO_{f,k}\}, 1 \leq k \leq 3, \quad (3.17)$$

where  $BG_f$  and  $UIO_{f,i}$  are the repaired background and the  $i$ -th UIO, respectively. Let  $BG_f^R$  be the resized  $BG_f$  to match the screen size, and  $UIO_{f,i}^R$  be the enlarged  $UIO_i$  by a scaling factor  $r_{f,i}$ . Conceptually, they are the “projected” images of the video objects on the screen surface (cf. Figure 3.8). Since the aspect ratio of each UIO is kept fixed to avoid shape distortion, the same scaling factor is applied to both of its dimensions. In the following, we describe in detail and formulate each of the adopted aesthetic principles.

1) *Principle of Object Closure*: When showing only part of an object on the screen, we must frame the objects so that the viewer can easily fill in the missing parts and perceive the whole. That is, enough parts of the enlarged UIOs should be visible on the screen  $S$  to be faithfully recognized. This principle can be formulated as

$$\frac{|UIO_{f,i}^R \cap S|}{|UIO_{f,i}^R|} > \delta, \quad (3.18)$$

where  $\delta$  is empirically set to 0.9 and  $|A|$  denotes the size of a video object  $A$ .

2) *Principle of Overlapping Planes*: When an object is partially covered by another, we perceive that the one that is doing the covering must be in front of the one that is partially covered. That is, the enlarged UIOs should not be overlapped since they are originally non-occluded. This principle can be formulated as

$$UIO_{f,i}^R \cap UIO_{f,j}^R = \phi, \quad i \neq j. \quad (3.19)$$

3) *Principle of Relative Size*: When the relative size of an object is smaller than another, we perceive the smaller one as being farther away and the larger one as being closer. That is, the relative size of the enlarged UIOs should be consistent with that of the originals to keep their inter-relationship. This principle can be formulated as

$$\frac{|UIO_{f,i}|}{|UIO_{f,j}|} = \frac{|UIO_{f,i}^R|}{|UIO_{f,j}^R|}, \quad i \neq j. \quad (3.20)$$

It implicitly implies that  $r_{f,i} = r_{f,j}$ .

4) *Principle of On-screen Continuity*: When the visual setting of a scene has been established, we must keep its consistency in the following frames to maintain the viewer's mental map. That is, the size changing pattern of the enlarged UIOs should be consistent with that of the originals. This principle can be formulated as

$$\frac{|UIO_{f+1,i}|}{|UIO_{f,i}|} = \frac{|UIO_{f+1,i}^R|}{|UIO_{f,i}^R|}, \quad (3.21)$$

where  $UIO_{f+1,i}$  corresponds to the same object of  $UIO_{f,i}$  in its next frame. It implicitly implies that  $r_{f,i} = r_{f+1,i}$ .

From the principles 3 and 4, we know that values of the scaling factor are identical for all UIOs within the same shot, which effectively reduces the solution space for exploration. To avoid biasing, we compute a valid value range  $[r_{min}, r_{max}]$  for each frame (as described in the following) and take the maximum from their intersections as our final result  $r^*$ . In this way, all enlarged UIOs are promised to satisfy the adopted aesthetic criteria. Obviously, the smaller scaling factor of the background would be the natural lower bound, i.e.,

$$r_{min} = \min \left( \frac{Width_S}{Width_F}, \frac{Height_S}{Height_F} \right). \quad (3.22)$$

If the obtained scaling factor  $r^*$  is roughly equal to  $r_{min}$ , the whole frame seems to be directly resized while the aspect ratio of UIOs is kept constant. Generally,



we search the possible valid maxima  $r_{max}$  for each frame by linearly increasing the value of  $r_{min}$ . Specifically, the value  $r_{max}$  of a frame is the maximum of possible scale factors that satisfying the adopted aesthetic principles. The search precision ( $sp$ ) depends on the speed efficiency requirement of the application. In this work, it is empirically set to 0.1. Let  $a$  and  $b$  be the indices of the first and the last frames of a shot, respectively. The process to obtain  $r^*$  of the shot can be summarized as follows.

- Step 1 (Initialization):  $[r_{min}, r_{max}] \Leftarrow [\min(\text{Width}_S/\text{Width}_F, \text{Height}_S/\text{Height}_F), \infty]$
- Step 2:
 

```

for  $f = a$  to  $b$  do
   $r_{temp} \Leftarrow r_{min}$ 
  while  $r_{temp} < r_{max}$  and  $r_{temp}$  satisfies ALL the adopted aesthetic cri-
  teria of the  $f$ -th frame do
     $r_{temp} \Leftarrow r_{temp} + sp$ 
  end while
   $r_{max} \Leftarrow r_{temp}$ 
end for

```
- Step 3:  $r^* = r_{max}$

It should be noted that if the additive incremental mechanism take a long time to find the solution, i.e., the selected search precision is very high, other fast algorithms like the variant binary search [CLRS01] can be applied instead.

**Table 3.2:** Screen sizes used in the experiments.

Type	Size (pixel <sup>2</sup> )	Aspect Ratio (AR)
1	240 × 180	4:3
2	208 × 156	4:3
3	168 × 126	4:3
4	120 × 90	4:3

## 3.5 Experimental Results

In this section, we conduct several experiments and compare our results with those of the conventional approaches [CV05, XLS05]. Then, we carry out user studies to verify the effectiveness of the proposed framework. Finally, the time efficiency of our approach is analyzed. Here, the technology of spatial resizing is chosen as the conventional approaches for the following two reasons. First, it is currently the most popular and dominant solution for adaptive video delivery [AWSZ05]. Second, to our best knowledge, although some improved solutions other than the spatial resizing are proposed, there is no relevant work that focuses on video-based applications. Most efforts are put on still images as described in Section 3.2.

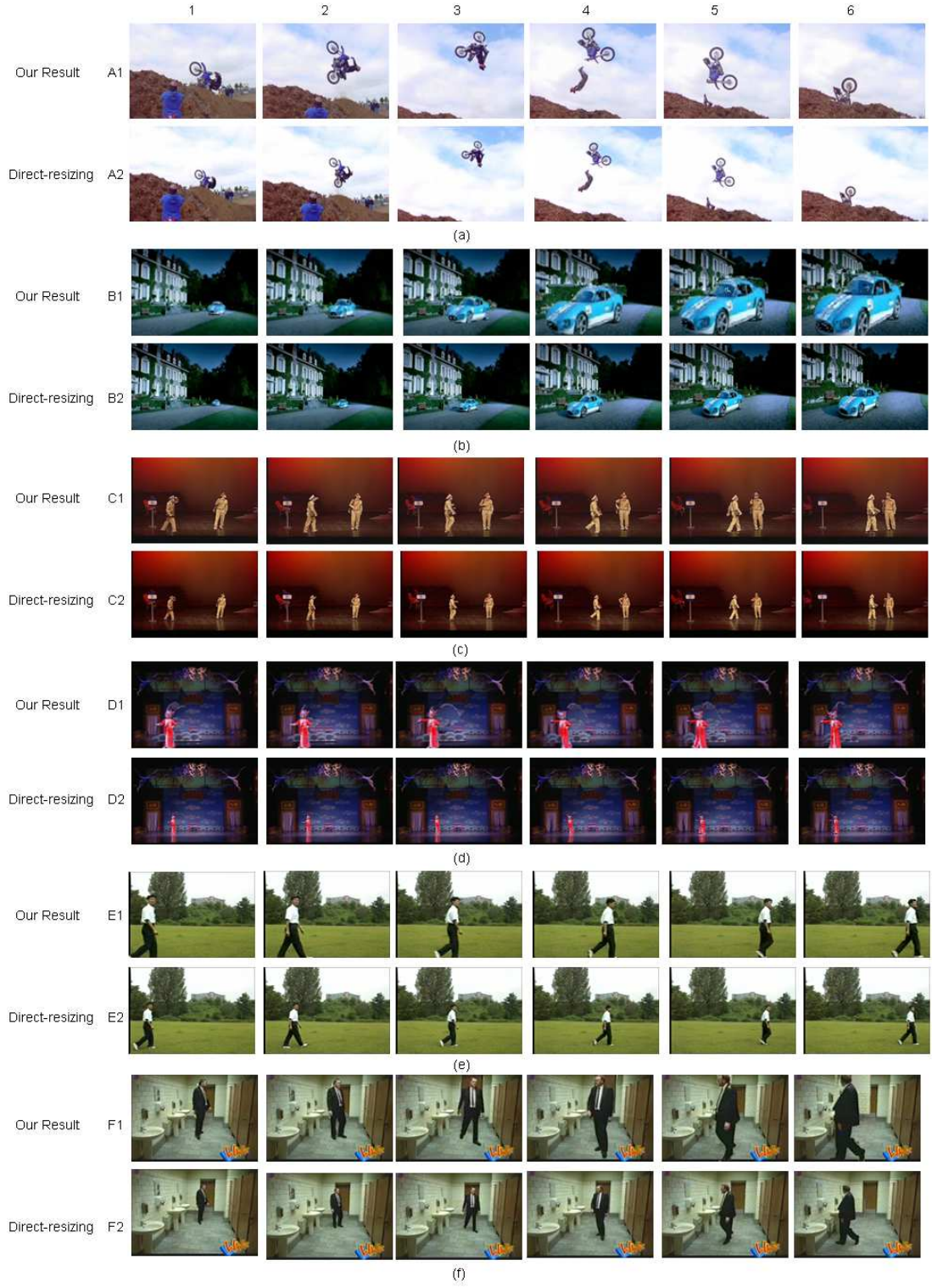
As listed in Table 3.2, four typical screen sizes of hand-held devices are adopted in our work [KMS05]. All of them have a 4:3 AR. On the other hand, we have eight source clips as described in Table 3.3. They are all expert-produced and each of them is about three to five minutes long. Some sample frames of each clip are illustrated in Figure 3.9 and Figure 3.10. Note that we use short clips rather than long sequences in the experiments, because observers' viewing fatigue has been reported to have a severe interference with user study [KMS05]. In this way, it would be affordable to cover more kinds of video data. Further, taking into account the effect of video changes in ARs, the source clips are divided into two subgroups according to their ARs. Each clip of the subgroup 1 (i.e., clips A

to F) is direct-resized into four testing clips with different resolutions as shown in Table 3.2. Also, four testing clips are automatically generated by our approach. In addition to the direct-resizing and recombination, linear-resizing is another adapting choice for the clips of subgroup 2 (i.e., clips G to H) since they have a different 16:9 AR with the adopted screen sizes. By definition, linear-resizing keeps the original video AR intact but direct-resizing changes it to match the adopted screen AR. Therefore, each clip of subgroup 2 will have two kinds of spatial resized testing clips for one specific resolution. In the rest of this section, the term conventional approaches will be used interchangeably to indicate both the direct-resizing and the liner-resizing approaches.

### 3.5.1 Recomposition Results

Figure 3.9 and Figure 3.10 illustrate partial results of our approach and the conventional approaches for clips of the subgroups 1 and 2, respectively. The frames are taken at the resolution format of screen type 3 (i.e.,  $168 \times 126$ ). For explanation, each frame is depicted by the adopted clip number followed by its virtual temporal index, e.g., C2-3. Generally, our approach outperforms the conventional ones based on the following observations:

1. Although the background contexts are faithfully preserved in all results, the key subjects are more effectively emphasized in our approach. In terms of visibility, our approach provides more useful information to the viewers.
2. It can be found that some of our results have lots of gain in the subject visibility and others have moderate improvement. For example, the UIO of Figure 3.9(b) (i.e., the automobile) is emphasized more clearly than that of Figure 3.9(a) (i.e., the motorcycle). The difference is due to the considera-



**Figure 3.9:** Comparison of our approach with the conventional approach (direct-resizing) for the clips of subgroup 1. 62



**Figure 3.10:** Comparison of our approach with the conventional approaches (direct-resizing and linear-resizing) for the clips of subgroup 2.

tion of visual rationality. For example, in frame A1-4, the two UIOs (i.e., the man and the motorcycle) are originally located at a close distance to each other. According to the aesthetic principle of overlapping planes (i.e., Equation (3.19)), mild enlargement is made to keep their spatial interrelationship. Similar phenomenon is observed in the results of Figures 3.9(c) and (d).

3. The UIOs in clips C and D (i.e., the main actors and the actress) have fine gestures, facial expressions, and plentiful body language. They are important visual cues in dramatic performance but most of them are almost unrecognizable with the conventional approach, cf. Figures 3.9(c) and (d). Furthermore, some small things that carry specific meaning are also invisible, e.g., in Figure 3.9(d), the white long feather on top of the actress's head. In contrast, the visibility of these details are improved with our approach. Even if the results are not perfect, they are at least "visible" to the viewers.
4. Our approach is flexible to deal with the case of video AR changes. We fully utilize the valuable space resource of a target screen and keep the UIO AR to avoid degradation of the viewer's visual comfort. Both advantages of the linear- and direct-resizing are effectively integrated in our approach. Figure 3.10 demonstrates two real examples.
5. Sometimes our approach generates new visual artifacts in the recomposed video, e.g., the actress's incomplete feather ornament in the frame D1-2 and the defective boundary of the girl's head in the frame H1-6. Based on the experiments, these artifacts come from the imperfect UIO segmentation. Currently, as described later, we find that most viewers are not well aware of those artifacts. If visual artifact becomes a major concern in the application,

other more accurate segmentation algorithms can be applied to resolve this shortage.

### 3.5.2 User Studies

To evaluate our approach, two user studies (TRIAL-I, TRIAL-II) are separately carried out. The objective of TRIAL-I is to determine if the accompanied content changes of our approach (e.g., enlarged UIOs) are visually acceptable to users, and to determine which of our approach and the conventional approaches is visually preferred. In TRIAL-II, we aim to investigate the effectiveness of our approach in practical usage. The viewer's viewing experience of our approach is compared with that of the conventional approaches on hand-held devices. For reference, the test conditions are listed in Table 3.4. The detailed methodology of each experiment is explained in the following:

#### TRIAL-I

For our purpose, two aspects of the results need to be investigated. One is whether UIOs are effectively emphasized, and another is whether recomposed videos look reasonable. Furthermore, the usefulness of our approach depends on whether it is actually preferred by viewers. Therefore, we apply a pair-comparison technique [CXF<sup>+</sup>03, LG05] to study viewers' visual acceptability and preference. That is, at each time, an observer will be shown two different adapted results of the same video, and asked to subjectively decide which one would be better based on some predefined questions. In this way, it allows us to know the relative advantage and disadvantage of our approach.

In this study, the used testing clips are at the resolution format of screen type 3 as prescribed. Twenty participants are randomly invited in our campus. They





**Figure 3.11:** Example of the displayed web page for TRIAL-I. For reality, both the pair of testing clips are presented on a virtual cellular phone. (See Subsection 3.5.2 for details.)

are in the ages of 20 to 27, all with Chinese as their native language. Before joining the study, they have no ideas about our research work. Since the study will be conducted via the web, every participant is assigned a 17-inch LCD at the viewing distance of 40 centimeters.

Initially, the testing goal, process, and relevant details are explained to the participants, such as descriptions of the predefined questions. For fair comparison, they are not told any details about our video adaptation algorithm, e.g., UIOs are extracted and reintegrated with the background. In addition, they are required to conceal personal interests in different video clips since the study focuses on viewer's visual experience rather than his/her emotional perception. After making sure that all participants understood the instructions clearly, we begin the experiment. At each time, a pair of testing clips (one is generated from our approach and another is from the conventional approaches) is displayed side by side on a web page, cf. Figure 3.11. To be fair, the source videos are not presented to the participants in advance, and names of the corresponding approaches will not be



prompted. Both the order of presentation and which clip to be appeared on the left or right side are independently randomized for each participant. After browsing the pair of clips, the participants are asked to answer the following eight predefined questions (Q1-Q8):

- Q1: In terms of UIOs, which clip is more visible to be recognized?
- Q2: In terms of UIOs, which clip appears with better motion and shape continuity on the screen?
- Q3: In terms of UIOs, which clip would you visually prefer?
- Q4: According to the relative size of video objects, which clip looks more reasonable?
- Q5: According to the interactive behavior of video objects, which clip looks more natural?
- Q6: According to the scene composition, which clip would you visually pleasant?
- Q7: Generally, for content comprehension, which clip would be more informative?
- Q8: Generally, for browsing on your hand-held device, which clip would you prefer to receive?

For each of the questions, three given comments are allowed for participants to choose as their answer: “the left one is better”, “no difference”, and “the right one is better”. Note that the answering time is unrestricted and the pair of clips are allowed to be repeated. The same process continues until all combinations of possible clips are tested for each participant.

For our testing purpose, Q1 to Q3 concentrate on the UIO itself. Specifically, Q1, Q2, and Q3 examine whether our UIOs are visually emphasized, acceptable, and preferred to viewers, respectively. Here, the acceptance refers to user-perceived motion smoothness and shape consistency of UIOs, which is affected by adopted underlying algorithms, such as the UIO segmentation. Therefore, Q2 in some sense serves as a performance index of our system. Next, Q4 to Q6 relate to the user-perceived visual rationality of the whole recomposition. The static and dynamic visual perceptions are individually explored in Q4 and Q5. Further, Q7 explores the assistance in content comprehension and helps us to know the functional role of our approach. Finally, Q8 investigates whether viewers would like to receive recomposed videos in practical applications, which demonstrates the usefulness of our approach.

Table 3.5 shows the statistical results of our approach. According to the categories of clip subgroups and competitive approaches, the results are further divided into three sub-tables (Tables 3.5(A)-(C)). Note that the fourth “worse” column denotes the percentage that the competitive approaches are chosen as better by viewers. For reference, we compute the weighted value  $\mu_{RP}$  as an index of the user’s relative preference (RP) to our approach, that is

$$\mu_{RP} = ((+1) \cdot f_b + 0 \cdot f_n + (-1) \cdot f_w) / 100, \quad (3.23)$$

where  $f_b$ ,  $f_n$ , and  $f_w$  are the “better”, “no difference”, and “worse” percentages for a specific question in a sub-table, respectively. Clearly,  $\mu_{RP}$  is in the range of  $[-1, 1]$ . If the value is positive, our approach would be more preferred by users, otherwise the conventional approaches. The RP strength is measured by its absolute magnitude, i.e.,  $|\mu_{RP}|$ . Meanwhile, a corresponding RP variance  $\sigma_{RP}$  is estimated.

According to Q1's statistics in Table 3.5, our approach is really helpful to improve the visibility of UIOs for viewers. As shown in Q3's statistics, most of the viewers also prefer such an improvement, but there is a 10%-40% decrease in the "better" percentage and the RP variance is high. Based on our observations, it is mainly caused by two reasons: First, the motion and shape continuity of emphasized UIOs is not perfect in our approach, e.g., shape inconsistency of the actress's feather tail in clip D1 as prescribed. As shown in Q2's statistics, some viewers are displeased to this kind of artifacts (e.g., there is a 13.33% "worse" in Table 3.5(A)) and would rather visually prefer the conventional approaches with smaller UIOs. Second, the effectiveness of UIO emphasis is content dependent. For example, in Figure 3.10, it is useful to emphasize the boy of clip H for showing his important details to viewers, such as the facial expression. However, in Figure 3.9, it becomes less meaningful for the car of clip B since viewers can easily recognize its appearance even in a smaller form. In this case, our approach is not specially preferred by viewers. That is also the reason why we have a lower "better" percentage (53.33%) and a higher "no difference" percentage (30.00%) of Q3 in Table 3.5(A) than those in Tables 3.5(B) and (C).

