ICIAP 2003

Content-Based Video Summarization and Adaptation for Ubiquitous Media Access

Shih-Fu Chang

Digital Video and Multimedia Lab
Columbia University
Sept. 19th 2003
http://www.ee.columbia.edu/dvmm
Example Driving Applications

- Broadcast users (summary, navigation)
- Mobile users (streaming, skim, highlight)
- Internet users (search, browsing)

Research Activities

Systems and Testbeds
- MPEG-7 and MPEG-21
- NSF Digital Library II: PERSIVAL Health Care DL
- ARDA VACE Information Analysis
- Consumer Media Management

Content Analysis & Management
- Audio-Video Event/Structure Mining
- Multimedia Highlight/Skim Generation
- Multimodal Fusion
- Interactive Retrieval
- Multimedia Semantic Ontology Construction

Pervasive Media Delivery
- Content-Adaptive Video Streaming
- Utility-Based Video Transcoding
- Spatio-Temporal Optimal Scalable Video
- Distributed Network Caching with Content-Aware QoS

Media Security
- Robust Content Authentication/Watermarking
- Information Hiding
- Watermarking for Error Concealment
DVMM @ Columbia: Digital Video and Multimedia Research Lab

Example Projects

Adaptive Video Streaming and Event Summary

VideoQ

Important segments: video mode

Video Skimming

Non-important segments: still frame + audio + captions

Video Object Search by Motion Sketch

Robust Content Feature

Digital Signature/ Watermark

Camera/Transmission

Image/Video

Content-Based Authentication/ Watermarking

http://www.ee.columbia.edu/dvmm
Research Activities

Systems, Applications, Projects
- MPEG-7 and MPEG-21
- NSF Digital Library II: PERSIVAL Health Care DL
- ARDA VACE Information Analysis
- Consumer Media Album with Industry

Content Analysis & Management
- Audio-Video Event/Structure Mining
- Multimedia Highlight/Skim Generation
- Multimodal Fusion
- Interactive Retrieval
- Multimedia Semantic Ontology Construction

Pervasive Media Delivery
- Content-Adaptive Video Streaming
- Utility-Based Video Transcoding
- Spatio-Temporal Optimal Scalable Video
- Distributed Network Caching with Content-Aware QoS

Rights Management & Security
- Robust Content Authentication/Watermarking
- Information Hiding
- Watermarking for Error Concealment

Driving Applications

Broadcast
- Mobile users (summary, navigation)
- Internet users (search, ontology)

Content management & exchange
- Broadcast production aggregation capturing

Internet
- Broadcast users (summary, navigation)
- Mobile users (streaming, skim, highlight)

Example Projects

VideoQ
- Adaptive Video Streaming and Event Summary
- Video Object Search by Motion Sketch

Video Skimming
- Important segments: video mode
- Non-important segments: still frame + audio + captions

Robust Content Feature
- PKI Encryption
- Digital Signature/ Watermark
- Content-Based Authentication/Watermarking

Camera/Transmission
- Image/Video

Image/Video
- Patient Record and Clinical Report
- Remote Medicine

Medical Video Indexing and Summarization

Echocardiogram Video
- Echo Video Acquisition
- Diagnosis/ Prognosis

Digital Echo Video Library
- Summary
- Views
- Links

Medical Video
- Indexing and Summarization
Focus of today’s talk
Video Adaptation in UMA

- Heterogeneous users, networks, and terminals -- one solution does not fit all
- *Content analysis* to assist video adaptation decision
Levels of Video Adaptation

- Semantic level
  - Event filtering – *show videos of highlight only*
  - Alert generation – *send alert video of abnormity immediately*

- Perceptual level
  - Transcoding in format, bit rate, frame rate, resolution, etc
  - Condense the video in time, size, or details
  - Modality conversion –
    - Key frames, slide shows, video posters, spatial summaries
  - Goal: maximize *perceptual quality*

- Rest of the talk
  - Techniques and examples in each level
Semantic-level Adaptation
Video Highlight Filtering

