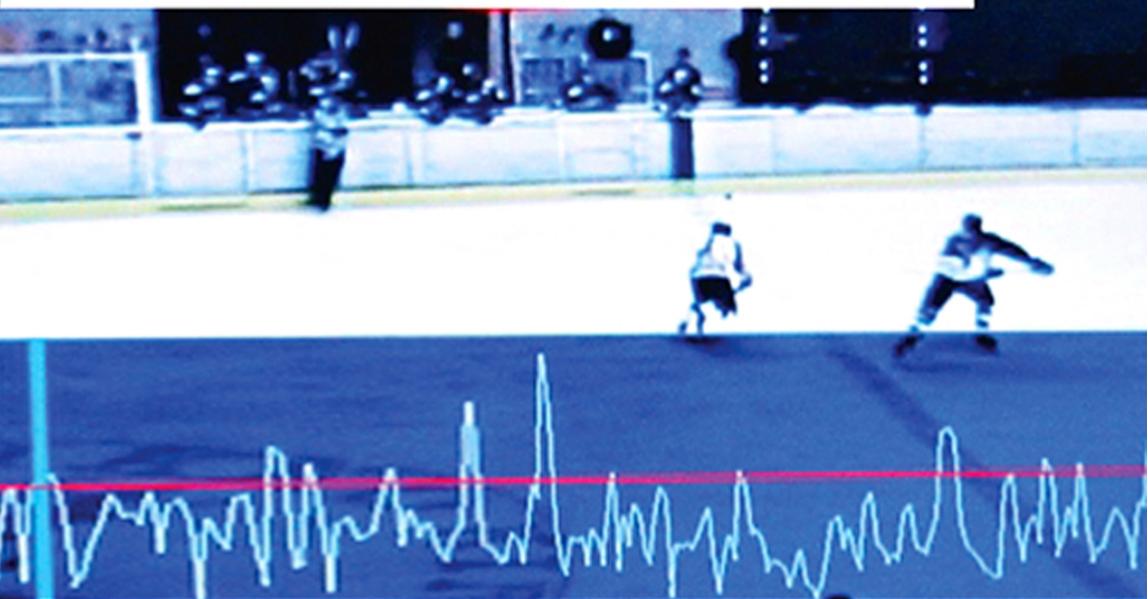


ZIYOU XIONG • REGUNATHAN RADHAKRISHNAN
AJAY DIVAKARAN • YONG RUI • THOMAS HUANG



A UNIFIED FRAMEWORK FOR VIDEO SUMMARIZATION, BROWSING AND RETRIEVAL

with APPLICATIONS TO CONSUMER AND SURVEILLANCE VIDEO



**A UNIFIED FRAMEWORK
FOR VIDEO SUMMARIZATION,
BROWSING, AND RETRIEVAL**

This Page Intentionally Left Blank

A UNIFIED FRAMEWORK FOR VIDEO SUMMARIZATION, BROWSING, AND RETRIEVAL

with Applications to Consumer and Surveillance Video

Ziyou Xiong,

Regunathan Radhakrishnan,

Ajay Divakaran,

Yong Rui, and

Thomas S. Huang



ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

Academic Press is an imprint of Elsevier



Elsevier Academic Press
30 Corporate Drive, Suite 400, Burlington, MA 01803, USA
525 B Street, Suite 1900, San Diego, California 92101-4495, USA
84 Theobald's Road, London WC1X 8RR, UK

This book is printed on acid-free paper. (∞)

Copyright © 2006, Elsevier Inc. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the publisher.

Permissions may be sought directly from Elsevier's Science & Technology Rights Department in Oxford, UK: phone: (+44) 1865 843830, fax: (+44) 1865 853333, E-mail: permissions@elsevier.co.uk. You may also complete your request online via the Elsevier homepage (<http://elsevier.com>), by selecting "Customer Support" and then "Obtaining Permissions."

Library of Congress Cataloging-in-Publication Data

A unified framework for video summarization, browsing, and retrieval with applications to consumer and surveillance video/Ziyou Xiong . . . [et al.].
p. cm.

Includes bibliographical references and index.

ISBN 13: 978-0-12-369387-7 (hardcover : alk. paper)

ISBN 10: 0-12-369387-X (hardcover : alk. paper) 1. Digital video—Indexes.

2. Video recordings—Indexes. 3. Automatic abstracting. 4. Database management.
5. Image processing—Digital techniques. I. Xiong, Ziyou.