Further, according to statistics of Q4 and Q5 in Table 3.5, the visual rationality of our approach is generally acceptable to viewers. We find an exception is in Q4's statistics of Table 3.5(C). One reason is due to the artificial essence of our approach. Since we recompose videos with software-based techniques rather than real video reshooting, the visual rationality of our approach could not be so realistic as that in the original. Another reason is that the object distortion caused by AR change is undesirable to viewers, which makes the perceived relative size of video objects visually unreasonable. It is found that the shape distortion seems more intolerable to viewers. For example, compared with the Q4 statistics

in Table 3.5(A), the “worse” percentage decreases to 2.50% in Table 3.5(B), but increases to 52.50% in Table 3.5(C). The more the video objects are distorted in AR, the more the relative size of them looks unreasonable. An interesting phenomenon is that even if the viewers are aware of those visual imperfection, in average, they still prefer the scene composition of our approach, cf. Q6’s statistics in Table 3.5. Notice that the terms “better” and “worse” in Table 3.5 do not mean absolute success or failure but the relative performance of the proposed or the conventional approaches. Therefore, we can say that although the visual rationality of recomposed videos is not perfect, it does not fall far short either when compared with the original one. The recomposed visual quality seems good enough to be accepted by most viewers.

Finally, in Table 3.5, Q8’s statistics show that most of the participants are willing to receive our results for practical usage. Furthermore, as shown in Q7’s statistics, our approach improves the viewer’s comprehension of video contents; however, the corresponding RP variance ( $\sigma_{RP}$ ) is high, as shown in Table 3.5(B). From the perspective of information delivery, this makes it plausible that viewers would prefer our approach for its informative benefits that attributed to the enhanced visibility of important details. Overall, while the results of our preliminary experiments may be inconclusive, we find it encouraging. The proposed approach seems really helpful to improve the video experience for mobile users. Also, the mobile users would prefer our approach to obtain the improvements.

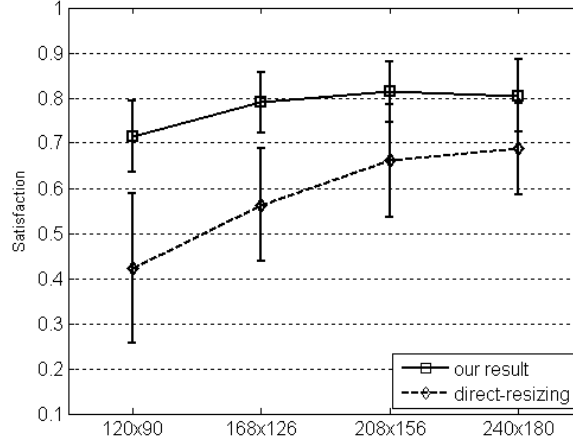
## TRIAL-II

To further evaluate the effectiveness of our approach in practical use, we carefully design the experiment to assess viewers’ subjective satisfaction with the viewing experience on hand-held devices. Specifically, the satisfaction refers to

the level that a viewer is satisfied with his/her viewing experience of an adapted video on a hand-held device when compared with that of the original video on a standard display. For our purpose, overall, the viewing experience is defined as the user-perceived detail visibility, visual rationality, and browsing ease of video content. Specifically, the detail visibility indicates the user-perceived clarity of small objects or things in a scene; the visual rationality relates to the user-perceived relative size and interactive behavior of video objects, cf. TRIAL-I; the browsing ease refers to if the user can comfortably view a whole video. In this way, we are able to measure how effective our approach would be in improving the viewing experience for mobile users.

In this study, the used testing clips are at all resolution formats as described in Table 3.2. To ensure the validity, another twenty participants different from TRIAL-I are randomly invited. A 17-inch LCD and a Dopod 900 Smartphone (with a 3.6-inch LCD) are assigned to each participant as the testing platforms. They are set at the viewing distance of 40 and 30 centimeters, and treated as the standard display and the hand-held device, respectively.

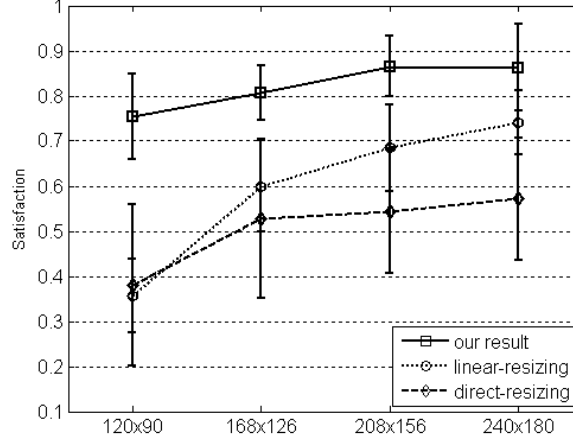
Initially, the testing purpose, process, and relevant details are explained to the participants, e.g., definitions of the viewing experience and subjective satisfaction. For fair comparison, they are not told any details about our video adaptation algorithm and required to conceal personal interests in different video clips, as like in TRIAL-I. Then, one of the source clips at its original format (cf. Table 3.3) is shown to participants through the standard display. To avoid viewers' misconception, the participants are asked to read the corresponding content descriptions in Table 3.3. The playing is repeated until all of the participants have well understood the video content. Next, all the corresponding testing clips of that source clip, one at a time, are presented on the hand-held device. For each of the test-



**Figure 3.12:** Comparison of the user study between our approach and the conventional approach (direct-resizing) for the clips of subgroup 1 at different resolution formats.

ing clips, when the playing is finished the participants have one minute to give a subjective score in the range of 0 to 1 with two decimal places at most, e.g., 0.75. The score value is proportional to the viewer's relative satisfaction with that clip as prescribed. For example, if the perceived viewing experience for a viewer is about the same as that on the standard display, the viewer will give a large score value, otherwise a small one instead. Note that the replay is inhibited and the answering time is restricted since we believe the viewer's first impression without reconsideration reveals his/her true satisfaction about the viewing experience. To avoid biasing, testing clips of the same resolution format are displayed in series and at a random order. In the following, the same process is conducted for all of the other source clips.

Figures 3.12 and 3.13 illustrate the statistical comparisons of the user studies between our approach and the conventional approaches for the clips of subgroup



**Figure 3.13:** Comparison of the user study between our approach and the conventional approaches (direct-resizing and linear-resizing) for the clips of subgroup 2 at different resolution formats.

1 and subgroup 2, respectively. In the figures, each of the points is obtained by averaging the participants' satisfaction of the adapted results of an approach at a fixed resolution format. The symmetric error bar indicates two standard deviation units in length. In addition, the techniques of hypothesis testing are applied to obtain the statistical significance ( $P$ -value) of our approach [Dev95]. Since the claim is that the viewer's satisfaction with our approach is higher than that of the conventional approaches, the  $P$ -value provides the probability that the difference (i.e., improvement) in the experiment happened by chance [pva]. For each resolution format in Figure 3.12 and Figure 3.13, we compute a  $P$ -value between our approach and one conventional approach using the upper-tailed  $t$ -test with  $n - 2$  degrees of freedom [Dev95], where  $n$  is the number of observed viewers' satisfaction. Specifically, for each resolution format, we obtain one  $P$ -value between our approach and the direct-resizing in Figure 3.12. Similarly, we

obtain two  $P$ -values (one between our approach and the linear-resizing, another between our approach and the direct-resizing) in Figure 3.13. The  $P$ -value results show that except for the cases at resolution  $240 \times 180$  with the linear-resizing and at  $168 \times 126$  with the direct-resizing in Figure 3.13 (i.e., 0.006 and 0.001, respectively), the other  $P$ -values are far less than 0.001.

Generally, according to the average satisfaction in Figure 3.12 and Figure 3.13, our approach outperforms the conventional approaches in all cases. It is found that the satisfaction of our approach remains high (above 0.7) throughout all resolution formats, but that of the conventional approaches drop rapidly down to an unacceptable level as the screen size decreases. This phenomenon indicates that important visual details (e.g., UIOs) have dominant effects on the viewing experience, which confirmed the statements given in [KMS05]. Figure 3.13 exhibits another fact that viewers prefer linear-resizing to direct-resizing when there is an AR mismatch between the source video and the target screen. As also indicated in TRIAL-I, it is interesting to find that although the linear-resizing wastes a large amount of screen space, the UIO distortion seems more intolerable to viewers. In summary, our approach is helpful to maintain acceptable video property and generates comfortable viewing results for viewers.

### 3.5.3 Time Efficiency Analysis

We analyze the time efficiency of our approach by logging the computational time costs. Without loss of generality, we only include the time costs for clips of the subgroup 1. The proposed framework is programmed using Matlab 6.5. Our test bed is Acer VT7600 PC with Intel P4 3.0 GHz CPU, 1.0 GB memory, and MS Windows XP system. The average processing time for recomposing a  $320 \times 240$  video frame is currently about 21 seconds. The time cost of each underlying



component is shown in Table 3.6. Obviously, the inpainting algorithm is the most time-consuming one. It takes more than 95% of the total time. However, as prescribed, we can reduce the time cost by merely repairing less than half parts of a scene hole. Moreover, with the help of some advanced techniques, such as the program porting to compiled languages like C/C++, code optimization, and system on chip (SoC) design, our approach could achieve real-time performance with confidence. For example, the technique of field-programmable gate array (FPGA) has been recently adopted by some researchers as a fast and low-cost way for creating real-time software applications [GNTD06, DRP<sup>+</sup>06]. In other words, the proposed framework is general and practical enough to be employed on various kinds of adaptive content delivery systems.

### 3.6 Summary

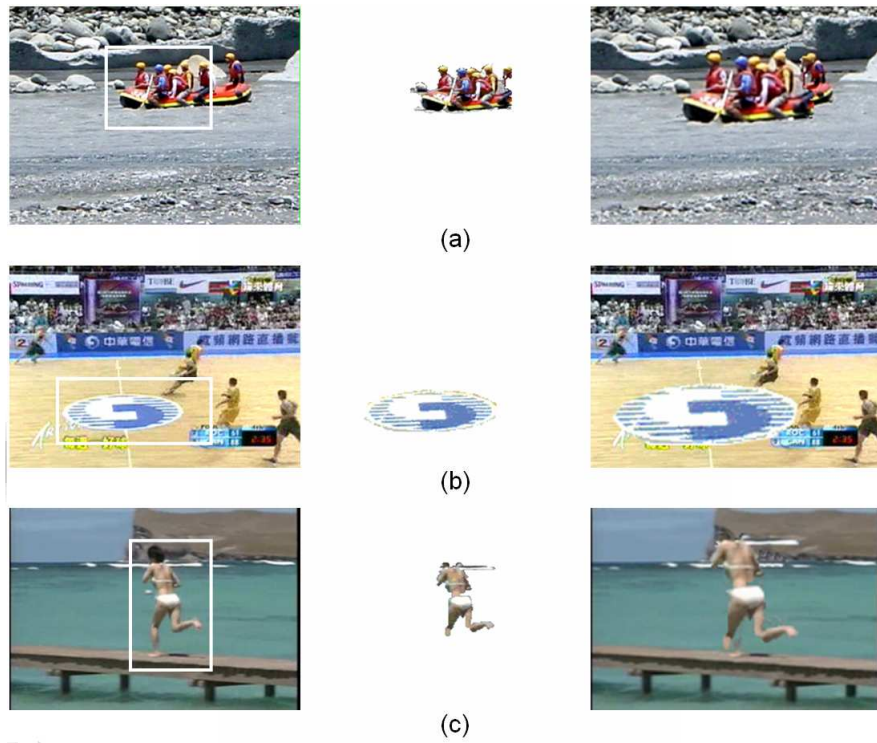
This chapter presents a novel framework for video adaptation based on content recomposition. Our approach is superior to existing schemes in that it emphasizes the important aspects of a scene while faithfully retaining the background context. It also considers the visual rationality of recomposed content and is robust to video changes in aspect ratio. Therefore, the proposed framework can provide more effective and informative video experience to viewers, in an automatic way.

Many aspects of our approach can be improved. For example, currently, we have a fixed expansion factor in the ROI determination module. A risk is that actual semantic objects may not be completely contained in a determined ROI, e.g., Figure 3.14(a). The phenomenon partially comes from the fact that the visually salient regions are not exactly corresponding to semantic objects. It is one essential limitation of the visual attention models [IKN98, PS00]. Therefore, a promising direction for future research is to integrate the proposed framework

with other semantic-level techniques of video understanding and computer vision, such as the “task-relevance map” [NI02] which tells where human eye’s attentions are voluntarily focused on one or more objects that are predefined or meaningful goals to the viewers. For example, in Figure 3.14(b), the detected ROI (i.e., the ground logo) does not match the viewer’s semantic attention. Another example is the UIOs extraction. Since the automatic and precise object segmentation from normal videos is extremely difficult and still an open problem [Sar, ZWL01], we use a simpler algorithm to trade segmentation accuracy for processing time. However, as showed in the user studies, the segmentation accuracy does have impacts on the viewer’s perceptual satisfaction. The development of a robust segmentation algorithm will be one of our future research directions. Another failure example is shown in Figure 3.14(c). Since the woman’s head color is more similar to that of the cliff surface rather than that of her body skin, it is segmented as a part of the background and erroneously separated from the body. Besides, speed efficiency is still a big issue in our approach. The underlying components should be effectively optimized and efficiently coupled together. Finally, it is obvious that the recomposed results will be better if we can “discuss” (i.e., interact) with the content authors in some ways. Therefore, the proposed framework should be integrated with the standardized description schemes for content authors to specify some usage rules. For example, the fifth part of MPEG-21 [BGP03, BdWH<sup>+</sup>03] specifies a machine-readable Rights Expression Language (REL) for declaring rights and permissions, which provides mechanisms to protect digital contents and honors the rights of content authors. In addition, the tenth part Digital Item Processing (DIP) specifies Digital Item Methods (DIMs) as a way for content authors to provide manipulation suggestions of a digital content.

More extensive and complete evaluation of our approach is of importance. One task is to assess the viewer's perceptual response to the recomposed content. We believe that UIOs should be highly emphasized to provide more important information, but we have no idea whether it is appropriate to all kinds of video data and where is the threshold limit value (TLV) of viewers' perceptual comfort. Specifically, the TLV represents the subjective limit that viewers would like to accept such an emphasis. Its investigation assists in clarifying the application scope of our approach. The influence of accompanying audio in the viewer's visual experience is another issue. The study of human visual and aural perceptual interaction would be very helpful. Besides, a fundamental problem is the lack of standardized testing video database. In the experiments, we have attempted to describe and illustrate all of our testing clips as clearly as possible. However, if the number is largely increased, it will be tedious to do this and hard to reproduce the experiments.

A limitation of our approach is that the modified spatial cues (e.g., scene depth and object size) of videos may not be acceptable to some applications, e.g., sports programs, medical teaching clips, astronomical observing videos, etc. Specifically, our approach is unsuitable for those accuracy-sensitive or distortion-intolerant applications. Another limitation is that the adaptation is performed only in the spatial domain. Although it is already highly useful in most existing application scenarios, more flexible and economic methods should be studied further. For example, spatio-temporal based recomposing techniques are effective to reduce the computational overhead. Besides, corresponding methods in the compressed domain are always required for practical demands. The synchronization between the adapted video and its original audio tracks is also an untouched issue. In the future, we will continue our investigation in these directions.



**Figure 3.14:** Failure examples of our approach. The columns from left to right are successively the original frames with ROIs, extracted UIOs, and recomposed frames.

**Table 3.3:** Source clips used in the experiments.

Clip	Resolution	AR	Duration	Content description
A	320 × 240	4:3	232 sec.	A motorcycle with a man flying into the sky and dropping down on a hill.
B	320 × 240	4:3	245 sec.	An automobile gradually approaches from a distant place.
C	320 × 240	4:3	196 sec.	Chinese comic dialogue: two actors interact with plentiful facial expressions and body languages.
D	320 × 240	4:3	277 sec.	Chinese opera: an actress performs with fine gesture on the stage.
E	320 × 240	4:3	168 sec.	A man walks leisurely in the park from the left to the right sides.
F	320 × 240	4:3	217 sec.	A man enters the bathroom and looks around.
G	640 × 360	16:9	258 sec.	Including one scene of the film <i>Team America - World Police</i> : a police fights against a terrorist.
H	640 × 360	16:9	323 sec.	Including three scenes of the film <i>Homerun</i> : a boy runs, two student soccer teams negotiate, and the boy and his sister happily walked together.

**Table 3.4:** Test conditions of the user studies. (See Subsection 3.5.2 for details.)

	TRIAL-I	TRIAL-II
Methodology	pair-comparison	score-rating
Display Device	17" LCD	17" LCD & 3.6" Smartphone
Viewing Distance	40 cm	40 cm & 30 cm
Video Resolution	screen type 3	all screen types (see Table 3.2)
Testee Number	20	20

**Table 3.5:** User study of the relative preference (RP) of our approach with regard to the conventional approaches.

	Better	No Diff.	Worse	$\mu_{RP}$	$\sigma_{RP}$
(A) OUR APPROACH VERSUS THE DIRECT-RESIZING for the clips of subgroup 1					
Q1	<b>95.00%</b>	3.33%	1.67%	+0.9333	0.0972
Q2	10.00%	<b>76.67%</b>	13.33%	-0.0333	0.2362
Q3	<b>53.33%</b>	30.00%	16.67%	+0.3667	0.5751
Q4	6.67%	<b>63.33%</b>	30.00%	-0.2300	0.3180
Q5	3.33%	<b>63.33%</b>	33.34%	-0.3000	0.2814
Q6	<b>51.67%</b>	38.33%	10.00%	+0.4167	0.4506
Q7	40.00%	<b>56.67%</b>	3.33%	+0.3667	0.3040
Q8	<b>83.33%</b>	10.00%	6.67%	+0.7667	0.3175
(B) OUR APPROACH VERSUS THE DIRECT-RESIZING for the clips of subgroup 2					
Q1	<b>90.00%</b>	7.50%	2.50%	+0.8750	0.1635
Q2	7.50%	<b>85.00%</b>	7.50%	+0.0000	0.1538
Q3	<b>77.50%</b>	20.00%	2.50%	+0.7500	0.2436
Q4	27.50%	<b>70.00%</b>	2.50%	+0.2500	0.2440
Q5	25.00%	<b>72.25%</b>	2.50%	+0.2250	0.2301
Q6	<b>70.00%</b>	25.00%	5.00%	+0.6500	0.3359
Q7	<b>62.50%</b>	17.50%	20.00%	+0.4250	0.6609
Q8	<b>92.50%</b>	5.00%	2.50%	+0.9000	0.1436
(C) OUR APPROACH VERSUS THE LINEAR-RESIZING for the clips of subgroup 2					
Q1	<b>97.50%</b>	2.50%	0.00%	+0.9750	0.0250
Q2	5.00%	<b>82.50%</b>	12.50%	-0.0750	0.1737
Q3	<b>72.50%</b>	7.50%	20.00%	+0.5250	0.6660
Q4	20.00%	27.50%	<b>52.50%</b>	-0.3250	0.6350
Q5	22.50%	<b>57.50%</b>	20.00%	+0.0250	0.4353
Q6	<b>62.50%</b>	22.50%	15.00%	+0.4750	0.5635
Q7	<b>70.00%</b>	20.00%	10.00%	+0.6000	0.4513
Q8	<b>87.50%</b>	10.00%	2.50%	+0.8500	0.1821

**Table 3.6:** Time efficiency analysis of the proposed framework for recomposing a  $320 \times 240$  video frame.

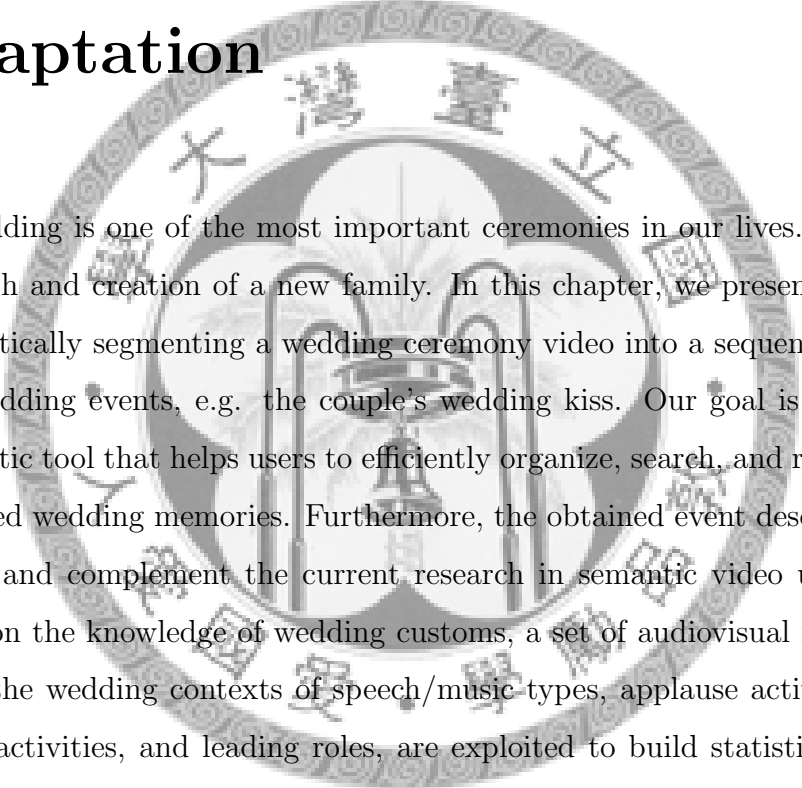
Components	Time (sec.)	Percentage (%)
Attention analysis	0.7500	3.45
ROI determination	0.0780	0.36
UIO extraction	0.1418	0.65
Background repairing	20.6661	95.02
VOs reintegration	0.1135	0.52
Total	21.7497	100





## Chapter 4

# Semantic Event Based Video Adaptation



Wedding is one of the most important ceremonies in our lives. It symbolizes the birth and creation of a new family. In this chapter, we present a system for automatically segmenting a wedding ceremony video into a sequence of recognizable wedding events, e.g. the couple's wedding kiss. Our goal is to develop an automatic tool that helps users to efficiently organize, search, and retrieve his/her treasured wedding memories. Furthermore, the obtained event descriptions could benefit and complement the current research in semantic video understanding. Based on the knowledge of wedding customs, a set of audiovisual features, relating to the wedding contexts of speech/music types, applause activities, picture-taking activities, and leading roles, are exploited to build statistical models for each wedding event. Thirteen wedding events are then recognized by a hidden Markov model, which takes into account both the fitness of observed features and the temporal rationality of event ordering to improve the segmentation accuracy. We conducted experiments on a collection of wedding videos and the promising results demonstrate the effectiveness of our approach. Comparisons with condi-

tional random fields show that the proposed approach is more effective in this application domain.

