

Multimodal Video Characterization And Summarization

THE KLUWER INTERNATIONAL SERIES IN VIDEO COMPUTING

Series Editor

Mubarak Shah, Ph.D.

*University of Central Florida
Orlando, USA*

Other books in the series:

3D FACE PROCESSING: MODELING, ANALYSIS AND SYNTHESIS

Zhen Wen, Thomas S. Huang ; ISBN: 1-4020-8047-6

EXPLORATION OF VISUAL DATA

Xiang Sean Zhou, Yong Rui, Thomas S. Huang ; ISBN: 1-4020-7569-3

VIDEO MINING

Edited by Azriel Rosenfeld, David Doermann, Daniel DeMenthon; ISBN: 1-4020-7549-9

VIDEO REGISTRATION

Edited by Mubarak Shah, Rakesh Kumar; ISBN: 1-4020-7460-3

MEDIA COMPUTING: COMPUTATIONAL MEDIA AESTHETICS

Chitra Dorai and Svetha Venkatesh; ISBN: 1-4020-7102-7

**ANALYZING VIDEO SEQUENCES OF MULTIPLE HUMANS: Tracking, Posture
Estimation and Behavior Recognition**

Jun Ohya, Akita Utsumi, and Junji Yanato; ISBN: 1-4020-7021-7

VISUAL EVENT DETECTION

Niels Haering and Niels da Vitoria Lobo; ISBN: 0-7923-7436-3

**FACE DETECTION AND GESTURE RECOGNITION FOR HUMAN-COMPUTER
INTERACTION**

Ming-Hsuan Yang and Narendra Ahuja; ISBN: 0-7923-7409-6

Multimodal Video Characterization And Summarization

Michael A. Smith

*AVA Media Systems
Carnegie Mellon University*

Takeo Kanade

Carnegie Mellon University

KLUWER ACADEMIC PUBLISHERS

NEW YORK, BOSTON, DORDRECHT, LONDON, MOSCOW

eBook ISBN: 0-387-23008-4
Print ISBN: 1-4020-7426-3

©2005 Springer Science + Business Media, Inc.

Print ©2005 Kluwer Academic Publishers
Boston

All rights reserved

No part of this eBook may be reproduced or transmitted in any form or by any means, electronic, mechanical, recording, or otherwise, without written consent from the Publisher

Created in the United States of America

Visit Springer's eBookstore at:
and the Springer Global Website Online at:

<http://ebooks.springerlink.com>
<http://www.springeronline.com>

Table of Contents

Series Foreword	ix
Acknowledgements	xi
1. Introduction	1
1.1 Video Characterization	2
1.2 Browsing Digital Video	8
1.3 Digital Video Libraries	8
1.4 Characterization and Summarization Research	9
1.5 Book Overview	12
1.6 Bibliography	12
2. Video Structure and Terminology	17
2.1 Video Terminology	18
2.2 Video Categories Used in this Book	21
2.2.1 Documentaries	21
2.2.2 Broadcast News	22
2.2.3 Feature-Films	23
2.2.4 Sports	24
2.3 Video Production and Editing Standards	31
2.3.1 Video Cuts - Shot Changes	31
2.3.2 Video Captions and Graphics	36
2.3.3 Motion Video	38
2.3.4 Subject Positioning	41
2.3.5 Angle Shots	42
2.3.6 Camera Focus	43
2.3.7 Lighting and Mattes	44
2.3.8 Grayscale Video	46
2.3.9 Audio Effects	47
2.3.10 Word Selection	48
2.4 Visualization Systems in Use Today	49
2.5 Higher Level Interpretation for Summarization	52
2.6 Conclusions	52
2.7 Bibliography	56
3. Multimodal Video Characterization	61
3.1 Video Characterization Representations	62
3.2 Video Features	63
3.3 Audio and Language Understanding	64

3.3.1 Speech Transcription	65
3.3.2 Language Characterization	65
3.3.3 Audio Segmentation and Keyword Extraction	70
3.4 Image Understanding	71
3.4.1 Image Difference and Histogram Analysis	71
3.4.2 Histograms for Segmentation and Image Correspondence	76
3.4.3 Texture and Edge Features	78
3.4.4 Camera Motion	79
3.4.5 Motion for Segmentation	85
3.4.6 Object Motion Detection	86
3.4.7 Caption Detection	88
3.4.8 Object Detection	94
3.5 Compressed Video Analysis	98
3.6 Embedded Content Features	104
3.7 Conclusions	105
3.8 Bibliography	105
4. Video Summarization	111
4.1 Video Skims	112
4.2 Automatic Video Surrogates	114
4.3 Audio Skim Selection	116
4.4 Image Skim Selection	123
4.5 Primitive Skim Selection Rules	123
4.6 Feature Integration Meta-Rules	129
4.7 Super-Rules for Genre and Visual Presentation	135
4.8 Rule Hierarchy and Tests for Visual Quality	140
4.9 Conclusions	143
4.10 Bibliography	146
5. Visualization Techniques	149
5.1 Visualization Categories	150
5.2 Evaluation Methods for Visualization	151
5.3 Text Titles	152
5.4 Thumbnail Images	154
5.5 Storyboards	156
5.6 Storyboard Plus Text	158
5.7 Skim Visualizations	160
5.8 Lessons from Single Document Surrogates	161
5.9 Temporal and Spatial Visualizations	165