TK6680.5.U55 2006

006.3'7—dc22

2005027690

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN 13: 978-0-12-369387-7

ISBN 10: 0-12-369387-X

For all information on all Elsevier Academic Press publications
visit our Web site at www.books.elsevier.com

Printed in the United States of America

05 06 07 08 09 10 9 8 7 6 5 4 3 2 1

Working together to grow
libraries in developing countries

www.elsevier.com | www.bookaid.org | www.sabre.org

ELSEVIER

BOOK AID
International

Sabre Foundation

To my parents Anmei, Xiong and Jinlan Lu, my brothers, Zisong and Zixiang, and my sister, Tian'e.

— Ziyou Xiong

To my parents, Malathi and Radhakrishnan, my sisters, Rup and Krithika, and my professor and friends.

— Regunathan Radhakrishnan

To my daughter, Swathi, my wife, Padma, and my parents, Bharathi and S. Divakaran.

— Ajay Divakaran

To Dongquin and Olivia.

— Yong Rui

To my students: past, present, and future.

— Thomas S. Huang

This Page Intentionally Left Blank

Contents

<i>List of Figures</i>	<i>xi</i>
<i>List of Tables</i>	<i>xvii</i>
<i>Preface</i>	<i>xix</i>
<i>Acknowledgments</i>	<i>xxi</i>
Chapter 1 Introduction	1
1.1 Introduction	1
1.2 Terminology	3
1.3 Video Analysis	6
1.3.1 Shot Boundary Detection	6
1.3.2 Key Frame Extraction	6
1.3.3 Play/Break Segmentation	7
1.3.4 Audio Marker Detection	7
1.3.5 Video Marker Detection	7
1.4 Video Representation	7
1.4.1 Video Representation for Scripted Content	8
1.4.2 Video Representation for Unscripted Content	9
1.5 Video Browsing and Retrieval	11
1.5.1 Video Browsing Using ToC-Based Summary	11
1.5.2 Video Browsing Using Highlights-Based Summary	11
1.5.3 Video Retrieval	12
1.6 The Rest of the Book	12
Chapter 2 Video Table-of-Content Generation	15
2.1 Introduction	15
2.2 Related Work	17
2.2.1 Shot- and Key Frame-Based Video ToC	17
2.2.2 Group-Based Video ToC	18
2.2.3 Scene-Based Video ToC	19
2.3 The Proposed Approach	20
2.3.1 Shot Boundary Detection and Key Frame Extraction	20
2.3.2 Spatiotemporal Feature Extraction	20
2.3.3 Time-Adaptive Grouping	21
2.3.4 Scene Structure Construction	24

2.4	Determination of the Parameters	30
2.4.1	Gaussian Normalization	30
2.4.2	Determining W_C and W_A	31
2.4.3	Determining <i>groupThreshold</i> and <i>sceneThreshold</i>	32
2.5	Experimental Results	33
2.6	Conclusions	37
 Chapter 3 Highlights Extraction from Unscripted Video		 39
3.1	Introduction	39
3.1.1	Audio Marker Recognition	39
3.1.2	Visual Marker Detection	39
3.1.3	Audio-Visual Marker Association and Finer-Resolution Highlights	41
3.2	Audio Marker Recognition	42
3.2.1	Estimating the Number of Mixtures in GMMs	42
3.2.2	Evaluation Using the Precision-Recall Curve	44
3.2.3	Performance Comparison	46
3.2.4	Experimental Results on Golf Highlights Generation	47
3.3	Visual Marker Detection	52
3.3.1	Motivation	52
3.3.2	Choice of Visual Markers	52
3.3.3	Robust Real-Time Object Detection Algorithm	60
3.3.4	Results of Baseball Catcher Detection	62
3.3.5	Results of Soccer Goalpost Detection	64
3.3.6	Results of Golfer Detection	68
3.4	Finer-Resolution Highlights Extraction	71
3.4.1	Audio-Visual Marker Association	71
3.4.2	Finer-Resolution Highlights Classification	71
3.4.3	Method 1: Clustering	72
3.4.4	Method 2: Color/Motion Modeling Using HMMs	73
3.4.5	Method 3: Audio-Visual Modeling Using CHMMs	82
3.4.6	Experimental Results with DCHMM	85
3.5	Conclusions	96
 Chapter 4 Video Structure Discovery Using Unsupervised Learning		 97
4.1	Motivation and Related Work	97
4.2	Proposed Inlier/Outlier-Based Representation for “Unscripted” Multimedia Using Audio Analysis	98
4.3	Feature Extraction and the Audio Classification Framework	101
4.3.1	Feature Extraction	102
4.3.2	Mel Frequency Cepstral Coefficients (MFCC)	102