## 4.1 Introduction

A wedding ceremony is an occasion that a couple's families and friends gather together to celebrate, witness, and usher the beginning of their marriage. It is a public announcement of the couple's transition from two separate lives to a new family unit. Often, the couples invite some videographers, whether professional or amateur, to document the wedding as their treasured memento of the ceremony. In this chapter, wedding videos refer to the raw, unedited footage recorded for wedding. Since a wedding video usually spans hours, the development of automatic tools for efficient content classification, indexing, searching, and retrieval becomes crucial.

In this chapter, we focus on the recognition of a wedding's group actions, namely wedding events, whereby a wedding is interpreted as a series of meaningful interactions among participants. Based on the knowledge of wedding customs [Spa01, War06], we define thirteen wedding events, such as the couple's wedding vows, ring exchange, and so forth. Our goal is to automatically segment a wedding video into a sequence of recognizable wedding events. Without loss of generality, we focus on one of the most popular wedding styles, the western wedding, that follows the basic *western tradition* [Spa01, War06] and takes place in a church-style venue. Based on our observations, a wedding video typically consists of four parts: preparation, guest seating, main ceremony, and reception. For simplicity, we deal with the third part alone because of its relative significance. In the rest of this chapter the term wedding refers to the main ceremony.

In the literature, the study of wedding video analysis has long been ignored. The wedding video is simply to be treated as one of various content sources in research on home videos [GPLS03, ZS05, AYK06]. Although the wedding ceremony video shares some common properties with other kinds of home videos, such as frequent poor-quality contents and unintentional camera operations [GPLS03, ZS05], several characteristics make it much more challenging to be processed and analyzed as indicated in the following:

- **Restricted spatial information:** Since most of the wedding events occur in a single place (e.g. the front of a church altar) and participants basically stay motionless during the ceremony, the conventional techniques based on scene, color, and motion information [GPLS03, ZS05, RS05] are not applicable to pre-partition a wedding video or to group “similar” shots into basic units for further event recognition. Likewise, most of the other content-generic visual features such as texture and edge are not reliable to be utilized.
- **Temporally continuous capture:** The extraction of broken time stamps is a widely used technique for generating shot candidates or event units of home videos [YHZ02, CFGW03]. However, to avoid missing anything important, videographers usually capture a wedding, especially the main ceremony, in a temporally continuous manner without any interruption. As a result, the temporal logs are not useful for wedding segmentation.
- **Implicit event boundary:** Although a wedding ceremony proceeds following a definite schedule, the boundaries between wedding events are often implicit and unclear. For example, a groom’s entering to the venue is sometimes overlapped with the start of the bride’s entering. It is not easy to determine an accurate change point for separating two events. This phenomenon not

only increases the difficulty of accurate video segmentation but also adds uncertainties to annotate the event ground truth.

To recognize the thirteen wedding events, we adopt a set of audiovisual features, relating to the wedding contexts of speech/music types, applause activities, picture-taking activities, and leading roles, as the basic event features to build our wedding video segmentation framework. Each wedding event is represented by a set of statistical models in terms of the extracted features. Since these features are selected based on the understanding of wedding customs [Spa01, War06], they are more discriminative in distinguishing wedding events than the aforementioned features, such as motion and textures, do. To effectively segment a wedding video, we develop a hidden Markov model (HMM) [Bis06], in which every hidden state is associated with a wedding event and a state transition is governed by how likely two corresponding wedding events take place in succession. The event sequence is, therefore, automatically determined by finding the most probable path. In summary, our event recognition framework not only uses the model similarity of extracted features, but simultaneously takes the temporal rationality of event ordering into account.

The main contributions of our work are twofold. First, an automatic system is proposed and realized for event-based wedding segmentation. To the best of our knowledge, this work is the first one to analyze and structure wedding videos at the semantic-event level. Actually, for any type of home videos, it might also be the first one for conducting the semantic event analysis. The methodology could be extensively applied to the other kinds of home videos that possess similar characteristics as wedding, such as the birthday party and school ceremonies. Second, a taxonomy is developed to categorize the wedding events, whereby we adopted a set of carefully selected audiovisual features for robust event modeling

and recognition. The true power of these features is that they are effective in discriminating various wedding events but their extractions from videos are as easy as the conventional ones. Furthermore, the obtained high-level descriptions could benefit and complement the current research in semantic video understanding.

The rest of this chapter is organized as follows. After a discussion of related work, Section 4.3 presents the taxonomy of wedding events. The extraction of event features and the modeling and segmentation of wedding videos are described in Section 4.4 and Section 4.5, respectively. Section 4.6 depicts the experimental results, and Section 4.7 presents our concluding remarks and the directions of future work.

## 4.2 Related Work

In this section, we review previous studies on home video analysis. According to their applications, they are classified into four major categories: scene-based segmentation, capture-intent detection, photo-assisted summarization, and highlight extraction. Meanwhile, their pros and cons as compared with our approach will be briefly discussed as well.

**Scene-based Segmentation.** A basic segmentation process is to cluster relevant shots into groups called scenes. A scene is defined as a subdivision of a video in which either the physical setting is fixed, or when it presents a continuous action in one place [ZS05, RS05]. Since the home video content tends to be close in time, the clustering can be simply confined to adjacent shots. Gatica-Perez *et al.* [GPLS03] proposed a greedy algorithm that initially treats each shot as a cluster and successively merges adjacent ones until a Bayesian criterion is violated. The merging order is determined by both the visual and the temporal similarities, such as color, edge, and shot duration. Zhai *et al.* [ZS05] located the scene boundaries

using the optimization technique – Markov chain Monte Carlo (MCMC). A color-based similarity matrix is constructed for video shots, from which the clusters with high intra- and low inter-similarities are detected as the desired scenes.

**Capture-intent Detection.** A capture-intent refers to an idea, a feeling, theme, or message that makes us to capture certain video segments [AYK06, MHZL07], e.g. a sentimental sunset or baby laughing. Since the user’s capture-intent is often expressed through the use of cinematic principles, some researchers exploit the theory of computational media aesthetics for capturing such intents [DV01]. Achanta *et al.* [AYK06] proposed a framework for modeling the capture-intents of four basic emotions, i.e. cheer, serenity, gloom, and excitement. An emotion delivery system is also developed for helping users to enhance the original or to convey a new emotion to a given home video. Mei *et al.* [MHZL07] further integrated the knowledge of psychology to classify the capture-intents into seven categories, such as close-up view, beautiful scenery, just record, etc. A learning-based mechanism for classifying the capture-intents is then presented using two kinds of feature sets: attention-specific and content-generic features.

**Photo-assisted Summarization.** Personal photo albums can be viewed as an excellent abstract of the corresponding home videos. Both capture most of important moments but photo albums are relatively concise in presenting the contents. Since a still image can be applied to search videos, the summarization task can be casted as the problem of template matching between the two media. Auer-Wolf *et al.* [AWW02] targeted on wedding videos. They represented each shot with one or several mosaics that are used to be aligned with the wedding photos. All shots with successful alignments are collected to generate a summarized wedding video. Similar ideas are adopted by Takeuchi *et al.* [TS06], but they instead estimated the user’s general preferences on the summarization. On

the other hand, Pan *et al.* [PN04] analyzed home videos in a finer unit called a snippet that corresponds to a meaningful camera motion pattern, such as a long static followed by a fast zoom.

**Highlight Extraction.** Highlights are the video segments with relatively higher semantic or perceptual attractions to users. Since the true understanding of video semantics cannot be achieved by the current computing technologies, the study of human attention models provides an alternative way for detecting perceptual highlights [MLZL02, CWW07]. Hua *et al.* [HLZ04] proposed a home video editing system, in which attention-based highlight segments are selected to be aligned with a given piece of incidental music to generate an edited highlight video. Meanwhile, a set of professional editing rules is utilized to optimize the editing quality, e.g. motion activity should match with music tempo. Abowd *et al.* [AGL03] presented a semi-automatic approach for highlight browsing. Home videos need to be manually annotated with a predefined tag hierarchy that helps to group the highlight segments with similar semantic meanings, e.g. clips of all the child's birthday wishing.

Some observations are made from the above discussions. First, the so-called event is a more semantic unit for video segmentation as compared with the conventional ones such as frame, subshot, shot, and scene [CCSW06, LTM03]. It represents a stand-alone human activity during a period of time. However, studies on semantic event analysis of home media are extremely rare as compared with the other kinds of content sources like sports [CCSW06]. Second, the analysis of home media are mostly from the perspective of a viewer or a videographer but not the media owner or event participants. Helping them to explicitly identify what had happened in a video often seems more crucial than simply indicating where

**Table 4.1:** Taxonomy of wedding events

Code	Event	Definition
<i>ME</i>	Main Group Entering <sup>†</sup>	Members of the main group walking down the aisle.
<i>GE</i>	Groom Entering	Groom (with the best man) walking down the aisle.
<i>BE</i>	Bride Entering	Bride (with her father) walking down the aisle.
<i>CS</i>	Choir Singing	Choir (with participants) singing hymns.
<i>OP</i>	Officiant Presenting	Officiants giving presentations, e.g. invocation, benediction, and homily.
<i>WV</i>	Wedding Vows	Couple exchanging wedding vows.
<i>RE</i>	Ring Exchange	Couple exchanging wedding rings.
<i>BU</i>	Bridal Unveiling	Groom unveiling his bride's veil.
<i>MS</i>	Marriage License Signing	Couple (with officiants) signing the marriage license.
<i>WK</i>	Wedding Kiss	Groom kissing his bride.
<i>AP</i>	Appreciation	Couple thanking to certain people, e.g. their parents or all participants.
<i>ED</i>	Ending	Couple (followed by the main group) walking back down the aisle.
<i>OT</i>	Others	Any events not belonging to the above, e.g. lighting a unity candle.

<sup>†</sup> The main group indicates all persons, except the ones in *GE* and *BE*, who are invited to walk down the aisle, e.g. flower girls, ring bearers, groomsmen, bridesmaids, honorary attendants, officiants, etc.

would be more significant. These observed phenomena motivate our development of a comprehensive scheme for event-based video analysis and segmentation.

### 4.3 Wedding Event Taxonomy

According to the western tradition [Spa01, War06], a wedding ceremony, whether religious or secular, begins when an assigned attendant (such as an officiant or bride's mother) is entering down the aisle and ends while the couple is walking out of the wedding venue. The mid-process may vary depending on countries, religions, local customs, and the wishes of the couple, but the basic elements that constitute the western weddings are almost the same [Spa01, War06]. Therefore, we define thirteen wedding events as listed in Table 4.1. They are carefully





**Figure 4.1:** Sample key-frames of the thirteen wedding events.

specified to be mutually exclusive and collectively exhaustive [Dra67]. The corresponding sample key-frames for these events are illustrated in Figure 4.1.

In addition to the traditions, the common perception of the relative event importance is also taken into account in the development of our taxonomy for further applications such as highlight extraction or video summarization. For example, the three entering events (*ME*, *GE*, *BE*) are traditionally to be viewed as a unity called a processional [Spa01, War06], but they should be explicitly separated because the couple's arriving is generally much more exciting than others. By contrast, we classify all of the officiants' formal presentations like invocation and benediction into a single wedding event (*OP*), because they are often invariable in form and the verbal expressions are basically predictable, often not beyond the scope of invoking the God's blessing upon the marriage or inspiring the attendants' religious spirits. It is evident that they are not as important as compared to other events.

Furthermore, as shown in Table 4.1, the taxonomy roughly follows the procession of a wedding ceremony, i.e. from the *ME* event to the *ED* event. However, it should be noted that the actual event ordering is based on each couple's own wedding program and certain events could be repeated or removed in the ceremony. For example, the *OP* and the *CS* events are often interweaved with other

ones. In addition, a simplified ceremony could only contain four events of *WV*, *RE*, *MS*, and *WK*.

## 4.4 Event Features Development and Extraction

Effective event modeling is built on top of reliable event features. The understanding of wedding customs [Spa01, War06] gives valuable insights to the process of feature exploration. Several key observations, which are found to be useful in discriminating the wedding events, are first presented in Section 4.4.1. In Section 4.4.2, guided by these findings, we develop corresponding audiovisual features, including four audio features and two visual features. They are collected together as event features for later event modeling.

### 4.4.1 Key Observations

According to the western traditions [Spa01, War06], wedding events are observed to behave differently in four main aspects: speech/music types, applause activities, picture-taking activities, and leading roles. In the following, we explain in detail for each of the key observations and then give corresponding guidance on the development of relevant event features.

#### Speech/Music Types

Traditionally, some wedding events contain purely speech and others are always accompanied with music [War06]. For example, in the *OP* and the *WV* events, all participants keep quiet to listen to an officiant or the couple speaking. In the *CS* and the *BE* events, a choir is singing with piano accompaniment or the selected background music (e.g. Mozart's Wedding March) is played during the event.

**Table 4.2:** The tendency of wedding events in their behavior of speech/music types, applause activities, picture-taking activities, and leading roles (from the second to the fifth columns, respectively).\*

	S/M <sup>a</sup>	App. <sup>b</sup>	Pic. <sup>c</sup>	Leading Roles <sup>d</sup>
<i>ME</i>	–	N	L <sup>+</sup>	main group
<i>GE</i>	–	N	–	groom, (best man)
<i>BE</i>	M	–	H <sup>+</sup>	bride, (bride's father)
<i>CS</i>	M	–	L <sup>–</sup>	choir, (wedding participants)
<i>OP</i>	S	N	–	officiants
<i>WV</i>	S	N	H <sup>–</sup>	bride, groom, officiants
<i>RE</i>	S	N	H <sup>–</sup>	bride, groom, officiants
<i>BU</i>	S	–	H <sup>–</sup>	bride, groom
<i>MS</i>	–	N	–	bride, groom, (officiants)
<i>WK</i>	–	Y	H <sup>+</sup>	bride, groom
<i>AP</i>	–	Y	–	bride, groom, (wedding participants)
<i>ED</i>	M	Y	H <sup>–</sup>	bride, groom, (main group)
<i>OT</i>	–	–	–	–

\* “–” in the blanks means no obvious tendency.

<sup>a</sup> S: speech events, M: music events.





<sup>b</sup> Y: applause events, N: non-applause events.

<sup>c</sup> L<sup>–</sup>, L<sup>+</sup>, H<sup>–</sup>, H<sup>+</sup>: events with the activity of picture-taking from low to high.

<sup>d</sup> People in parentheses are optional.

The tendency of wedding events in speech/music types is shown in Table 4.2. Obviously, the discrimination between speech and music types from recorded audio plays a key role in wedding event recognition. However, because the quality of the recorded audio is generally poor and often interfered with environmental sound and background noise, the selected audio features related to the speech/music discrimination have to be robust enough to survive such a low-SNR audio input.

**Table 4.3:** Examples of flash distributions of four successive wedding events in a ceremony.\*

1. OP	2. WV	3. RE	4. WK
			
674 (sec)	234 (sec)	142 (sec)	12 (sec)
19 (times)	55 (times)	8 (times)	73 (times)
0.0282 (Hz)	0.2350 (Hz)	0.0563 (Hz)	6.0833 (Hz)

\* The third to the fifth rows are the durations, flash numbers (manually counted), and flash densities of the corresponding wedding events, respectively.

### Applause Activities

Applause is usually expected from wedding attendants as the expression of approval or admiration at certain moments during the ceremony. For example, in the *WK* and the *ED* events, the couple routinely receives a burst of applause at the moments when they are kissing or walking back down the aisle. By contrast, in the *OP* and the *WV* events, wedding attendants rarely applaud in order to keep the solemnity and avoid interfering with the ongoing wedding speech. Thus, effective applause detection is beneficial to the recognition of wedding events, cf. Table 4.2. Note that, for our applications, the applause especially refers to the ones created by a group of people rather than by an individual. Specifically, the applause is generated by the group act of hands clapping and naturally the group members tend to clap at slightly different rates. This phenomenon makes the sound of applause difficult to be analyzed without the use of prior knowledge [CCW03, Rep87]. Therefore, a common technique is to exploit the physical properties of applause [Rep87, PECV07] to identify its appearance in the audio track of wedding videos.

### Picture-taking Activities

Wedding attendants, especially the couple's family members and close friends, often take pictures during the ceremony, and the number of pictures taken roughly represents the relative importance of a wedding event. Table 4.2 illustrates a relative comparison for the generally observed frequency of taking pictures during various wedding events. Since the occurrence of camera flashes correlates closely with the activity of picture-taking [TV01], the estimation of flash density could be an effective visual cue for wedding event discrimination. Table 4.3 shows an example of flash distributions for four successive wedding events in a ceremony. We observed high variations of flash distributions among events. For example, the *WK* event is merely 12 seconds long, but there are 73 flashes. Its density reaches six times per second, on average. By contrast, the *OP* event is of relatively less importance to the audiences, as described in Section 4.3, and it contains a small number of flashes even if it lasts for a much longer duration.