5.10 Multiple Document Summarization	165
5.11 Visualization Interface Settings	167
5.12 Conclusions	169
5.13 Bibliography	169
6. Evaluation	173
6.1 Experimentation and Evaluation	175
6.2 Skims used for User Studies	175
6.3 Skim Study - Experiment I	178
6.4 Skim Study - Experiment II	182
6.5 Poster Frame Study	193
6.6 Conclusions	194
6.7 Bibliography	195
7. Conclusions	197
Index	201

Series Foreword

Traditionally, scientific fields have defined boundaries, and scientists work on research problems within those boundaries. However, from time to time those boundaries get shifted or blurred to evolve new fields. For instance, the original goal of computer vision was to understand a single image of a scene, by identifying objects, their structure, and spatial arrangements. This has been referred to as image understanding. Recently, computer vision has gradually been making the transition away from understanding single images to analyzing image sequences, or video understanding. Video understanding deals with understanding of video sequences, e.g., recognition of gestures, activities, facial expressions, etc. The main shift in the classic paradigm has been from the recognition of static objects in the scene to motion-based recognition of actions and events.

Video understanding has overlapping research problems with other fields, therefore blurring the fixed boundaries. Computer graphics, image processing, and video databases have obvious overlap with computer vision. The main goal of computer graphics is to generate and animate realistic looking images, and videos. Researchers in computer graphics are increasingly employing techniques from computer vision to generate the synthetic imagery. A good example of this is image-based rendering and modeling techniques, in which geometry, appearance, and lighting is derived from real images using computer vision techniques. Here the shift is from synthesis to analysis followed by synthesis. Image processing has always overlapped with computer vision because they both inherently work directly with images. One view is to consider image processing as low-level computer vision, which processes images, and video for later analysis by high-level computer vision techniques. Databases have traditionally contained text, and numerical data. However, due to the current availability of video in digital form, more and more databases are containing video as content. Consequently, researchers in databases are increasingly applying computer vision techniques to analyze the video before indexing. This is essentially analysis followed by indexing.

Due to MPEG-4 and MPEG-7 standards, there is a further overlap in research for computer vision, computer graphics, image processing, and databases. In a typical model-based coding for MPEG-4, video is first analyzed to estimate local and global motion then the video is synthesized using the estimated parameters. Based on the difference between the real video and synthesized video, the model parameters are updated and finally coded for transmission. This is essentially analysis followed by synthesis, followed by model update, and followed by coding. Thus, in order to solve research problems in the context of the MPEG-4 codec, researchers from different video computing fields will need to collaborate. Similarly, MPEG-7

is bringing together researchers from databases, and computer vision to specify a standard set of descriptors that can be used to describe various types of multimedia information. Computer vision researchers need to develop techniques to automatically compute those descriptors from video, so that database researchers can use them for indexing. Due to the overlap of these different areas, it is meaningful to treat video computing as one entity, which covers the parts of computer vision, computer graphics, image processing, and databases that are related to video. This international series on Video Computing will provide a forum for the dissemination of innovative research results in video computing, and will bring together a community of researchers, who are interested in several different aspects of video.

Mubarak Shah
University of Central Florida, Orlando

Acknowledgements

I would like to thank my family, Mr. and Mrs. Richard Smith Jr., Veneka, Nyasha, Caymen, and Carmen. The authors wish to thank the supporting members of the Carnegie Mellon University Vision and Autonomous Systems Center, including Toshio Sato, Shin'ichi Sato, Yuichi Nakamura, Henry Rowley, Henry Schneiderman, Tsuhan Chen, and Rawesak Tanawongsuwan. The authors also wish to thank the supporting members from the Carnegie Mellon University Informedia Digital Video Library Project, including Howard Wactlar, Michael Christel, Alex Hauptmann, and Scott Stevens.

A portion of the material in this chapter was developed by AVA Media Systems. The Informedia material is based on work supported by the National Science Foundation (NSF) under Cooperative Agreement No. IRI-9817496. A portion of this research is also supported in part by the advanced Research and Development Activity (ARDA) under contract number MDA908-00-C-0037. CNN and WQED Communications in Pittsburgh, PA supplied video to the Informedia library for sole use in research. Their video contributions as well as video from NASA, the U.S. Geological Survey, and U.S. Bureau of Reclamation are gratefully acknowledged.

Michael A. Smith
Current Affiliation - Director, Digital Content Management
France Telecom R&D, San Francisco, CA.