4.3.3	Modified Discrete Cosine Transform (MDCT) Features from AC-3 Stream	103
4.3.4	Audio Classification Framework	109
4.4	Proposed Time Series Analysis Framework	111
4.4.1	Problem Formulation	112
4.4.2	Kernel /Affinity Matrix Computation	113
4.4.3	Segmentation Using Eigenvector Analysis of Affinity Matrices	114
4.4.4	Past Work on Detecting “Surprising” Patterns from Time Series	117
4.4.5	Proposed Outlier Subsequence Detection in Time Series	119
4.4.6	Generative Model for Synthetic Time Series	121
4.4.7	Performance of the Normalized Cut for Case 2	122
4.4.8	Comparison with Other Clustering Approaches for Case 2	127
4.4.9	Performance of Normalized Cut for Case 3	135
4.5	Ranking Outliers for Summarization	141
4.5.1	Kernel Density Estimation	141
4.5.2	Confidence Measure for Outliers with Binomial and Multinomial PDF Models for the Contexts	142
4.5.3	Confidence Measure for Outliers with GMM and HMM Models for the Contexts	149
4.5.4	Using Confidence Measures to Rank Outliers	153
4.6	Application to Consumer Video Browsing	154
4.6.1	Highlights Extraction from Sports Video	154
4.6.2	Scene Segmentation for Situation Comedy Videos	171
4.7	Systematic Acquisition of Key Audio Classes	179
4.7.1	Application to Sports Highlights Extraction	179
4.7.2	Event Detection in Elevator Surveillance Audio	185
4.8	Possibilities for Future Research	192
Chapter 5 Video Indexing		199
5.1	Introduction	199
5.1.1	Motivation	199
5.1.2	Overview of MPEG-7	199
5.2	Indexing with Low-Level Features: Motion	200
5.2.1	Introduction	200
5.2.2	Overview of MPEG-7 Motion Descriptors	201
5.2.3	Camera Motion Descriptor	201
5.2.4	Motion Trajectory	203
5.2.5	Parametric Motion	203
5.2.6	Motion Activity	204
5.2.7	Applications of Motion Descriptors	206
5.2.8	Video Browsing System Based on Motion Activity	208
5.2.9	Conclusion	212
5.3	Indexing with Low-Level Features: Color	212
5.4	Indexing with Low-Level Features: Texture	213
5.5	Indexing with Low-Level Features: Shape	214
5.6	Indexing with Low-Level Features: Audio	215

5.7	Indexing with User Feedback	217
5.8	Indexing Using Concepts	218
5.9	Discussion and Conclusions	219
Chapter 6	A Unified Framework for Video Summarization, Browsing, and Retrieval	221
6.1	Video Browsing	221
6.2	Video Highlights Extraction	223
6.2.1	Audio Marker Detection	223
6.2.2	Visual Marker Detection	224
6.2.3	Audio-Visual Markers Association for Highlights Candidates Generation	225
6.2.4	Finer-Resolution Highlights Recognition and Verification	226
6.3	Video Retrieval	227
6.4	A Unified Framework for Summarization, Browsing, and Retrieval	229
6.5	Conclusions and Promising Research Directions	235
Chapter 7	Applications	237
7.1	Introduction	237
7.2	Consumer Video Applications	238
7.2.1	Challenges for Consumer Video Browsing Applications	241
7.3	Image/Video Database Management	242
7.4	Surveillance	244
7.5	Challenges of Current Applications	247
7.6	Conclusions	247
Chapter 8	Conclusions	249
	Bibliography	253
	About the Authors	261
	Index	265

List of Figures

Figure 1.1	Relations between the five research areas.	2
Figure 1.2	A hierarchical video representation for scripted content.	5
Figure 1.3	A hierarchical video representation for unscripted content.	5
Figure 2.1	A hierarchical video representation for scripted video content.	16
Figure 2.2	An example video ToC.	27
Figure 2.3	Merging scene 1 to scene 0.	29
Figure 2.4	The Gaussian $N(0, 1)$ distribution.	32
Figure 2.5	Proposed approach for detecting documentary scenes.	36
Figure 2.6	Video ToC for Movie1 (scene level).	36
Figure 2.7	Video ToC for Movie1 (group level).	37
Figure 3.1	Our proposed framework: an overview.	40
Figure 3.2	Audio markers for sports highlights extraction.	40
Figure 3.3	Examples of visual markers for different sports.	41
Figure 3.4	$MDL(K, \theta)$ (Y-axis) with respect to different numbers of GMM mixtures K (X-axis) to model (a) applause, (b) cheering, (c) music, (d) speech, and (e) speech with music sound shown in the raster-scan order. $K = 1 \cdots 20$. The optimal mixture numbers at the lowest positions of the curves are 2, 2, 4, 18, and 8, respectively.	45
Figure 3.5	Precision-recall curves for the test golf game. Left: by the current approach; right: by the previous approaches; Y-axis: precision; X-axis: recall.	49
Figure 3.6	The interface of our system displaying sports highlights. The horizontal line imposed on the curve is the threshold value the user can choose to display those segments with a confidence level greater than the threshold.	50
Figure 3.7	A snapshot of the interface of the results in our previous approach. The horizontal line imposed on the curve is the threshold.	51
Figure 3.8	Some examples of the typical view of the squatting baseball catcher.	53
Figure 3.9	A typical video camera setup for live soccer broadcast.	54
Figure 3.10	Some examples of the typical view of the first view of the goalpost.	55
Figure 3.11	Some examples of the typical view of the second view of the goalpost.	56
Figure 3.12	Some examples of the first view of the golfer.	57
Figure 3.13	Some examples of the second view of the golfer.	58
Figure 3.14	Some examples of the third view of the golfer.	59