### Leading Roles

As shown in Table 4.2, the leading roles involved in various wedding events are different. For example, groom and the best man are the main characters in the *GE* event; the groom, his bride, and officiants are the main focuses in the *RE* event. The main characters' occurrence pattern gives a visual hint for the event category. A naïve solution would be to recognize all roles in videos. This is, however, not a trivial task with today's technology. Fortunately, there are some simple tricks to detect the bride, inarguably the most important focus of a wedding. According to the western tradition [Spa01, War06], the bride invariably wears white gown and veil as a symbol of purity but the other roles could have some flexibility in their

dress color. Therefore, it is more reliable to indicate the bride's appearance using the truth of her wearing white.

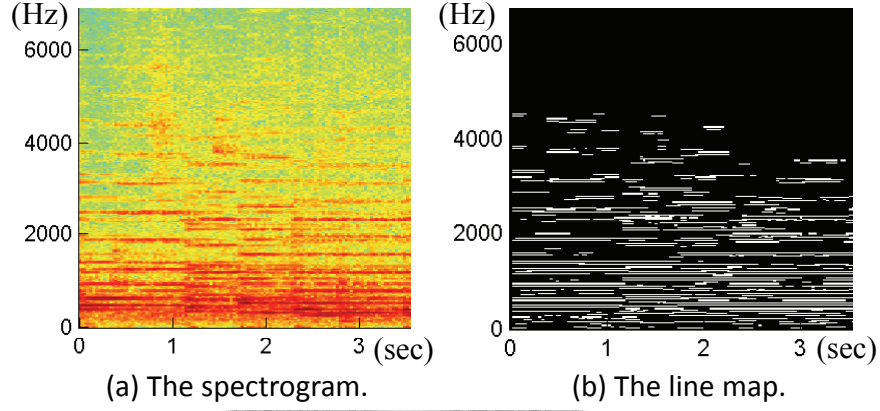
#### 4.4.2 Selected Features for Event Modeling

Based on the observations of Section 4.4.1, four kinds of audiovisual features, related to the scopes of speech/music discrimination, applause detection, flash detection, and bride indication, are developed as basic features for event modeling. In the following, we detail the development for each adopted event features and give their definitions in mathematical forms.

##### Event Features Related to Speech/Music Discrimination

As mentioned in Section 4.4.1, the audio recordings of weddings are often with poor quality. Thus, the selected audio features have to be still distinguishable between speech/music types for the given low-SNR inputs. However, in the literature, most studies address the speech/music discrimination problem only for clean data or with the assumption of known noise types [CCW03, ZK01]. To identify the audio features that are resistant to noises, we first collect a comprehensive set of candidate features from the previous work [CCW03, ZK01, PT05] and determine the more reliable ones using feature selection algorithms [CT06, PLD05].

Initially, tens of audio features are collected to form a candidate set, including the short-time energy, energy crossing, band energy ratio, root mean square (RMS), normalized RMS variance, zero crossing (ZC), joint RMS/ZC, bandwidth, silent interval frequency, mel-frequency cepstral coefficients (MFCCs), frequency centroid, maximal mean frequency, harmonic degree, music component ratio, and so forth [CCW03, ZK01, PT05]. Each of the collected audio features is assessed by information theoretical measures [CT06, PLD05], so as to estimate its discrim-



**Figure 4.2:** Example of a music signal with (a) its spectrogram using short-time Fourier transform and (b) its corresponding line map.

inability between the speech and the music types. At the end, three of them are chosen for their stable performances under various noise types. They are the one-third energy crossing (OEC), the silent interval frequency (SIF), and the music component ratio (MCR), as detailed below. Note that, for extracting the audio features, the audio track of a wedding video is converted to 44,100-Hz mono-channel format first. For simplicity, let  $x(n)$  be a discrete-time audio signal with time index  $n$  and  $N$  denotes the total number of samples in the interval from which features are extracted.

- **One-third Energy Crossing (OEC).** One of the characteristics of a speech signal is that the corresponding amplitude has more obvious variations than that of the music. Given a fixed threshold  $\delta$ , the number of audio energy waveform's crossings over  $\delta$  is often higher in a speech than that in a music. For each audio track, we empirically set  $\delta$  to one-third of the whole range of its average amplitude. Therefore, OEC is defined as a measurement

of the audio's energy-spectral content as follows:

$$\text{OEC} \triangleq \frac{1}{2} \cdot \sum_{n=2}^N |\text{sign}_\delta(x^2(n)) - \text{sign}_\delta(x^2(n-1))| \quad (4.1)$$

where

$$\text{sign}_\delta(a) = \begin{cases} 1, & a > \delta \\ 0, & a = \delta \\ -1, & a < \delta \end{cases} \quad (4.2)$$

As suggested by previous work [PT05, LD06], the audio track is uniformly segmented into non-overlapping 1-second audio frames. For each audio frame, one feature value is computed in every 20-ms interval and these 50 short-time feature values are averaged to generate the representative OEC feature for that 1-second frame. The same mechanism is used in SIF extraction, as described in the following paragraph.

- **Silent Interval Frequency (SIF).** Since a speech signal is a concatenation of a series of syllables, it contains more pronouncing pauses than a music signal does. Therefore, SIF is defined to measure the silent intervals of an audio signal as follows [PT05]:

$$\text{SIF} \triangleq I((ZC = 0) \text{ or } (E < \theta_l) \text{ or } (E < 0.1E_{max} \text{ and } E < \theta_h)) \quad (4.3)$$

where  $I(\cdot)$  is the indicator function,  $E$  is RMS of the signal amplitude, and  $E_{max}$  is the maximum RMS value of the whole audio track. To be precise,

$$E = \sqrt{\sum_{n=1}^N x^2(n)} \quad (4.4)$$

and

$$\text{ZC} \triangleq \frac{1}{2} \cdot \sum_{n=2}^N |\text{sign}_0(x(n)) - \text{sign}_0(x(n-1))|. \quad (4.5)$$



In addition, the two thresholds  $\theta_l$  and  $\theta_h$  are empirically set to 0.5 and 2, respectively. As described in OEC extraction, we compute a representative SIF feature for each 1-second audio frame by taking average of 50 short-time SIF values.

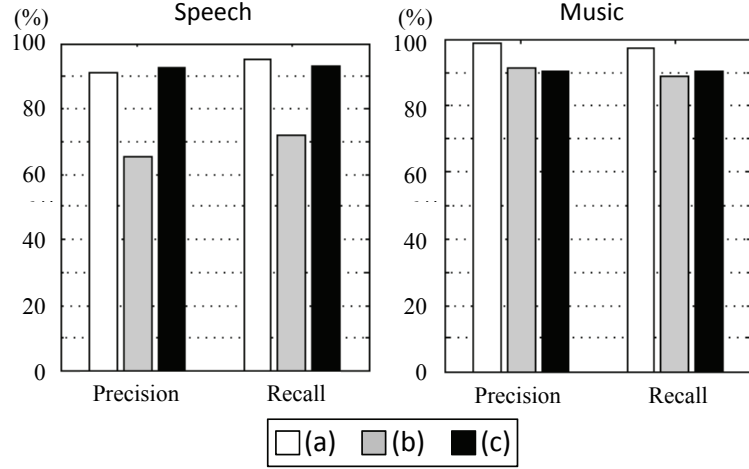
- **Music Component Ratio (MCR).** Harmonicity is the most prominent characteristic of a music signal. A music signal often contains spectral peaks at certain frequency levels and the peaks last for a period of time. This can be observed from the “horizontal lines” in the spectrogram of a music signal, as shown in Figure 4.2. MCR is then defined as the average horizontal line number of an audio spectrogram within a second, and the line extraction algorithm is as follows:

1. Segment the given audio track into 40-ms audio frames with a 10-ms overlap between two successive frames.
2. Compute the spectrogram (Figure 4.2(a)) of the audio frames using short-time Fourier transform.
3. Convert the spectrogram to a corresponding gray-level image by taking the absolute values of the Fourier coefficients.
4. Construct a line map (Figure 4.2(b)) from the image using the Sobel operation [GW01], and a 7-order median filter is applied to remove outliers along each row of the map.
5. Identify all horizontal lines in the line map using the Hough transform [GW01].
6. For each 1-second frame, calculate the line number from every 4-pixel-wide windows with 2-pixel advance in the line map, and take the average of the line numbers as the final MCR value.

As a result, we use OEC, SIF, and MCR to practically realize a multi-class SVM classifier for speech/music discrimination [FCL05]. The classifier has been evaluated on three small audio datasets, each containing approximately three-hour sources. The first dataset is collected from Internet radio and the second is obtained by adding 5 dB white noises to the first one. In addition, we constitute the third one from audio tracks of two kinds of home videos, i.e. the wedding and the birthday party. Here, sound of birthday party is included because its audio contents have higher variations and contain more diversified sound effects. For example, some of the birthday parties are taken place at a quiet home, and others are in a very noisy environment, such as the restaurants with crowd laughing, talking, and cheering. Then, a fivefold cross-validation experiment [Bis06] is conducted for the classifier on each of the datasets and the results measured by average precisions and recalls are illustrated in Figure 4.3. The classification performance shows that the proposed audio features discriminate music/speech quite well even for the audio with a substantial amount of noises.

#### **Event Features Related to Applause Detection**

The same feature selection mechanisms, as described in Section 4.4.2, are applied to identify the noise-resistant audio features for detecting the presence of applause in low-SNR audio recordings. However, based on our experiments, the collected audio features in our candidate set (cf. Section 4.4.2) generally do not perform very well. Instead, a specific audio feature is developed by exploiting the physical properties of applause, as indicated in Section 4.4.1. That is, when applause is coming up in the audio signal, a significant increase in magnitude can be observed over the whole power spectrum [Rep87, PECV07]. An example is illustrated in Figure 4.4(a). For comparison, two power spectrums taken from



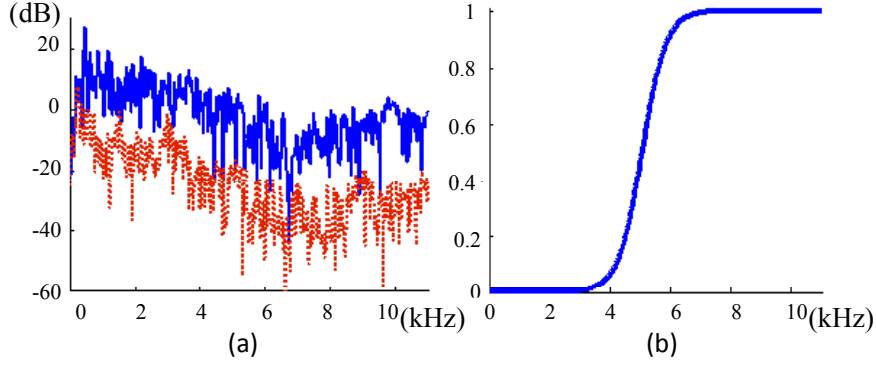
**Figure 4.3:** Classification results of the audio types of speech (the left subplot) and music (the right subplot) on three audio datasets of (a) Internet radio, (b) Internet radio with added white noises (5 dB), and (c) audio tracks from home videos, using a multi-class SVM classifier built upon the three audio features proposed in Section 4.4.2.

consecutive time instances of a wedding audio are depicted in the same figure. The spectrum with applause (the top solid curve) is around 20 dB larger in magnitude than the one without applause (the bottom dotted curve) for almost all frequencies. To capture the global variations of audio magnitudes, an audio feature of the weighted short-time energy (WSE) is employed.

- **Weighted Short-time Energy (WSE).** The feature value of weighted short-time energy is defined as the weighted sum over the spectrum power (in decibels) of an audio signal at a given time as follows:

$$\text{WSE} \triangleq \frac{1}{\text{WSE}_{\max}} \int_0^{\omega_s} W(\omega) \cdot 10 \log(|SF(\omega)|^2 + 1) d\omega \quad (4.6)$$

where  $SF(\omega)$  is the short-time Fourier transform coefficient of the frequency component  $\omega$ , and  $W(\omega)$  is the corresponding weighting function. In addition,  $\omega_s$  denotes the sampling frequency and  $\text{WSE}_{\max}$  is the maximum WSE

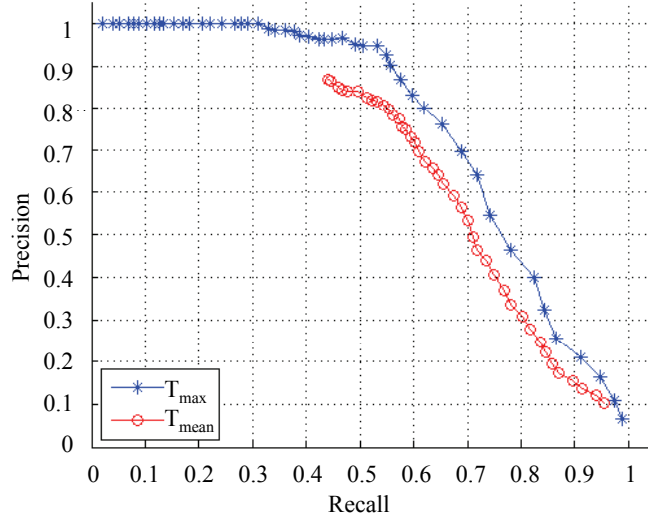


**Figure 4.4:** Examples of (a) two power spectrums of a wedding audio from consecutive time instances, one with applause (the top solid curve) and another without applause (the bottom dotted curve), and (b) a sigmoidal filter function.

in the audio track as a normalization factor. The calculation of WSE is special in that the spectrum power is in a logarithmic unit of decibels. Summation in the decibel domain is the same as multiplication in the energy domain. The logarithmic nature makes that a large WSE value is coming from a global trend of high power over the whole spectrum but not few dominant frequencies. Furthermore, since human speech is commonly observed in a wedding and the speech signals are bandlimited to around 3.2 kHz [ZK01],  $W(\omega)$  is chosen to be a sigmoidal function (cf. Figure 4.4(b)) in order to suppress the contributions from low frequencies. Specifically,

$$W(\omega) = \frac{1}{1 + e^{-\omega_1(\omega - \omega_2)}}, \quad (4.7)$$

where  $\omega_1$  and  $\omega_2$  are control parameters and are respectively set to 2.5 (kHz) and 5.0 (kHz). Therefore, as mentioned in Section 4.4.2, the input audio track is first segmented into non-overlapping 1-second audio frames. For each audio frame, one feature value is computed for every 50-ms interval with 10-ms overlap. A median filter is then applied to diminish possible noises. Instead of aggregation, based on our experiments, the maximum of









**Figure 4.5:** Precision-recall curves of the applause detection results using two different thresholds. (See Section 4.4.2 for details.)

these 25 feature values is selected as the representative WSE feature for that 1-second frame.

To verify the capability of WSE, a simple trial is conducted to detect the applause presented in audio recordings using two different thresholds:  $T_{\max}$  and  $T_{\text{mean}}$ . That is, given a series of WSE values, we compute two thresholds by individually multiplying the maximum value and their mean to a numerical factor between  $[0,1]$ . Then applause can be located at the positions with higher WSE values than the chosen threshold. Figure 4.5 illustrates the precision-recall curves of the average detection results on 15 audio tracks from a set of collected home videos, including wedding and birthday parties. Overall, the performance is well acceptable and it shows that WSE can capture applause effectively even for noisy home video recordings.

**Table 4.4:** The collection of six wedding videos used in our experiments.

Clip	A	B	C	D	E	F
						
<b>Duration</b>	2215 (sec)	410 (sec)	4122 (sec)	3790 (sec)	1062 (sec)	1350 (sec)
<b>Event #</b>	17	8	35	23	15	14

### Event Features Related to Flash Detection

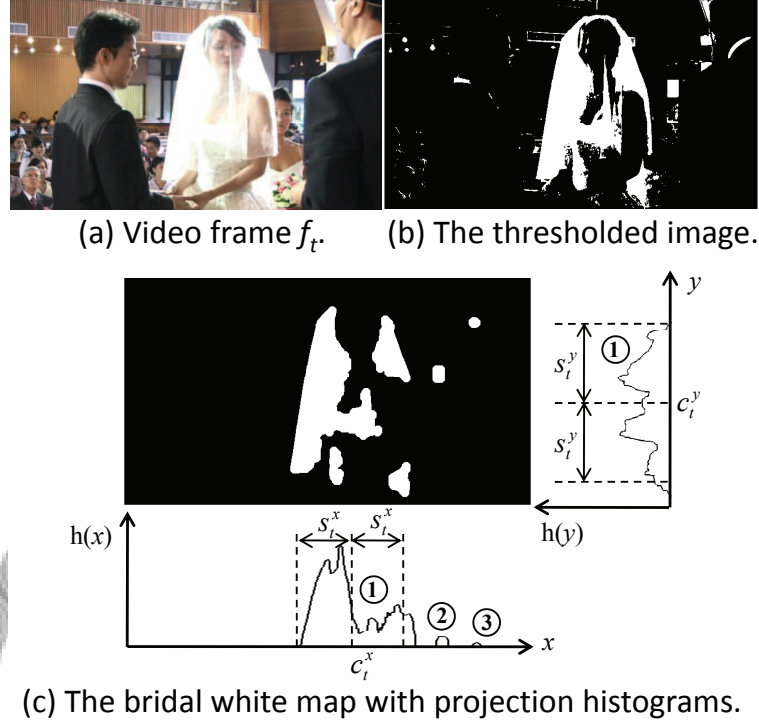
Flashes of picture-taking can be detected from abrupt and short increases of the global intensity in a video frame. A visual feature of the flash density, as suggested in Section 4.4.1, can then be defined in the following.

- **Flash Density (FLD).** In home videos, the durations of observed flashes are seldom longer than two video frames. In every 1-second interval, we compute a feature value of the flash density as follows:

$$\text{FLD} \triangleq \sum_{t=2}^{M-1} I((\hat{f}_t^I - \hat{f}_{t-1}^I \geq \epsilon) \text{ and } (\hat{f}_t^I - \hat{f}_{t+1}^I \geq \epsilon)) \quad (4.8)$$

where  $M$ ,  $\hat{f}_t^I$  are respectively the total number of video frames and the value of average intensity of the frame  $f_t$ , and the threshold  $\epsilon = 5$  was suggested by previous work [TV01] for flash detection.

To get more insight into the feature of FLD, we apply the flash detection algorithm to one wedding video used in later experiments, i.e. the Clip-A in Table 4.4. In terms of flash numbers, 457 flashes are correctly detected among the 482 truth ones, and there are 17 false positives. The detecting precision and recall are 94.81% and 96.41%, respectively. The detecting performance shows that flashes can be robustly captured.



**Figure 4.6:** Examples of (a) a video frame with (b) the thresholded image and (c) the bridal white map with projection histograms.

#### Event Features Related to Bride Indication

As mentioned in Section 4.4.1, the bride is an important leading role in wedding events and her appearance can be detected by the color of “bridal white”. However, due to various lighting conditions, the determination of real bridal white is extremely difficult and often needs a laborious training process similar to that of the skin color detection [PBC05]. Instead, our current implementation approximates bridal white map for each video frame, whereby a corresponding visual feature, bridal white ratio (BWR), can then be defined. The bridal white map is generated using the following procedure:

1. Convert a video frame  $f_t$  to the HSI color space [GW01], in which the values are within the range of  $[0, 255]$ .
2. Set empirically two thresholds  $\phi_t^I$  and  $\phi_t^S$  for the intensity and the saturation respectively for the bridal white:

$$\phi_t^I = \min(240, \hat{f}_t^I + 80) \text{ and } \phi_t^S = 75. \quad (4.9)$$

3. Construct a thresholded image  $\bar{\Gamma}_t$  from the video frame using the above two thresholds, cf. Figure 4.6(b). The thresholded image is defined as

$$\bar{\Gamma}_t(\mathbf{p}) = \begin{cases} 1, & \text{if } f_t^I(\mathbf{p}) \geq \phi_t^I \text{ and } f_t^S(\mathbf{p}) < \phi_t^S \\ 0, & \text{otherwise} \end{cases} \quad (4.10)$$

where  $\mathbf{p}$  is a pixel, and  $f_t^I(\mathbf{p})$  and  $f_t^S(\mathbf{p})$  denote  $\mathbf{p}$ 's intensity and saturation values, respectively.

4. Obtain a bridal white map  $\Gamma_t$  (cf. Figure 4.6(c)) by removing outliers of  $\bar{\Gamma}_t$  using a morphological closing (i.e., erosion followed by dilation) [GW01]. That is

$$\Gamma_t = \bar{\Gamma}_t \circ Se \quad (4.11)$$

where  $Se$  is a disk structuring element whose radius is 5-pixel wide and  $\circ$  denotes the closing operation.

After constructing bridal white map, the feature, bridal white ratio, is then defined as follows:

- **Bridal White Ratio (BWR).** To obtain BWR, the technique of histogram projection [HQS00] is applied to improve the reliability of  $\Gamma_t$ . Specifically, based on the observation that the bride roughly appears in the shape of a white vertical bar (cf. Figure 4.6(a)), we add a spatial constraint that



the white distribution in the vertical direction should be wider than that in the horizontal one. Therefore, we project the bridal white map along the  $x$  and the  $y$  directions to construct two 1-D histograms (cf. Figure 4.6(c)), from which the isolated component with the maximum white ratio is individually selected. For example, in Figure 4.6(c), there are three isolated components in the horizontal histogram but only one in the vertical one. We compute standard deviations,  $s_t^x$  and  $s_t^y$ , of the white distributions for the maximum components along both axes. In every 1-second interval, a feature value of BWR is defined as

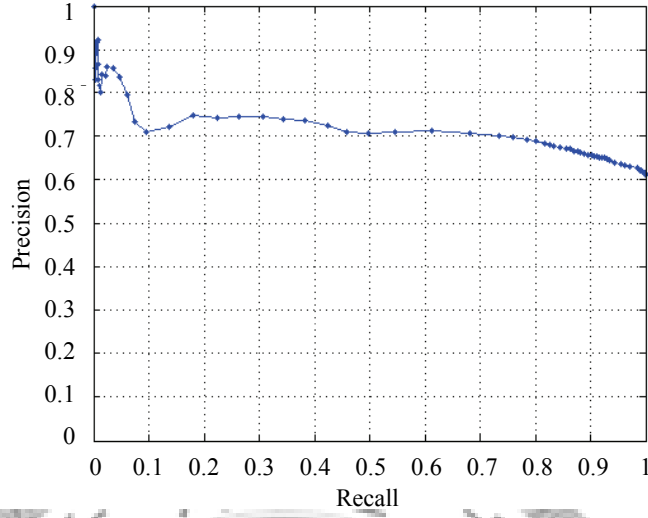
$$\text{BWR} \triangleq \frac{1}{M} \sum_{t=1}^M \Phi(\Gamma_t) \cdot I(s_t^x < s_t^y) \quad (4.12)$$

where  $\Phi(\Gamma_t)$  returns the white ratio of  $\Gamma_t$  in terms of white pixel number with respect to the map size. Note that we use the average white percentage to avoid making the hard-decision on whether the bride is present in video frames or not.

For understanding its performance, a simple trial is carried out for the bride indication by making binary decisions (i.e. presence or absence) on the basis of the obtained BWR values. Given a predefined threshold, a higher BWR value corresponds to the bride's presence, otherwise her absence. Figure 4.7 illustrates precision-recall curves of the detecting results for a wedding video, i.e. the Clip-A in Table 4.4. The “hard-decision” performance is promising and we believe that the resulted “soft-decision” BWR is helpful for our modeling task.

## 4.5 Wedding Modeling

The objective of wedding modeling is to estimate the event sequencing of a wedding video. At each time instance, extracted event features are exploited to



**Figure 4.7:** Precision-recall curves of the bride indication results. (See Section 4.4.2 for details.)

recognize the wedding events. In addition, a wedding video is a kind of sequential data. The occurrence of a wedding event highly depends on the category of its preceding neighbors. Thus, in wedding modeling, it needs not only to consider how likely the acquired features match an event candidate but also the temporal rationality whether the candidate is appropriate to follow the existing sequence immediately. Therefore, we use an effective learning tool, the hidden Markov model (HMM), to describe the spatio-temporal relations of events within a wedding video [Bis06]. In Sections 4.5.1 and 4.5.2, we first build statistical models of the feature similarity and the temporal ordering for each of the wedding events. Section 4.5.3 then devises an integrated HMM framework for both the event-based analysis and the wedding segmentation.

Before proceeding, note that we uniformly divide the wedding video into a sequence of 1-second units. The main reason for this uniform pre-segmentation is that we can not use conventional video units, such as shots, as the basic analysis units. This is because shots of a wedding video can't be reliably obtained using

conventional techniques as mentioned in Section 4.1. In addition, the simplicity of uniform segmentation makes online processing possible. For convenience, let  $E$  denotes an index set [GKP94] of the wedding events, where the indexing consists of a *bijective* mapping from the event set  $E_S = \{ME, GE, \dots, OT\}$  to a set of natural numbers, i.e.  $E = \{1, 2, \dots, |E_S|\}$ . Similarly,  $F$  is an index set corresponding to the collection of event features  $F_S = \{OEC, SIF, MCR, WSE, FLD, BWR\}$ . For the  $t$ -th video unit, let  $\mathbf{e}_t \in E$  be the corresponding state variable that indicates the occurrence of a specific wedding event, and let  $\mathbf{x}_t = (x_t^1, \dots, x_t^{|F|})$  be the feature vector associated with the specific event features  $x_t^j$ ,  $j \in F$ .

#### 4.5.1 Wedding Event Modeling

For each of the wedding events, a statistical feature model is constructed for each of the adopted event features. Specifically, a feature model is a probability distribution describing the likelihood of feature values. The use of statistical histograms [GW01] is a naïve approach, but their discrete nature often causes unwanted discontinuity in results, especially when a feature value locates near the boundaries of histogram bins. Instead, we accumulate the probability by regarding each feature sample as a Gaussian centered at the sample. Assume that, for the  $i$ -th event, we have  $N$  samples for the  $j$ -th feature  $\{x_1^j, \dots, x_N^j\}$  extracted from the training clips. The distribution  $p_{i,j}$  of the  $j$ -th feature for the  $i$ -th event can then be obtained as

$$p_{i,j}(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^N \frac{1}{\lambda_j \sqrt{2\pi}} e^{-(\mathbf{x} - x_n^j)^2 / 2(\lambda_j)^2}, \quad \forall i \in E, \quad \forall j \in F, \quad (4.13)$$

where  $\int_{\mathbf{x}=-\infty}^{\infty} p_{i,j}(\mathbf{x}) d\mathbf{x} = 1$  and  $\lambda_j$  is a confidence parameter specifying how we trust the extracted values of the  $j$ -th feature. That is, if the extracted feature samples are more accurate and reliable, we can set  $\lambda_j$  to a smaller value.

Since the feature models are used for discriminating the wedding events, the divergence among feature models of different wedding events should be as large as possible. Quantitatively, the divergence of two probability distributions can be defined by the symmetric Kullback-Leibler (SKL) distance [CT06]:

$$D_{SKL}(\mathbf{p}, \mathbf{q}) = \frac{1}{2} \int_y \left[ \mathbf{p}(y) \log \frac{\mathbf{p}(y)}{\mathbf{q}(y)} + \mathbf{q}(y) \log \frac{\mathbf{q}(y)}{\mathbf{p}(y)} \right] dy \quad (4.14)$$

For the  $j$ -th feature, the confidence parameter  $\lambda_j$  is chosen to maximize the sum of divergences among the same kind of feature models. That is,

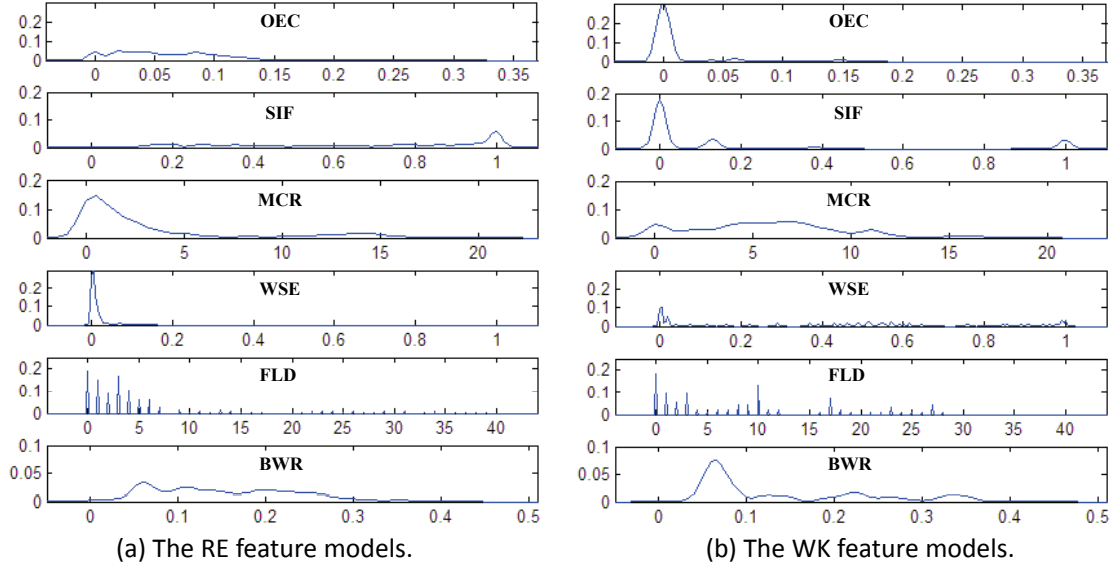
$$\lambda_j = \arg \max_{\lambda} \sum_{i,k \in E, i < k} D_{SKL}(p_{i,j}, p_{k,j}) \quad (4.15)$$

To find the optimal  $\lambda_j$ , we use exhausted search and empirically set a search range (e.g.  $[0, 1]$ ) with a desired precision (e.g. 0.05). The optimal confidence parameters we found are  $\lambda_{OEC} = 0.005$ ,  $\lambda_{SIF} = 0.015$ ,  $\lambda_{MCR} = 0.5$ ,  $\lambda_{WSE} = 0.0025$ , and  $\lambda_{BWR} = 0.01$ . It is worthy to notice that FLD is an exception because its values are discrete. As a result, we manually set  $\lambda_{FLD} = 0$  and apply a 9-point normalized filter to the sample sequences of FLD feature values as an alternative to the Gaussian-based smoothing.

Therefore, given a video unit (e.g. the  $t$ -th one), we can compute the probability that we observe  $\mathbf{x}_t$  given that this video unit belongs to the  $i$ -th wedding event:

$$p(\mathbf{x}_t | \mathbf{e}_t = i) = \prod_{j=1}^{|F|} p_{i,j}(x_t^j) \quad (4.16)$$

Note that, in practice, we compute the log-likelihood by taking logarithm of the expression, and thus obtain a contributive weight  $\kappa_j$  to the  $j$ -th feature model, where  $\sum_j \kappa_j = 1$ . In our experiments, we give a default set of the weights, i.e.  $\kappa_{OEC} = 0.25$ ,  $\kappa_{SIF} = 0.2$ ,  $\kappa_{MCR} = 0.1$ ,  $\kappa_{WSE} = 0.1$ ,  $\kappa_{FLD} = 0.1$ , and  $\kappa_{BWR} = 0.25$ . They are automatically specified by optimizing the recognition



**Figure 4.8:** Examples of wedding event models of (a) the *RE* event and (b) the *WK* event.

accuracy of wedding events through a cross-validation process (cf. Section 4.6) that is iteratively repeated among training clips. An interesting phenomenon is that the audio-based event features take as high as two-thirds of the weights. This implies that audio information seems more crucial for the wedding analysis.

Overall, the proposed event modeling has the following advantages. First, it has good tolerance to inaccuracy and uncertainty of the extracted event features. The Gaussian component helps to reduce and diversify the influence of an inaccurate feature value. Second, it avoids the artifacts due to quantization errors in the constructed feature models. The distribution of feature values is faithfully represented without approximation. Figure 4.8 gives examples of feature statistical models for two wedding events, *RE* and *WK*.

### 4.5.2 Event Transition Modeling

The event transition model (ETM) is constructed to describe the probability that a wedding event is immediately followed by another in a wedding ceremony. In other words, it evaluates whether a temporal transition is to be allowed between each pair of the wedding events. Therefore, ETM can be defined by an  $|E| \times |E|$  matrix  $A$  as follows:

$$A_{i,k} = Pr(\mathbf{e}_t = k | \mathbf{e}_{t-1} = i), \forall i, k \in E \quad (4.17)$$

where  $A_{i,k}$  is the entry of the  $i$ -th row and the  $k$ -th column of  $A$ , and  $t-1, t$  are two successive time instances in seconds. Since all possible transitions are enumerated in  $A$ , the marginal probability along each row is unity, that is

$$\sum_{k=1}^{|E|} A_{i,k} = 1, \forall i \in E. \quad (4.18)$$

In fact, given a training set of wedding videos with the event ground truth, we can tabulate an approximation of ETM, namely  $\tilde{A}$ . However, the obtained probability distributions are often extremely biased. That is, most of the probabilities are prone to centralize on the diagonal entries, i.e.  $\tilde{A}_{i,i}$ . This phenomenon is due to the fact that transitions are counted in seconds. For example, assuming that we have two successive events which are both 100 seconds long, only one event transition will be accounted during this 200-second period. Therefore, for each row of  $\tilde{A}$  (e.g. the  $i$ -th one), we exploit a regularization to balance the probabilities as follows:

$$A_{i,k} = \begin{cases} \gamma_i \tilde{A}_{i,k} & , i = k \\ (1 - \gamma_i \tilde{A}_{i,i}) / (1 - \tilde{A}_{i,i}) \cdot \tilde{A}_{i,k} & , i \neq k \end{cases}, \forall k \in E \quad (4.19)$$

where  $\gamma_i$  is the regularization factor in the range of  $[0, 1]$ . To be precise, we shift some of the diagonal probabilities to the off-diagonal ones but keep their relative

**Table 4.5:** An even transition model of the wedding events.

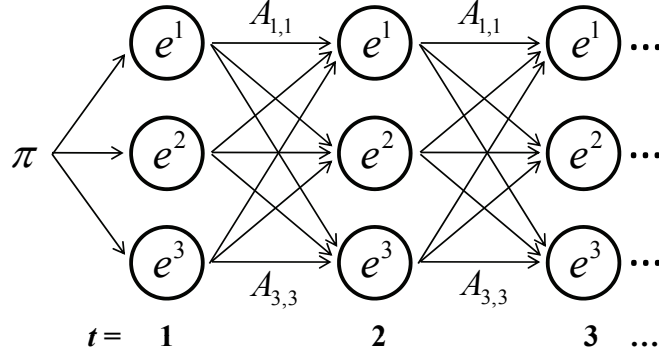
	<i>ME</i>	<i>GE</i>	<i>BE</i>	<i>CS</i>	<i>OP</i>	<i>WV</i>	<i>RE</i>	<i>BU</i>	<i>MS</i>	<i>WK</i>	<i>AP</i>	<i>ED</i>	<i>OT</i>
<i>ME</i>	0.80	0.11	0.09										
<i>GE</i>	0.12	0.80	0.08										
<i>BE</i>			0.80	0.04	0.16								
<i>CS</i>				0.80	0.16							0.01	0.03
<i>OP</i>				0.07	0.80	0.03	0.01			0.01	0.02	0.02	0.04
<i>WV</i>						0.80	0.13	0.03					0.03
<i>RE</i>						0.03	0.80	0.13					0.03
<i>BU</i>								0.80		0.20			
<i>MS</i>					0.07				0.80		0.07	0.07	
<i>WK</i>				0.03	0.11				0.03	0.80			0.03
<i>AP</i>				0.12					0.04		0.80	0.04	
<i>ED</i>												1.00	
<i>OT</i>				0.05	0.14				0.02				0.80

ratios unchanged. Empirically, all of the diagonal entries are regularized to take approximate 80% probabilities along each row, i.e.  $A_{i,i} \approx 0.8$ , after regularization.

Table 4.5 shows the ETM we learnt from training videos, in which the blank entries represent zero probabilities. Sparsity of the ETM shows that few types of event transitions are allowed. It also demonstrates the occurrence of wedding events has a strong temporal correlation. This fact helps to reduce the computation cost and to increase the reliability of the determined event sequencing.

### 4.5.3 Wedding Segmentation Using HMM

HMM is a specific instance of state space models, in which the concept of hidden states is introduced to recognize the temporal pattern of a Markov process [Bis06]. Since the sequence of wedding events can be viewed as a first-order Markov data, as shown in Section 4.5.2, we exploit an HMM framework for segmenting wedding videos, in which the wedding event statistical models (Section 4.5.1) and the event transition model (Section 4.5.2) are integrated together.



**Figure 4.9:** A simplified example of the HMM for wedding segmentation. (See Subsection 4.5.3 for details.)

Specifically, given an input wedding video  $V$ , it is first partitioned into  $N$  1-second video units,  $V = \{\mathbf{v}_1, \dots, \mathbf{v}_N\}$ . For each video unit  $\mathbf{v}_t$ ,  $t \in \{1, \dots, N\}$ , we have a set of  $|F|$  event features associated with it, i.e.  $\mathbf{x}_t = (x_t^1, \dots, x_t^{|F|})$ . Collecting all the observations  $X = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ , our goal is to find the most probable event sequencing  $S$  for  $V$ , where  $S = \{\mathbf{e}_1, \dots, \mathbf{e}_N\}$ . Therefore, we develop a left-to-right HMM of  $|E|$  states  $\{e^i | i \in E\}$ , in which each state corresponds to one of the adopted event categories. The HMM is governed by a set of parameters,  $\theta = \{\pi, A, \phi\}$ , where  $\pi$ ,  $A$ , and  $\phi$  define the initial state probabilities, the state transition probabilities, and the emission probabilities, respectively [Bis06]. Figure 4.9 illustrates a trellis representation of a simplified HMM with only three states. Clearly,  $\phi$  and  $A$  have been explicitly described by the wedding event models and the event transition model, respectively. Without loss of generality,  $\pi$  is presumed to be a uniform distribution, i.e.  $p(\mathbf{e}_1 = i | \pi) = 1/|E|, \forall i \in E$ .



Accordingly, our goal for finding the optimal sequencing  $S$  can be formulated as

$$\begin{aligned}
 S &= \arg \max_s Pr(X, S|\theta) \\
 &= \arg \max_s p(\mathbf{e}_1|\pi) \left[ \prod_{t=2}^N p(\mathbf{e}_t|\mathbf{e}_{t-1}, A) \right] \prod_{t=2}^N p(\mathbf{x}_t|\mathbf{e}_t, \phi) \\
 &= \arg \max_s p(\mathbf{e}_1|\pi) \left[ \prod_{t=2}^N A_{\mathbf{e}_{t-1}, \mathbf{e}_t} \right] \prod_{t=2}^N \prod_{j=1}^{|F|} p_{\mathbf{e}_t, j}(x_t^j)
 \end{aligned} \tag{4.20}$$

where the second and the third terms are derived from Equations (4.16) and (4.17), respectively. Because the HMM trellis is equivalent to a directed tree (as shown in Figure 4.9), the solution of  $S$  can be efficiently obtained using the Viterbi algorithm [Bis06].

After labeling each 1-second unit of the input video, the temporal extent of a detected wedding event, or called an event segment, is defined by collecting successive video units with the same event labeling. Finally, a smoothing scheme is applied to reduce possible labeling errors. Since, in general, a wedding event lasts for at least tens of seconds, we remove the short ones (less than 10 seconds in duration) by merging it into its neighbors. If its proceeding and succeeding neighbors belong to different event categories, it is merged into the left one; otherwise, all the three events are merged into one event.