Figure 3.15	Example rectangle features shown relative to the enclosing detection window. The sum of the pixels that lie within the white rectangles are subtracted from the sum of pixels in the gray rectangles. Two-rectangle features are shown in (a) and (b). Figure (c) shows a three-rectangle feature, and (d) shows a four-rectangle feature.	60
Figure 3.16	The AdaBoost algorithm.	61
Figure 3.17	Schematic depiction of the detection cascade. A series of classifiers are applied to every subwindow. The initial classifier eliminates a large number of negative examples with very little processing. Subsequent layers eliminate additional negatives but require additional computation. After several stages of processing, the number of subwindows has been reduced radically. Further processing can take any form, such as additional stages of the cascade or an alternative detection system.	62
Figure 3.18	The first few weak classifiers learned by the AdaBoost algorithm for the catcher model.	63
Figure 3.19	The precision-recall curve of baseball catcher detection.	64
Figure 3.20	Two examples of the preprocessing step for the two soccer goalpost views. The thumbnail images (bottom row) are taken from the “whitened” images (middle row).	66
Figure 3.21	The first few weak classifiers learned by the AdaBoost algorithm for the first view of the goalpost model.	67
Figure 3.22	The first few weak classifiers learned by the AdaBoost algorithm for the second view of the goalpost model.	67
Figure 3.23	The precision-recall curve of goalpost detection for view 1.	68
Figure 3.24	The first few weak classifiers learned by the AdaBoost algorithm for the first view of the golfer model.	69
Figure 3.25	The first few weak classifiers learned by the AdaBoost algorithm for the second view of the golfer model.	70
Figure 3.26	An example of the change of color characteristics in a baseball hit.	73
Figure 3.27	The scaled version of each video frame’s average hue value over time for the 18 training “putt” sequences. The scaling factor is $1000/\text{MAX}(\cdot)$. X-axis: video frames; Y-axis: scaled average hue values.	75
Figure 3.28	The scaled version of each video frame’s average hue value over time for the 32 training “swing” sequences. The scaling factor is $1000/\text{MAX}(\cdot)$. X-axis: video frames; Y-axis: scaled average hue values.	76
Figure 3.29	The scaled version of each video frame’s average hue value over time for the 95 test sequences.	77
Figure 3.30	The scaled version of each video frame’s average motion intensity value over time for the 18 training “putt” sequences. The scaling factor is $1000/\text{MAX}(\cdot)$. X-axis: video P-frames; Y-axis: scaled average motion intensity values.	78
Figure 3.31	The scaled version of each video frame’s average motion intensity value over time for the 32 training “swing” sequences. The scaling factor is $1000/\text{MAX}(\cdot)$. X-axis: video P-frames; Y-axis: scaled average motion intensity values.	79