## 4.6 Experimental Results

This section first presents experimental results for the evaluation of the proposed framework in wedding event recognition (Section 4.6.1) and wedding ceremony video segmentation (Section 4.6.2). The performance comparisons with another well-known algorithm, linear-chain conditional random fields (LCRF),

and an extension of our system to a practical scenario are respectively described in Sections 4.6.3 and 4.6.4.

In our experiments, we used a total of six wedding video clips. Each of them contains a complete recording of a wedding ceremony. Three observers (none of the clip owners) collaboratively annotated the event ground truth. Table 4.4 summarizes the statistics of the videos used in the experiments and also reports durations and numbers of the annotated events for all six videos. Our experiments were performed using a leave-one-out cross-validation strategy, in which models were trained from five clips and tested on the remaining one, and the whole training-testing procedure was iterated six times. In addition, our current system is programmed using Matlab 7.2 without code optimization, and running on a machine with Intel P4 3.0 GHz CPU, 1.0 GB memory, and MS Windows XP Professional x32 Edition. Based on the experiments below in Section 4.6.1, the average testing time for a clip is about 15 times longer than its original video length, and the extraction of audiovisual features accounts for around 96% of the time.

#### 4.6.1 Event Recognition Analysis

Table 4.7 summarizes the event recognition results in unit of seconds, presented in the form of confusion matrix [LD06], where the leftmost column represents the actual event categories while the top-most row indicates the resultant ones recognized by the HMM framework. The confusion matrix is accumulated from results of all clips in the collection. The recognition precision (RP) and the recognition recall (RR) for each of the event categories are reported in Table 4.7. As described in Section 4.1, since the actual boundaries between wedding events are not always precise, the recognition result of a video unit is claimed to be correct if it hits the

**Table 4.6:** The statistics of means  $\mu$  and variances  $\sigma^2$  of event duration for each of the event categories in our video collection (unit: seconds).

(a) From all event samples							
Event	ME	GE	BE	CS	OP	WV	RE
$\mu_i$	92.00	42.33	114.00	139.90	130.91	163.33	135.50
$\sigma_i$	38.11	36.25	67.73	104.62	182.28	61.71	13.20
Event	BU	MS	WK	AP	ED	OT	
$\mu_i$	47.33	166.00	11.60	68.33	75.20	149.08	
$\sigma_i$	6.66	62.60	1.14	6.66	13.48	67.13	
(b) From half of the event samples with shorter durations							
Event	ME	GE	BE	CS	OP	WV	RE
$\tilde{\mu}_i$	45.33	19.00	37.00	56.64	54.24	88.50	111.67
$\tilde{\sigma}_i$	15.95	5.57	1.41	32.08	32.16	26.16	23.63
Event	BU	MS	WK	AP	ED	OT	
$\tilde{\mu}_i$	38.67	132.50	10.00	61.33	51.33	97.63	
$\tilde{\sigma}_i$	8.39	33.23	1.00	5.51	24.01	40.17	

ground truth within a tolerant range. Instead of setting a universal range value, we adopt a dynamic setting scheme based on the recognized event categories because the event durations vary greatly among different wedding events as shown in Table 4.6(a). Initially, for each event category, all of the event samples are sorted by duration in descending order. We then compute a truncated mean  $\tilde{\mu}_i$  of the event duration (Table 4.6(b)) by ignoring the samples of the first half (i.e. the longer ones in the top half), and the range value is set to  $\min(0.2\tilde{\mu}_i, \xi)$ , where we set  $\xi = 10$  so that the tolerant ranges vary according to event categories but do not exceed 10 seconds. Here, we use a truncated mean but not the standard mean because of its better statistical reliability. That is, for most of the event categories, a large variance is observed with durations of all of the event samples, as shown in Table 4.6(a). By contrast, as shown in Table 4.6(b), the truncated variances are generally much smaller than the standard ones in Table 4.6(a), which implies that durations of the shorter samples would be more consistent. More importantly, by

ignoring the longer samples, a smaller tolerant range can be naturally obtained to enforce a stricter standard for recognition hits.

Overall, as shown in Table 4.7, large amounts of the detected wedding events reach over 70% in both RP and RR values. Some of them even achieve the level of 85%, such as *BU* and *ED* events. Several observations could be made from this table: 1) A few recognition errors are associated with *CS* and *OP* events, especially the later one. This phenomenon is usually unavoidable because a wedding event, such as *OP* or *MS*, is sometimes arranged to be accompanied with choirs singing and the whole ceremony is generally hosted by wedding officiants who often give some short presentations within a wedding event. They also cause severe degradations in RP values for both *RE* and *AP* events. 2) The confusion matrix is sparse and the recognition errors show grouping effects. That is, the wedding events of a similar group are prone to be mis-classified to each other, e.g. the set of the entering events (*ME*, *GE*, *BE*) and the set of the couple's committing events (*WV*, *RE*). From Table 4.5, we can find that the events of each event set correspond to the ones that are more probable to occur in succession. Thus, the recognition errors partially come from the implicit event boundaries. 3) The RR value of *OT* event is relatively low. This is due to the fact that *OT* event is inherently varied in forms. For example, it could be 'reading of poetry' or 'lighting of the unity candle'. Compared with other kinds of wedding events, *OT* event is the most difficult one to be modeled. Moreover, it severely influences the overall recognition performance by spreading out the recognition errors over various event categories.

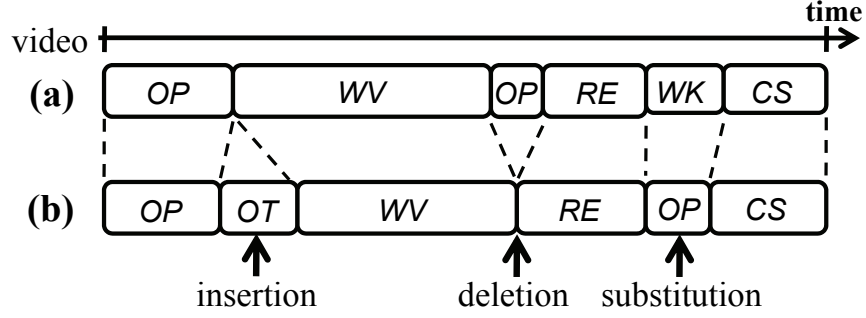
As a comparison with the HMM-based modeling, we also perform event recognition solely based on the maximum similarity of audiovisual features among the wedding events without exploiting the temporal relation of event transitions. Ta-

ble 4.8 shows the results when using (a) the four audio features only, (b) the two visual features only, and (c) all six audiovisual features. Generally, the use of both audio and visual features together outperforms the use of either unimodal features alone. The adopted event features from both modalities could complement each other in the recognition task. However, the results in Table 4.8(c) are still not as good as those in Table 4.7, in which event transition modeling is augmented. This shows evidences to support the effectiveness of the HMM framework.

### 4.6.2 Video Segmentation Analysis

In this section, we further evaluate the segmentation performance of our approach. Since, in practice, the temporal extent of a wedding event is perceived as a whole by users, the segmentation results are compared at the ‘event’ level but not at the ‘second’ level. We follow a similar idea exploited in the longest common substring problems [CLRS01]. That is, we represent a wedding video as a symbol string where the alphabet consists of the event codes given in Table 4.1. Note that the symbol string is generated in unit of detected events, and each symbol corresponds to an event segment of the wedding video. Therefore, for each of the tested wedding clips, the segmentation performance is measured by the number of the required edit operations (substitution, insertion, and deletion) for transforming the reference string corresponding to the ground truth into the string corresponding to the recognition result. Figure 4.10 shows an example of transforming strings. The less the edit operations are needed, the better the segmented videos match with the ground truth.

Table 4.9 shows the statistics. We claim an event segment as correct if it hits the ground truth in more than 80% of its duration. The segmentation precision (SP) and the segmentation recall (SR) of a resultant video are then defined as



**Figure 4.10:** Edit operations for transforming (a) a reference event string to (b) the one for comparison.

follows:

$$SP = \frac{\text{Corrects}}{\text{Corrects} + \text{Substitutions} + \text{Insertions}} \cdot 100\%, \quad (4.21)$$

$$SR = \frac{\text{Corrects}}{\text{Corrects} + \text{Substitutions} + \text{Deletions}} \cdot 100\%. \quad (4.22)$$

In addition, the F-measure,  $SF = 2 \cdot SP \cdot SR / (SP + SR)$ , is provided as a metric for evaluating the integral performance.

From Table 4.9, we can see that SR values generally achieve 80% high, i.e. most of the event segments are correctly identified. A low value of Clip-B comes mostly from its small number of events as shown in Table 4.4. By contrast, the overall SP values are relatively low, at the level of 60%. Compared with the ground truth, a large amount of redundant events are erroneously “inserted” in the segmentation results by our approach. These are mainly caused by the following two reasons. First, the erroneous events are generated in a one-to-many pattern. A single event that has been deleted from the ground truth usually turns into a series of successive erroneous ones in the resultant event sequence. This phenomenon is partly relating to the use of our HMM modeling. For example, consider an event subsequence of the ground truth,  $WK-CS-ED$ . If  $CS$  is not detected and the direct transition from  $WK$  to  $ED$  is not allowed (i.e. the transition probability equals zero), the

HMM framework would be forced to go through a longer path of erroneous events to connect *WK* and *ED*, such as *WK-OT-MS-ED*. Also, when a succession of two events has never been observed in the training data, its zero transitive probability could cause the same problem. Second, the erroneous events are prone to exist around an event boundary of the ground truth. The same phenomenon has been observed from the recognition errors, as reported in Section 4.6.1.

Since the erroneous events are “mutated” from parts of the original event segments, in general, they have a shorter duration as compared with the same kind of wedding events. Therefore, we use a duration-based filtering scheme to identify and correct the abnormal ones. Specifically, for each of the event categories, we exploit the truncated models (Section 4.6.1 and Table 4.6(b)) to determine a lower bound of the reasonable event duration, i.e.  $\Omega_i = \tilde{\mu}_i - \alpha_i \tilde{\sigma}_i$ , where a rational scalar  $\alpha_i$  is empirically set within the range of  $[1.5, 2]$ . If an event segment is recognized as the  $i$ -th event category and its duration is less than  $\Omega_i$ , we merge it into its left neighbor in our current implementation. Table 4.10 summarizes the segmentation results after applying the duration-based filtering. Compared with Table 4.9, the number of inserted erroneous events is effectively reduced and on average a 10% improvement is obtained for SP values. This improvement is accompanied by a slight decrease in SR values because some correct events would be filtered out at the same time.

Overall, as shown in Table 4.10, the performance of our system is satisfactory. It achieves the level of 70% in terms of the SF metrics. Furthermore, with the assist of the duration-based filter, the tendencies of both SP and SR behaviors are much more balanced and consistent. The statistical results may not be comprehensive but it is encouraging. It gives us support and confidence that, as long as we capture well the content characteristics, it is possible to conduct high-level

semantic analysis of home videos through the use of generic and easily extracted audiovisual features. That is also an advantage of the proposed framework, making it plausible for real applications.

### 4.6.3 Performance Comparisons with LCRF Models

To further evaluate the validity of HMM approach, we compare the performance of HMM with that of the linear-chain conditional random fields (LCRF) [LMP01, GT06]. LCRF is a well-known probabilistic framework for labeling and segmenting sequence data. In the terminology of statistical relational learning, HMM and LCRF are known as a *generative-discriminative pair* [GT06], in the sense that HMM measures the joint probability of sequential observations and the corresponding label sequences but LCRF is to estimate the conditional probability of associated label sequences given the observations.

As a comparison to HMM framework (cf. Section 4.5.3), in LCRF modeling the goal to find the optimal sequence  $S$  given observations  $X$  is differently formulated as

$$S = \arg \max_s Pr(S|X, \theta') \quad (4.23)$$

where  $\theta'$  denotes the model parameters. A quasi-Newton method (i.e. BFGS [GT06]) is then adopted to optimize the estimation of  $\theta'$  from the training data. Following the same experimental procedures as described in Sections 4.6.1 and 4.6.2, both the event recognition and the video segmentation results of the LCRF model are obtained and summarized in Tables 4.12, 4.13, and 4.14.

In Table 4.12, some observations can be made from the LCRF recognition statistics: 1) As compared with the results of HMM in Table 4.7, LCRF performs much worse in RR values than RP values. A half of the RR values is below the 50% level and some are even down to the level of 20%, such as  $WK$ ,  $AP$ ,



and *OT* events. Also, the events with low RR values often have relatively lower RP values. This phenomenon might partly come from two reasons. One is the inherent event properties and the other is the unbalanced amount of training samples. For example, as shown in Table 4.11, the summed percentage of total event duration for these low-RR events is less than that of a single *OP* event in our video collection. By contrast, HMM's performance (in both the RP and RR values) are more consistent and stable, as discussed in Section 4.6.1. It seems that HMM approach could be more robust for unbalanced classification. 2) The low RR values are inappropriate for real applications. For example, the events of *WV*, *RE*, *BU*, and *WK* are arguably the most important moments in a wedding ceremony and also the most frequent pieces users would like to review in wedding videos [Spa01]. However, a large number of those events are not detected by the LCRF model, cf. Table 4.12. It is especially worthy to note that *WK* event has both its RP and RR values at the level of merely 20%. By contrast, HMM approach performs better in the RR values, e.g. both *BU* and *WK* events are higher than 95%, although the corresponding RP values are comparably lower. From the user's perspective, they would more like to see "fakes" rather than totally miss anything important. 3) Similar to HMM results, *OT* event is still a main culprit for bad recognition performance, and the performance is even worse for LCRF model. Specifically, not only the *OT* event tends to be incorrectly detected as the other event categories, but also events from the other categories are prone to be recognized as *OT*. In Table 4.12, the effects can be observed from the widespread errors associated with the *OT* event.

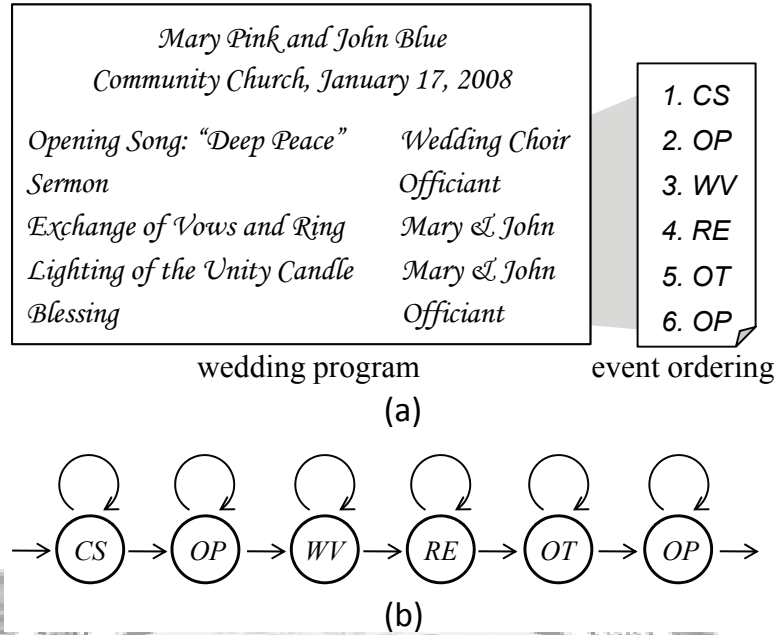
Tables 4.13 and 4.14 give the video segmentation results of LCRF model, with and without duration-based filtering. From Table 4.13, we can see that the most SR values are only around 60% and 70% levels. In comparison with HMM results

(cf. Table 4.9), the degradation is due to that more ground-truth events failed to be detected. On the other hand, an interesting thing is the burst increase in number of deletions after duration-based filtering, as shown in Table 4.14. For example, the deletions for clip C rise to near three times as the original. Based on our observations, it is also caused by the low RRs as described above. This fact makes the detected duration of events tend to be shorter than their actual lengths in the ground truth. The “abnormality” then raises their possibilities to be removed during the filtering. In brief, comparing HMM approach to the LCRF model, the HMM framework is more effective in the recognition of wedding events, especially the highlights such as *WK*. More importantly, in terms of duration, HMM approach would be more accurate to include complete contents in the detected events.

#### 4.6.4 Extension to the Scenario with Known Event Ordering

In this section, we investigate an extension of our work to the scenario when the actual event ordering of a wedding video is available. The investigation is conducted for two purposes. First, by reducing the temporal uncertainty, it is more reliable for us to examine the true capability of the proposed audiovisual features in discriminating various wedding events. Second, the scenario creates an opportunity for users to interact with our system so as to possibly improve the segmentation accuracy. For example, the ordering information can be obtained by manual input from users or semi-automatic transcription from the couple’s wedding programs [War06], such as the one shown in Figure 4.11(a).

Under the assumption of known event ordering, our original task is in some sense converted into the type of *change-point problem* [ZS05, CLRS01]. That



**Figure 4.11:** (a) A sample wedding program accompanied with the transcribed event ordering, and (b) the state diagram in form of a Markov chain built according to the above event ordering.

is, the problem is to determine the set of boundaries where event transitions happen. Therefore, instead of using the proposed HMM framework, a modified state space model is built for each wedding video, in which each state corresponds to one of the known events and the states are arranged in the form of a Markov chain according to the given event ordering, as illustrated in Figures 4.11(a) and (b). Note that the directed edges are simply used to indicate allowable transitions between states but not assigned with any transition weights in order to account for the contributions from the wedding event models alone. The most probable event sequence is then computed by exploiting dynamic programming [Bis06, CLRS01], and the event boundaries can be automatically located at the points of transition among different states.

Table 4.15 summarizes the segmentation results, in which “Detects” are defined as the number of detected event segments and “Corrects” (cf. Section 4.6.2) indicate the number of correct ones among “Detects”. The statistics of precision (P), recall (R), and F-measure (F) are also reported in this table. As a reference, Table 4.16 shows the second-based recognition ratio for each of the testing clips. The overall performance is satisfactory. Both the precisions and recalls reach a high level of more than 80%. The results are very encouraging. It would not only demonstrate the effectiveness of our audiovisual features but also imply that the minor requirement of user intervention could further advance the proposed framework for practical applications.

## 4.7 Summary

In this chapter, we have proposed and realized a system for event-based wedding video analysis and segmentation. According to the wedding customs, we developed a taxonomy for classifying wedding events, whereby a set of discriminative audiovisual event features are exploited for robust event modeling. Combined with a hidden Markov model, the resulted system shows good performance on event recognition and video segmentation for wedding videos. Thus, it can help users to access, organize, and retrieve his/her treasured contents in an automatic and more efficient way. To the best of our knowledge, this work is the first one to analyze and structure wedding videos on the basis of semantic events. Actually, it might also be the first one for semantic event analysis on the general domain of home videos.

Many aspects of our approach can be improved. First, it is possible to explore more semantic features for event recognition. For example, speaker change detection or identification would be helpful in discriminating the events with dense

speech, such as *WV* and *RE* events. Next, the modeling mechanisms could be improved. For example, on one hand, advanced fusion schemes of the feature models can be adopted. One example is hierarchical classification that combines homogeneous features as mid-level concepts and then builds event models on top of these concepts [FGLJ08, SWS05]. On the other hand, the development of a time-variant event transition model could produce more reasonable event sequences. Finally, more extensive and thorough evaluation of our system is a must. Moreover, since home videos are private data and usually hard to be acquired, it is beneficial to have a common database and relevant evaluation benchmarks for wedding videos. In the future, we will continue our investigation in these directions.