Figure 3.32	The scaled version of each video frame's average motion intensity value over time for the 95 test sequences. The scaling factor is $1000/\text{MAX}(\cdot)$. X-axis: video P-frames; Y-axis: scaled average motion intensity values.	80
Figure 3.33	Comparison results of three different modeling approaches in terms of precision-recall curves. Solid line: audio modeling alone; dashed line: audio + dominant color modeling; dotted line: audio + dominant color + motion modeling. X-axis: recall; Y-axis: precision.	81
Figure 3.34	The graphical model structure of the DCHMM. The two rows of squares are the coupled hidden state nodes. The circles are the observation nodes.	83
Figure 3.35	Precision-recall curves for the test golf game. The four highlights extraction methods compared here are (1) audio classification followed by long, contiguous applause selection; (2) HMM classification using the models trained from audio labels of the highlight and non-highlight examples; (3) HMM classification using the models trained from video (motion) labels of the highlight and nonhighlight examples; and (4) coupled HMM classification using the models trained from both audio and video (motion) labels of the highlight and non-highlight examples. X-axis: recall; Y-axis: precision. The first two curves coincide with each other (see text for detail).	92
Figure 3.36	Precision-recall curves for the test soccer game. The four highlights extraction methods compared here are (1) audio classification followed by long, contiguous cheering selection; (2) HMM classification using the models trained from audio labels of the highlight and non-highlight examples; (3) HMM classification using the models trained from video (motion) labels of the highlight and nonhighlight examples; and (4) coupled HMM classification using the models trained from both audio and video (motion) labels of the highlight and nonhighlight examples. X-axis: recall; Y-axis: precision.	94
Figure 3.37	A snapshot of the interface of our system. The horizontal line imposed on the curve is the threshold.	96
Figure 4.1	A hierarchical video representation for unscripted content.	98
Figure 4.2	Proposed event discovery framework for analysis and representation of unscripted content for summarization.	99
Figure 4.3	MFCC feature extraction.	103
Figure 4.4	MFCC feature extraction.	104
Figure 4.5	MDCT coefficients for the first block.	105
Figure 4.6	MDCT coefficients for the second block.	106
Figure 4.7	MDCT coefficients for a frame from the mean of six MDCT blocks without absolute value operation.	107
Figure 4.8	MDCT coefficients for a frame from the mean of six MDCT blocks with the absolute value operation.	108
Figure 4.9	Energy normalization for MDCT coefficients.	108
Figure 4.10	η Vs q for effective rank.	109
Figure 4.11	μ Vs q for effective rank.	110
Figure 4.12	Audio classification framework.	111

Figure 4.13	Input time series. (a) Time along X-axis and symbols (1, 2, 3) along Y-axis. (b) The computed affinity matrix.	115
Figure 4.14	Zoomed-in view of the input time series from P_1 and P_2 , time along X-axis and symbols (1, 2, 3) along Y-axis.	116
Figure 4.15	Proposed outlier subsequence detection framework.	120
Figure 4.16	Generative model for synthetic time series with one background process and one foreground process.	122
Figure 4.17	Cases of interest.	123
Figure 4.18	Performance of normalized cut on synthetic time series for case 2. (a) X-axis for time, Y-axis for symbol. (b) Affinity matrix. (c) X-axis for candidate normalized cut threshold, Y-axis for value of normalized cut objective function. (d) X-axis for time index of context model, Y-axis for cluster indicator value. (e) X-axis for time, Y-axis for symbol.	124
Figure 4.19	Performance of k -means on synthetic time series for case 2. (a) X-axis for time index of context model, Y-axis for cluster indicator value. (b) X-axis for time, Y-axis for symbol.	129
Figure 4.20	(a) Performance of dendrogram cut on synthetic time series for case 2. (b) X-axis for time, Y-axis for symbol.	130
Figure 4.21	Structured “salient” foreground in unstructured background.	131
Figure 4.22	Performance of modified normalized cut on synthetic time series for case 2. (a) X-axis for candidate threshold for cut, Y-axis for cut value. (b) X-axis for candidate threshold for modified normalized cut, Y-axis for modified normalized cut value. (c) X-axis for time index of context model, Y-axis for cluster indicator value. (d) X-axis for time, Y-axis for symbol.	133
Figure 4.23	Performance comparison of normalized cut and modified normalized cut on synthetic time series for case 3. (a) X-axis for time, Y-axis for symbol. (b) X-axis for candidate threshold for normalized cut, Y-axis for normalized cut value. (c) X-axis for time index of context model, Y-axis for cluster indicator value (normalized cut). (d) X-axis for time index of context model, Y-axis for cluster indicator value (foreground cut).	136
Figure 4.24	Performance of hybrid (normalized cut and foreground cut) approach on synthetic time series for case 4. (a) X-axis for time, Y-axis for symbol. (b) (<i>top</i>) X-axis for time index of context model, Y-axis for cluster indicator; (<i>bottom</i>) corresponding temporal segmentation. (c) (<i>top</i>) X-axis for time index of context model, Y-axis for cluster indicator (second normalized cut); (<i>bottom</i>) corresponding temporal segmentation. (d) Final temporal segmentation.	138
Figure 4.25	PDF of the defined distance metric for binomial PDF as a background model for context sizes of 200 and 400 symbols.	144
Figure 4.26	PDF of the defined distance metric for multinomial PDF as a background model for a context size of 200.	147