**Table 4.7:** The recognition results of all wedding events (unit: seconds).

Events	<i>ME</i>	<i>GE</i>	<i>BE</i>	<i>CS</i>	<i>OP</i>	<i>WV</i>	<i>RE</i>	<i>BU</i>	<i>MS</i>	<i>WK</i>	<i>AP</i>	<i>ED</i>	<i>OT</i>	RR(%)
<i>ME</i>	<b>547</b>	0	<b>32</b>	0	0	0	0	0	0	0	0	0	0	<b>94.47</b>
<i>GE</i>	25	<b>99</b>	18	0	0	0	0	0	0	0	0	0	0	<b>69.72</b>
<i>BE</i>	80	0	<b>350</b>	0	0	0	0	0	0	0	0	0	0	<b>81.40</b>
<i>CS</i>	0	0	0	<b>2320</b>	93	0	0	0	0	42	64	0	154	<b>86.79</b>
<i>OP</i>	1	0	5	212	<b>3622</b>	145	459	4	0	2	28	8	156	<b>78.03</b>
<i>WV</i>	0	0	0	43	77	<b>602</b>	73	0	0	0	0	0	0	<b>75.72</b>
<i>RE</i>	0	0	0	0	55	152	<b>442</b>	6	0	0	0	0	0	<b>67.48</b>
<i>BU</i>	0	0	0	0	0	0	0	<b>183</b>	0	2	0	0	0	<b>98.92</b>
<i>MS</i>	0	0	0	9	113	0	0	0	<b>143</b>	0	0	0	0	<b>53.96</b>
<i>WK</i>	0	0	0	0	0	0	0	0	0	<b>87</b>	0	0	0	<b>100.00</b>
<i>AP</i>	30	0	0	23	2	0	0	0	0	0	<b>164</b>	0	2	<b>74.21</b>
<i>ED</i>	0	0	0	0	3	0	0	0	0	0	0	<b>427</b>	0	<b>99.30</b>
<i>OT</i>	0	0	0	586	509	130	96	17	0	0	48	0	<b>436</b>	<b>23.93</b>
RP(%)	<b>80.09</b>	<b>100.00</b>	<b>86.42</b>	<b>72.66</b>	<b>80.96</b>	<b>58.50</b>	<b>41.31</b>	<b>87.14</b>	<b>100.00</b>	<b>65.41</b>	<b>53.95</b>	<b>98.16</b>	<b>58.29</b>	

**Table 4.8:** The recognition results solely based on the feature similarity of wedding events without exploiting the event transition modeling.

Events	ME	GE	BE	CS	OP	WV	RE	BU	MS	WK	AP	ED	OT
<b>(a) using audio features only</b>													
<b>RP(%)</b>	34.54	30.14	42.57	78.39	69.81	0	0	71.55	0	12.09	0	34.80	69.32
<b>RR(%)</b>	87.39	59.86	87.21	64.46	88.49	0	0	44.86	0	96.55	0	80.93	13.89
<b>(b) using visual features only</b>													
<b>RP(%)</b>	20.30	12.58	30.64	47.10	62.68	36.61	0	20.40	0	4.32	0	16.59	45.49
<b>RR(%)</b>	87.74	66.20	29.07	45.23	14.81	8.43	0	44.32	0	87.36	0	74.65	11.91
<b>(c) using audiovisual features all</b>													
<b>RP(%)</b>	44.12	21.62	55.13	75.04	76.01	81.48	0	28.72	0	14.38	0	38.52	71.51
<b>RR(%)</b>	93.96	69.72	73.72	73.55	91.29	5.53	0	45.95	0	100.00	0	83.49	21.08

**Table 4.9:** The segmentation results without duration-based filtering (unit: event segments).

Clip	Corr.	Sub.	Ins.	Del.	SP(%)	SR(%)	SF(%)
A	16	1	10	0	59.26	94.12	72.73
B	5	1	0	2	83.33	62.50	71.43
C	28	2	19	5	57.14	80.00	66.67
D	22	1	18	0	53.66	95.65	68.75
E	12	0	6	3	66.67	80.00	72.73
F	12	1	9	1	54.55	85.71	66.67
<b>Avg.</b>					<b>62.44</b>	<b>83.00</b>	<b>71.27</b>

**Table 4.10:** The segmentation results with duration-based filtering (unit: event segments).

Clip	Corr.	Sub.	Ins.	Del.	SP(%)	SR(%)	SF(%)
A	16	1	5	0	72.73	94.12	82.05
B	5	1	0	2	83.33	62.50	71.43
C	27	1	10	7	71.05	77.14	73.97
D	21	1	12	1	61.76	91.30	73.68
E	12	0	3	3	80.00	80.00	80.00
F	11	0	6	3	64.71	78.57	70.97
<b>Avg.</b>					<b>72.26</b>	<b>80.61</b>	<b>76.21</b>

**Table 4.11:** The percentage of total event duration for each of the event categories in our video collection.

Event	ME	GE	BE	CS	OP	WV	RE
	4.45%	1.09%	3.30%	20.53%	35.87%	6.11%	5.03%
Event	BU	MS	WK	AP	ED	OT	
	1.42%	2.04%	0.67%	2.19%	3.30%	14.00%	



**Table 4.12:** LCRF recognition results of all wedding events (unit: seconds).

Events	<i>ME</i>	<i>GE</i>	<i>BE</i>	<i>CS</i>	<i>OP</i>	<i>WV</i>	<i>RE</i>	<i>BU</i>	<i>MS</i>	<i>WK</i>	<i>AP</i>	<i>ED</i>	<i>OT</i>	RR(%)
<i>ME</i>	<b>394</b>	0	30	155	0	0	0	0	0	0	0	0	0	<b>68.05</b>
<i>GE</i>	0	<b>99</b>	18	25	0	0	0	0	0	0	0	0	0	<b>69.72</b>
<i>BE</i>	51	0	<b>339</b>	16	0	0	0	0	0	0	0	0	24	<b>78.84</b>
<i>CS</i>	0	0	0	<b>2221</b>	164	0	0	36	0	12	0	4	236	<b>83.09</b>
<i>OP</i>	0	0	0	403	<b>3967</b>	0	0	0	0	6	0	2	261	<b>85.51</b>
<i>WV</i>	0	0	0	0	356	<b>283</b>	37	0	0	0	95	0	24	<b>35.60</b>
<i>RE</i>	0	0	0	0	270	0	<b>300</b>	0	0	0	85	0	0	<b>45.80</b>
<i>BU</i>	0	0	0	0	11	0	20	<b>58</b>	0	0	9	18	69	<b>31.35</b>
<i>MS</i>	0	0	0	0	17	0	0	0	<b>137</b>	0	0	0	111	<b>51.70</b>
<i>WK</i>	0	0	0	0	39	0	0	5	0	<b>19</b>	0	11	13	<b>21.84</b>
<i>AP</i>	17	0	0	66	0	0	0	0	0	0	<b>48</b>	0	90	<b>21.72</b>
<i>ED</i>	0	0	0	0	0	0	0	0	83	26	0	<b>289</b>	32	<b>67.21</b>
<i>OT</i>	0	0	0	545	749	0	0	0	0	8	56	118	<b>346</b>	<b>18.99</b>
RP(%)	<b>85.28</b>	<b>100.00</b>	<b>87.60</b>	<b>64.73</b>	<b>71.18</b>	<b>100.00</b>	<b>84.03</b>	<b>58.59</b>	<b>62.27</b>	<b>26.76</b>	<b>16.38</b>	<b>65.38</b>	<b>28.69</b>	

**Table 4.13:** LCRF segmentation results without duration-based filtering (unit: event segments).

Clip	Corr.	Sub.	Ins.	Del.	SP(%)	SR(%)	SF(%)
A	16	1	11	0	57.14	94.12	71.11
B	5	1	0	2	83.33	62.50	71.43
C	26	3	17	6	56.52	74.29	64.20
D	20	0	16	3	55.56	86.96	67.80
E	10	1	2	4	76.92	66.67	71.43
F	11	0	15	3	42.31	78.57	55.00
<b>Avg.</b>					<b>61.93</b>	<b>77.19</b>	<b>68.72</b>

**Table 4.14:** LCRF segmentation results with duration-based filtering (unit: event segments).

Clip	Corr.	Sub.	Ins.	Del.	SP(%)	SR(%)	SF(%)
A	13	0	2	4	86.67	76.47	81.25
B	3	1	0	4	75.00	37.50	50.00
C	17	2	5	16	70.83	48.57	57.63
D	16	0	4	7	80.00	69.57	74.42
E	10	0	1	5	90.91	66.67	76.93
F	9	0	4	5	69.23	64.29	66.67
<b>Avg.</b>					<b>78.77</b>	<b>60.51</b>	<b>68.44</b>

**Table 4.15:** Segmentation results in the case when event orderings are available (unit: event segments).

Clip	Det.	Corr.	P(%)	R(%)	F(%)
A	17	16	94.12	94.12	94.12
B	6	5	83.33	62.50	71.43
C	30	28	93.33	80.00	86.15
D	23	23	100.00	100.00	100.00
E	12	12	100.00	80.00	88.89
F	14	11	78.57	78.57	78.57
<b>Avg.</b>			<b>91.56</b>	<b>82.53</b>	<b>86.81</b>

**Table 4.16:** The second-based recognition rate of wedding events for all clips in our video collection.

Clip	A	B	C	D	E	F
	92.55%	86.30%	72.32%	97.66%	73.63%	71.63%

## Chapter 5

# Conclusions and Future Work

### 5.1 Conclusions

The Universal Multimedia Access (UMA) is the final frontier in multimedia research, aiming at enabling unrestricted access to and consumption of multimedia contents, in support of the user from everywhere, at anytime, and along with any devices, networks, and preferences. The fulfillment of UMA is not an easy or a trivial task. It requires not only the availability of useful descriptions about both the multimedia contents and the usage environments but also the existence of effective systems capable of using those information to ensure and maximize the quality of adaptation.

In this dissertation, we respond to the above challenges by first presenting a generic framework for semantic multimedia content adaptation, with the purpose of improving automatic multimedia adaptation at the semantic level. Meanwhile, the design principles behind the framework are explicitly specified, such as the basic requirements of application awareness, content awareness, and semantics extraction, which can be served as guidelines for the development of practical

adaptation applications. To demonstrate the effectiveness of the framework, two technical realizations are presented:

- The object-based system for video recomposition provides effective small size videos that emphasize important aspects of the scene while faithfully retaining the background context.
- The event-based system for wedding analysis is a pioneering work to analyze home videos on the basis of semantic events, whereby to benefit the user's navigation in hours-long contents.

Our work would help to take one step closer to the UMA's objective, but a long and hard road remains ahead. As an example, let us look again at the Chinese classical painting given in Figure 1.5. In terms of semantic concepts, what can be effectively detected therefrom with today's technology falls far behind the human's interpretation, no matter in number or the perceived semantic quality. Specifically, using the current techniques of content analysis, it would probably describe the painting as an "outdoor" scene with some "people" there rather than a more semantic description, such as the scene of an "outdoor" "concert" with a "performer" and some "audience". One reason is that the content analysis heavily relies on the statistical learning of audiovisual patterns. If a semantic concept tends to be irregular in form (e.g. "vapor" and "Chinese wedding") or represents a more abstract notion (e.g. "friend" and "time"), it will be hard or even infeasible to build its statistical models. Thus the semantic gap is still there and becomes a challenge to be resolved.

Further, the study of semantic adaptation is characterized by the need of joint considerations with several closely related issues, such as the ontology, content analysis, and adaptation strategy. For example, the selection of a semantic concept

is often affected by its possibility to be automatically extracted and the frequency to be observed in actual contents. In addition, the collaborative research with other fields would arouse more interesting studies and novel applications. As shown in the example of Section 2.3.1, the incorporation of knowledge in computer graphics is helpful to improve the user's browsing experience in their own photos through 3D presentations.

Therefore, we obviously have a long way to go before making the dream of UMA come true. The way forward is full of opportunities but also challenges. We believe that much more research effort will gather together to enable the corresponding breakthrough. The improvements as a result would be able to change our digital life and further bring us into a whole new digital era.

## 5.2 Future Research

We now highlight a few possible directions for the future research.

- **Improvements to the Framework:** Many aspects of the framework can be improved. For example, the role of elements to play in the formulation of adaptive optimization needs to be further investigated to identify the interrelationship in different adaptation scenarios.
- **Multimodal Adaptation:** Adaptation of multimodal media often needs to perform different adapting operations on individual modalities. The investigation of approaches to make joint adaptation decisions that maximize the overall adaptation utility is necessary.
- **Interactive Adaptation:** A possible way to improve the quality of adaptation is by including users in the loop of making adaptation decisions. The

user feedback can be used to dynamically change the parameter setting or even the configuration of an adaptation engine.

- **Semi-Automatic Ontology Construction:** The semantic concept ontology is still many years away from fully automated construction, but hand-crafting is no longer practical for a large-scale one. The study of semi-automatic building schemes is a promising direction.
- **Format Standardization of Media Descriptions:** Multimedia contents can be manipulated with various operations along the content delivery path. To standardize the encoding formats of media descriptions would facilitate the communications between processing units.
- **General Content Analysis:** While limited-domain detection of semantic concepts has been shown to be effective, the general content analysis remains a challenging open problem. More research efforts should be gathered together to address this issue.



# Bibliography

- [ABBH08] Lora Aroyo, Pieter Bellekens, Martin Björkman, and Geert-Jan Houben. Semantic-based framework for personalised ambient media. *Multimedia Tools and Applications*, 36(1-2):71–87, 2008.
- [AGL03] Gregory D. Abowd, Matthias Gauger, and Andreas Lachenmann. The family video archive: An annotation and browsing environment for home movies. In *Proceedings of the International Workshop on Multimedia Information Retrieval (MIR)*, pages 1–8, 2003.
- [AWSZ05] Ishfaq Ahmad, Xiaohui Wei, Yu Sun, and Ya-Qin Zhang. Video transcoding: an overview of various techniques and research issues. *IEEE Transactions on Multimedia*, 7(5):793–804, 2005.
- [AWW02] Aya Aner-Wolf and Lior Wolf. Video de-abstraction or how to save money on your wedding video. In *Proceedings of the IEEE Workshop on Applications of Computer Vision (WACV)*, pages 264–268, 2002.
- [AYK06] Radhakrishna S.V. Achanta, Wei-Qi Yan, and Mohan S. Kankanhalli. Modeling intent for home video repurposing. *IEEE Multimedia*, 13(1):46–55, 2006.
- [BdWH<sup>+</sup>03] Ian Burnett, Rik Van de Walle, Keith Hill, Jan Bormans, and Fernando Pereira. MPEG-21: goals and achievements. *IEEE Multimedia*, 10(4):60–70, 2003.
- [BGP03] Jan Bormans, Jean Gelissen, and Andrew Perkis. MPEG-21: The 21st century multimedia framework. *IEEE Signal Processing Magazine*, 20(2):53–62, 2003.
- [Bis06] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

- [BKOK04] Noboru Babaguchi, Yoshihiko Kawai, Takehiro Ogura, and Tadahiro Kitahashi. Personalized abstraction of broadcasted american football video by highlight selection. *IEEE Transactions on Multimedia*, 6(4):575–586, 2004.
- [BMM99] Roberto Brunelli, Ornella Mich, and Carla Maria Modena. A survey on the automatic indexing of video data. *Journal of Visual Communication and Image Representation*, 10(2):78–112, 1999.
- [Bow02] Sing T. Bow. *Pattern Recognition and Image Preprocessing*. Marcel Dekker, 2002.
- [BPdWK06] Ian Burnett, Fernando Pereira, Rik Van de Walle, and Rob Koenen, editors. *The MPEG-21 Book*. John Wiley & Sons, 2006.
- [BT03] David Bordwell and Kristin Thompson. *Film Art: An Introduction*. McGraw-Hill, 7th edition, 2003.
- [CAL96] Shun Yan Cheung, Mostafa H. Ammar, and Xue Li. On the use of destination set grouping to improve fairness in multicast video distribution. In *Proceedings of IEEE INFOCOM’96*, pages 553–560, 1996.
- [CCSW06] Min Chen, Shu-Ching Chen, Mei-Ling Shyu, and Kasun Wickramaratna. Semantic event detection via multimodal data mining. *IEEE Signal Processing Magazine*, 23(2):38–46, 2006.
- [CCW03] Wen-Huang Cheng, Wei-Ta Chu, and Ja-Ling Wu. Semantic context detection based on hierarchical audio models. In *Proceedings of the International Workshop on Multimedia Information Retrieval (MIR)*, pages 109–115, 2003.
- [CCW05] Wen-Huang Cheng, Wei-Ta Chu, and Ja-Ling Wu. A visual attention based region-of-interest determination framework for video sequences. *IEICE Transactions on Information and Systems Journal*, (7):1578–1586, 2005.
- [CEJ<sup>+</sup>07] Shih-Fu Chang, Dan Ellis, Wei Jiang, Keansub Lee, Akira Yanagawa, Alexander C. Loui, and Jiebo Luo. Large-scale multimodal semantic concept detection for consumer video. In *Proceedings of the International Workshop on Multimedia Information Retrieval (MIR)*, pages 255–264, 2007.



- [CFGW03] Matthew Cooper, Jonathan Foote, Andreas Girgensohn, and Lynn Wilcox. Temporal event clustering for digital photo collections. In *Proceedings of the ACM International Multimedia Conference (MM)*, pages 364–373, 2003.
- [CFJ05] Vincent Cheung, Brendan J. Frey, and Nebojsa Jojic. Video epitomes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.
- [Cha02] Shih-Fu Chang. Optimal video adaptation and skimming using a utility-based framework. In *Proceedings of the International Tyrrhenian Workshop on Digital Communications (IWDC)*, 2002.
- [CHL<sup>+</sup>05] Wen-Huang Cheng, Chun-Wei Hsieh, Sheng-Kai Lin, Chia-Wei Wang, and Ja-Ling Wu. Robust algorithm for exemplar-based image inpainting. In *Proceedings of the International Conference on Computer Graphics, Imaging and Vision (CGIV)*, 2005.
- [Chu06] Wei-Ta Chu. Semantics-based content analysis and organization in movies and sports videos. *PhD Dissertation, Department of Computer Science and Information Engineering, National Taiwan University*, 2006.
- [CLRS01] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms*. MIT Press, 2nd edition, 2001.
- [cnn] CNN: <http://www.cnn.com/>.
- [CPT04] Antonio Criminisi, Patrick Perez, and Kentaro Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on Image Processing*, 13(9):1200–1212, 2004.
- [CSE05] Andrea Cavallaro, Olivier Steiger, and Touradj Ebrahimi. Semantic video analysis for adaptive content delivery and automatic description. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(10):1200–1209, 2005.
- [CSP01] Shih-Fu Chang, Thomas Sikora, and Atul Puri. Overview of the MPEG-7 standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6):688–695, 2001.

- [CT06] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2nd edition, 2006.
- [CV05] Shih-Fu Chang and Anthony Vetro. Video adaptation: Concepts, technologies, and open issues. *Proceedings of the IEEE*, 93(1):148–2005, 2005.
- [CV07] Rudi L. Cilibrasi and Paul M.B. Vitanyi. The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):370–383, 2007.
- [CWW07] Wen-Huang Cheng, Chia-Wei Wang, and Ja-Ling Wu. Video adaptation for small display based on content recomposition. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(1):43–58, 2007.
- [CXF<sup>+</sup>03] Liqun Chen, Xing Xie, Xin Fan, Wei-Ying Ma, Hong-Jiang Zhang, and Heqin Zhou. A visual attention model for adapting images on small displays. *Multimedia Systems Journal*, 9(4):353–364, 2003.
- [Dev95] Jay L. Devore. *Probability and Statistics for Engineering and the Sciences*. Wadsworth, 4rd edition, 1995.
- [Dje02] Chabane Djeraba. Content-based multimedia indexing and retrieval. *IEEE Multimedia*, 9(2):18–22, 2002.
- [DMK<sup>+</sup>05] Stamatia Dasiopoulou, Vasileios Mezaris, Ioannis Kompatsiaris, Vasileios-Kyriakos Papastathis, and Michael G. Strintzis. Knowledge-assisted semantic video object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(10):1210–1224, 2005.
- [Dra67] Alvin W. Drake. *Fundamentals of Applied Probability Theory*. McGraw-Hill College, 1967.
- [DRP<sup>+</sup>06] Javier Diaz, Eduardo Ros, Francisco Pelayo, Eva M. Ortigosa, and Sonia Mota. FPGA-based real-time optical-flow system. *IEEE Transactions on Circuits and Systems for Video Technology*, 16(2):274–279, 2006.
- [DV01] Chitra Dorai and Svetha Venkatesh. Computational media aesthetics: finding meaning beautiful. *IEEE Multimedia*, 8(4):10–12, 2001.

- [DV03] Chitra Dorai and Svetha Venkatesh. Bridging the semantic gap with computational media aesthetics. *IEEE Multimedia*, 10(2):15–17, 2003.
- [EZW97] Stephen Engel, Xuemei Zhang, and Brian Wandell. Colour tuning in human visual cortex measured with functional magnetic resonance imaging. *Nature*, 388(6637):68–71, 1997.
- [fac] Facebook: <http://www.facebook.com/>.
- [FCL05] Rong-En Fan, Pai-Hsuen Chen, and Chih-Jen Lin. Working set selection using the second order information for training svm. *Journal of Machine Learning Research*, 6:1889–1918, 2005.
- [Fel98] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.
- [FGLJ08] Jianping Fan, Yuli Gao, Hangzai Luo, and Ramesh Jain. Mining multilevel image semantics via hierarchical classification. *IEEE Transactions on Multimedia*, 10(2):167–187, 2008.
- [fil] Digital Recomposition System, FlikFX Pty GmbH Ltd., <http://www.widescreenmuseum.com/flikfx/>.
- [FLE04] Jianping Fan, Hangzai Luo, and A.K. Elmagarmid. Concept-oriented indexing of video databases: toward semantic sensitive retrieval and browsing. *IEEE Transactions on Image Processing*, 13(7):974–992, 2004.
- [fli] Flickr: <http://www.flickr.com/>.
- [Fre04] Linton C. Freeman. *The Development of Social Network Analysis: A Study in the Sociology of Science*. Empirical Press, 2004.
- [GKP94] Ronald L. Graham, Donald Knuth, and Oren Patashnik. *Concrete Mathematics: A Foundation for Computer Science*. Addison-Wesley, 2nd edition, 1994.
- [GLF06] Yuli Gao, Hangzai Luo, and Jianping Fan. Searching and browsing large scale image database using keywords and ontology. In *Proceedings of the ACM International Multimedia Conference (MM)*, pages 811–812, 2006.

- [GNTD06] Amit Kumar Gupta, Saeid Nooshabadi, David Taubman, and Michael Dyer. Realizing low-cost high-throughput general-purpose block encoder for JPEG2000. *IEEE Transactions on Circuits and Systems for Video Technology*, 16(7):843–858, 2006.
- [GPLS03] Daniel Gatica-Perez, Alexander Loui, and Ming-Ting Sun. Finding structure in home videos by probabilistic hierarchical clustering. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(6):539–548, 2003.
- [GT06] Lise Getoor and Ben Taskar. *Introduction to Statistical Relational Learning*. MIT Press, 2006.
- [GW01] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing*. Prentice-Hall, 2nd edition, 2001.
- [HAA97] Youichi Horry, Ken-Ichi Anjyo, and Kiyoshi Arai. Tour into the picture: using a spidery mesh interface to make animation from a single image. In *Proceedings of the ACM SIGGRAPH*, pages 225–232, 1997.
- [HCPW03] Chia-Chiang Ho, Wen-Huang Cheng, Ting-Jian Pan, and Ja-Ling Wu. A user-attention based focus detection framework and its applications. In *Proceedings of the Pacific-Rim Conference on Multimedia (PCM)*, 2003.
- [HCY08] Alexander G. Hauptmann, Michael G. Christel, and Rong Yan. Video retrieval based on semantic concepts. *Proceedings of the IEEE*, 96(4):602–622, 2008.
- [Her07] Luis Herranz. Integrating semantic analysis and scalable video coding for efficient content-based adaptation. *Multimedia Systems Journal*, 13(2):103–118, 2007.
- [HLZ04] Xian-Sheng Hua, Lie Lu, and Hong-Jiang Zhang. Optimization-based automated home video editing system. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(5):572–583, May 2004.
- [Ho03] Chia-Chiang Ho. A study of effective techniques for user oriented video streaming. *PhD Dissertation, Department of Computer Science and Information Engineering, National Taiwan University*, 2003.

- [HQS00] Niels Haering, Richard J. Qian, and M. Ibrahim Sezan. A semantic event-detection approach and its application to detecting hunts in wildlife video. *IEEE Transactions on Circuits and Systems for Video Technology*, 10(6):857–868, 2000.
- [HWC05] Chia-Chiang Ho, Ja-Ling Wu, and Wen-Huang Cheng. A practical foveation-based rate-shaping mechanism for mpeg videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(11):1365–1372, 2005.
- [HWG04] Miska M. Hannuksela, Ye-Kui Wang, and Moncef Gabbouj. Isolated regions in video coding. *IEEE Transactions on Multimedia*, 6(2):259–267, 2004.
- [HX05] Alan Hanjalic and Li-Qun Xu. Affective video content representation and modeling. *IEEE Transactions on Multimedia*, 7(1):143–154, 2005.
- [HZ04] Xian-Sheng Hua and Hong-Jiang Zhang. An attention-based decision fusion scheme for multimedia information retrieval. In *Proceedings of the Pacific-Rim Conference on Multimedia (PCM)*, pages 1001–1010, 2004.
- [IK99] Laurent Itti and Christof Koch. A comparison of feature combination strategies for saliency-based visual attention systems. In *Proceedings of the SPIE Human Vision and Electronic Imaging IV (HVEI)*, pages 473–482, 1999.
- [IKN98] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.
- [J91] Bernd Jähne. *Spatio-Temporal Image Processing: Theory and Scientific Applications*. Springer-Verlag, 1991.
- [Jim05] Ana Belén Benítez Jiménez. Multimedia knowledge: Discovery, classification, browsing, and retrieval. *PhD Dissertation, Graduate School of Arts and Sciences, Columbia University*, 2005.
- [JMF99] A.K. Jain, M.N. Murty, and P.J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, 1999.

- [JP08] Satu Jumisko-Pyykkö. “i would like to see the subtitles and the face or at least hear the voice”: effects of picture ratio and audio-video bitrate ratio on perception of quality in mobile television. *Multimedia Tools and Applications*, 36(1-2):167–184, 2008.
- [KMS05] Hendrik Knoche, John D. McCarthy, and M. Angela Sasse. Can small be beautiful? assessing image resolution requirements for mobile tv. In *Proceedings of the ACM International Multimedia Conference (MM)*, 2005.
- [LCC<sup>+</sup>01] Keansub Lee, Hyun Sung Chang, Seong Soo Chun, Hyungseok Choi, and Sanghoon Sull. Perception-based image transcoding for universal multimedia access. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, volume 2, pages 475–478, 2001.
- [LCS03] Chia-Wen Lin, Yung-Chang Chen, and Ming-Ting Sun. Dynamic region of interest transcoding for multipoint video conferencing. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(10):982–992, 2003.
- [LD06] Ying Li and Chitra Dorai. Instructional video content analysis using audio information. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6):2264–2274, 2006.
- [Len95] Douglas B. Lenat. CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38, 1995.
- [LG04] Jose A. Lay and Ling Guan. Retrieval for color artistry concepts. *IEEE Transactions on Image Processing*, 13(3):326–339, 2004.
- [LG05] Feng Liu and Michael Gleicher. Automatic image retargeting with fisheye-view warping. In *Proceedings of the ACM symposium on User Interface Software and Technology (UIST)*, 2005.
- [LL01] Lin-Shan Lee and Yumin Lee. Voice access of global information for broad-band wireless: technologies of today and challenges of tomorrow. *Proceedings of the IEEE*, 89(1):41–57, 2001.
- [LLH03] Ho Young Lee, Ho Keun Lee, and Yeong Ho Ha. Spatial color descriptor for image retrieval and video segmentation. *IEEE Transactions on Multimedia*, 5(3):358–367, 2003.

- [LMP01] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 282–289, 2001.
- [LS04] Hugo Liu and Push Singh. ConceptNet - a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226, 2004.
- [LTM03] Joo-Hwee Lim, Qi Tian, and Philippe Mulhem. Home photo content modeling for personalized event-based retrieval. *IEEE Multimedia*, 10(4):28–37, 2003.
- [LXMZ03] Hao Liu, Xing Xie, Wei-Ying Ma, and Hong-Jiang Zhang. Automatic browsing of large pictures on mobile devices. In *Proceedings of the ACM International Multimedia Conference (MM)*, pages 148–155, 2003.
- [May79] Peter S. Maybeck. *Stochastic Models, Estimation, and Control*, volume 1. Academic Press, 1979.
- [MHZL07] Tao Mei, Xian-Sheng Hua, He-Qin Zhou, and Shipeng Li. Modeling and mining of users capture intention for home videos. *IEEE Transactions on Multimedia*, 9(1):66–77, 2007.
- [MLZL02] Yu-Fei Ma, Lie Lu, Hong-Jiang Zhang, and Mingjing Li. A user attention model for video summarization. In *Proceedings of the ACM International Multimedia Conference (MM)*, pages 533–542, 2002.
- [MSL99] Rakesh Mohan, John R. Smith, and Chung-Sheng Li. Adapting multimedia internet content for universal access. *IEEE Transactions on Multimedia*, 1(1):104–114, 1999.
- [mys] MySpace: <http://www.myspace.com/>.
- [nba] NBA: <http://www.nba.com/>.
- [net] Netflix: <http://www.netflix.com/>.
- [NH02] Milind R. Naphade and Thomas S. Huang. Extracting semantics from audiovisual content: The final frontier in multimedia retrieval. *IEEE Transactions on Neural Networks*, 13(4):793–810, 2002.

- [NI02] Vidhya Navalpakkam and Laurent Itti. A goal oriented attention guidance model. *Lecture Notes in Computer Science*, 2525:453–461, 2002.
- [NPZ03] Chong-Wah Ngo, Ting-Chuen Pong, and Hong-Jiang Zhang. Motion analysis and segmentation through spatio-temporal slices processing. *IEEE Transactions on Image Processing*, 12(3):341–355, 2003.
- [NST<sup>+</sup>06] Milind Naphade, John R. Smith, Jelena Tesic, Shih-Fu Chang, Winston Hsu, Lyndon Kennedy, Lyndon Kennedy, Alexander Hauptmann, and Jon Curtis. Large-scale concept ontology for multimedia. *IEEE Multimedia*, 13(3):86–91, 2006.
- [NYHK05] Jeho Nam, Man Ro Yong, Youngsik Huh, and Munchurl Kim. Visual content adaptation according to user perception characteristics. *IEEE Transactions on Multimedia*, 7(3):435–445, 2005.
- [PB03] Fernando Pereira and Ian Burnett. Universal multimedia experiences for tomorrow. *IEEE Signal Processing Magazine*, 20(2):63–73, 2003.
- [PBC05] Son Lam Phung, Abdesselam Bouzerdoun, and Douglas Chai. Skin segmentation using color pixel classification: analysis and comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(1):148–154, 2005.
- [PECV07] Leevi Peltola, Cumhuri Erkut, Perry R. Cook, and Vesa Valimaki. Synthesis of hand clapping sounds. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):1021–1029, 2007.
- [PLD05] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.
- [PN04] Zailiang Pan and Chong-Wah Ngo. Structuring home video by snippet detection and pattern parsing. In *Proceedings of the International Workshop on Multimedia Information Retrieval (MIR)*, pages 69–76, 2004.
- [PRO98] R. Paramesan, P. Ramaswamy, and S. Omatu. Regular moments for symmetric images. *IEE Electronics Letters*, 34(15):1481–1482, 1998.



- [PS00] Claudio M. Privitera and Lawrence W. Stark. Algorithms for defining visual regions-of-interest: comparison with eye fixations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(9):970–982, 2000.
- [PSB05] Kedar A. Patwardhan, Guillermo Sapiro, and Marcelo Bertalmio. Video inpainting of occluding and occluded objects. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, volume 2, pages 69–72, 2005.
- [PT05] Costas Panagiotakis and George Tziritas. A speech/music discriminator based on rms and zero-crossings. *IEEE Transactions on Multimedia*, 7(1):155–166, 2005.
- [pva] [http://teachmefinance.com/Scientific\\\_Terms/p-value.html/](http://teachmefinance.com/Scientific\_Terms/p-value.html/).
- [Rep87] Bruno H. Repp. The sound of two hands clapping: an exploratory study. *Journal of the Acoustical Society of America*, 81(4):1100–1109, 1987.
- [RH99] Yong Rui and Thomas S. Huang. Image retrieval: Current techniques, promising directions, and open issues. *Journal of Visual Communication and Image Representation*, 10(1):39–62, 1999.
- [RJ05] Lawrence Rowe and Ramesh Jain. ACM SIGMM retreat report on future directions in multimedia research. *ACM Transactions on Multimedia Computing, Communications and Applications*, 1(1):3–13, 2005.
- [RL02] Eric C. Reed and Jae S. Lim. Optimal multidimensional bit-rate control for video communication. *IEEE Transactions on Image Processing*, 11(8):873–885, 2002.
- [RS05] Zeeshan Rasheed and Mubarak Shah. Detection and representation of scenes in videos. *IEEE Transactions on Multimedia*, 7(6):1097–1105, 2005.
- [Sal83] Gerard Salton. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.
- [Sar] Ramesh Sarukkai. Video search: opportunities and challenges. *The keynote speech at 2005 ACM International Workshop on Multimedia Information Retrieval (MIR)*.

- [Spa01] Lisl M. Spangenberg. *Timeless Traditions: A Couple's Guide to Wedding Customs Around the World*. Universe Publishing, 2001.
- [STG<sup>+</sup>04] Vidya Setlur, Saeko Takagi, Michael Gleicher, Ramesh Raskar, and Bruce Gooch. Automatic image retargeting. Technical report, Computer Science Department, Northwestern University, 2004.
- [Sun02] Hari Sundaram. Segmentation, structure detection and summarization of multimedia sequences. *PhD Dissertation, Graduate School of Arts and Sciences, Columbia University*, 2002.
- [SWS<sup>+</sup>00] Arnold W.M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.
- [SWS05] Cees G.M. Snoek, Marcel Worring, and Arnold W.M. Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the ACM International Multimedia Conference (MM)*, pages 399–402, 2005.
- [SWvG<sup>+</sup>06] Cees G.M. Snoek, Marcel Worring, Jan C. van Gemert, Jan-Mark Geusebroek, and Arnold W.M. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proceedings of the ACM International Multimedia Conference (MM)*, pages 421–430, 2006.
- [TLS04] Belle L. Tseng, Ching-Yung Lin, and John R. Smith. Using MPEG-7 and MPEG-21 for personalizing video. *IEEE Multimedia*, 11(1):42–52, 2004.
- [tre] Trecvid: <http://www-nlpir.nist.gov/projects/trecvid/>.
- [TS06] Yuichiro Takeuchi and Masanori Sugimoto. Video summarization using personal photo libraries. In *Proceedings of the International Workshop on Multimedia Information Retrieval (MIR)*, 2006.
- [Tun02] Yi-Shin Tung. The design and implementation of an MPEG-4 based universal scalable video codec in layered path-tree structure. *PhD Dissertation, Department of Computer Science and Information Engineering, National Taiwan University*, 2002.

- [TV01] Ba Tu Truong and Svetha Venkatesh. Determining dramatic intensification via flashing lights in movies. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 61–64, 2001.
- [TV07] Ba Tu Truong and Svetha Venkatesh. Video abstraction: a systematic review and classification. *ACM Transactions on Multimedia Computing, Communications and Applications*, 3(1):1–37, 2007.
- [TWC<sup>+</sup>08] Ming-Chun Tien, Yi-Tang Wang, Chen-Wei Chou, Kuei-Yi Hsieh, Wei-Ta Chu, and Ja-Ling Wu. Event detection in tennis matches based on video data mining. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2008.
- [vBSE<sup>+</sup>03] Peter van Beek, John R. Smith, Touradj Ebrahimi, Teruhiko Suzuki, and Joel Askelof. Metadata-driven multimedia access. *IEEE Signal Processing Magazine*, 20(2):40–52, 2003.
- [War06] Diane Warner. *Diane Warner's Contemporary Guide to Wedding Ceremonies*. New Page Books, 2006.
- [WB04] Greg Welch and Gary Bishop. An introduction to the kalman filter. Technical report, Department of Computer Science, University of North Carolina at Chapel Hill, 2004.
- [WC06] Hee Lin Wang and Loong-Fah Cheong. Affective understanding in film. *IEEE Transactions on Circuits and Systems for Video Technology*, 16(6):689–704, 2006.
- [WCC<sup>+</sup>07] Chia-Wei Wang, Wen-Huang Cheng, Jun-Cheng Chen, Shu-Sian Yang, and Ja-Ling Wu. Film narrative exploration through analyzing aesthetic elements. In *Proceedings of the International Multimedia Modeling Conference (MMM)*, 2007.
- [WKCK07] Yong Wang, Jae-Gon Kim, Shih-Fu Chang, and Hyung-Myung Kim. Utility-based video adaptation for universal multimedia access (uma) and content-based utility function prediction for real-time video transcoding. *IEEE Transactions on Multimedia*, 9(2):213–220, 2007.
- [WLC06] Huan Wang, Song Liu, and Liang-Tien Chia. Does ontology help in image retrieval?: a comparison between keyword, text ontology and multi-modality ontology approaches. In *Proceedings of the ACM International Multimedia Conference (MM)*, pages 109–112, 2006.

- [WLH00] Yao Wang, Zhu Liu, and Jin-Cheng Huang. Multimedia content analysis-using both audio and visual clues. *IEEE Signal Processing Magazine*, 17(6):12–36, 2000.
- [WOZ01] Yao Wang, Jörn Ostermann, and Ya-Qin Zhang. *Video Processing and Communications*. Prentice Hall, 2001.
- [XLS05] Jun Xin, Chia-Wen Lin, and Ming-Ting Sun. Digital video transcoding. *Proceedings of the IEEE*, 93(1):84–97, 2005.
- [yah] Yahoo!: <http://www.yahoo.com/>.
- [YHZ02] Pei Yin, Xian-Sheng Hua, and Hong-Jiang Zhang. Automatic time stamp extraction system for home videos. In *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 73–76, 2002.
- [you] YouTube: <http://www.youtube.com/>.
- [Zet98] Herbert Zettl. *Sight, Sound, Motion: Applied Media Aesthetics*. Wadsworth, 3rd edition, 1998.
- [ZK01] Tong Zhang and C.C. Jay Kuo. *Content-Based Audio Classification and Retrieval for Audiovisual Data Parsing*. Kluwer, 2001.
- [ZS05] Yun Zhai and Mubarak Shah. Automatic segmentation of home videos. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2005.
- [ZSX07] Amit Zunjarwad, Hari Sundaram, and Lexing Xie. Contextual wisdom: social relations and correlations for multimedia event annotation. In *Proceedings of the ACM International Multimedia Conference (MM)*, pages 615–624, 2007.
- [ZWL01] Hua Zhong, Liu Wenyin, and Shipeng Li. Interactive tracker - a semi-automatic video object tracking and segmentation system. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2001.