Interactive Event Browsing

- Highlights
- Pitches
- Runs
- By Player
- By Time

- Find semantic events in specific domains -- e.g., player/play/outcome in sports
- Match events to user preferences
- Save tremendous user time, bandwidth, and power
- Typical approaches for detection –
  - Detect fundamental syntactic units -- Scene composition model, and object tracking, spatio-temporal rules of objects
  - Fuse multi-modal metadata streams, e.g., VOCR, ASR, and Close Captions
A Simple Example: Use Regular Structures and Views

Production Syntax:
- canonical view \(\leftrightarrow\) recurrent semantic unit
- view transition pattern \(\leftrightarrow\) types of events

S.-F. Chang, Columbia U.
Detect canonical views using multi-level cues

(Zhong & Chang ’00)

- Easy to find discriminative features (e.g., color, motion, object, layout)
- Compressed-domain processing helps achieve real-time performance
- Multi-stage coarse-to-fine verification useful for enhancing accuracy

92%-98% detection accuracy for baseball/tennis
Fusing Multi-Stream Information - VOCR

Compressed-domain processing ➔ Real-Time

Explore domain-specific transition constraints

Compressed-domain processing (directional projection + unsupervised classification)

Transition constraint model

98% detection, 92% recognition (demo)

S.-F. Chang, Columbia U.
Demo

Sports Event Summary

- Random access to start of every play
- Random access to start of every score and other events

SCORING HIGHLIGHTS

<table>
<thead>
<tr>
<th>Innings</th>
<th>Score (ARI-NY) (SKIN)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 BOT</td>
<td>0-0 &gt; 0-1</td>
</tr>
<tr>
<td>3 BOT</td>
<td>0-1 &gt; 0-2</td>
</tr>
<tr>
<td>3 BOT</td>
<td>0-2 &gt; 0-3</td>
</tr>
</tbody>
</table>

This system uses single-modality analysis only!
A new concept of content-adaptive streaming

(Chang et al 2001)

- Send important segments at a bandwidth higher than the channel bandwidth
- Pay the price of buffer and latency

Important segments: video mode
Non-important segments: still frame + audio + captions

Event Detection
Structure Analysis
Resource Adaptation
Encode/Mux
Buffer

Live Video

Server or Gateway

Adaptive video rate
Channel rate
Time
Demo: View Detection and Adaptive Streaming

- Non-adaptive video @ 64Kbps
- Content-adaptive video @ 64Kbps
Playback latency is the main constraint
- cannot delay delivery of real-time event too long

Buffer size: not a constraint
- 8MB buffer can store 2000 sec (32Kbps) - 250 sec (256Kbps) of highlight

Client error more serious than server error
- How to handle client underflow error?
- Resume to normal quality, freeze and resume, or adaptive playback speed.
Multi-modal fusing is key to many tasks

- Example: TREC 2003 news story segmentation (120 hours from CNN/ABC)
- Detecting standard structures (anchor + news) is easy.

- But very often the structures are violated!

- Regular anchors may account for only 50-60% -- many exceptions
  - E.g., station logo, program preview, special effects, sports, interviews
- Every modality contributes, but when used alone, achieves insufficient accuracy

Exception example
A Clear Need of Multi-Modal Fusion (Hsu & Chang 03)

- No single modality is good enough!
- An ideal problem for statistical modeling and features combination

Multi-Modal observations, $x$

- {video, audio, VOCR, ASR}

Exponential posterior model

$$q(b \mid x) = \frac{1}{Z_{\lambda}(x)} e^{\sum_{i} \lambda_i f_i(x, b)} , b \in \{0, 1\}$$

Minimize divergence between training distribution and model

$$D(\tilde{p} \parallel q) = \sum_{x} \tilde{p}(x) \sum_{b} \tilde{p}(b \mid x) \log \frac{\tilde{p}(b \mid x)}{q(b \mid x)}$$

S.-F. Chang, Columbia U.
Results confirm multi-modal contributions

- (AV + speech) best, but AV alone better than audio alone
  - Prosodic cues, cue terms (Speech and VOCR), and visual all important
- Performance over TREC 2003 video (120 hours video)
  - 88% precision 68% recall for ABC, 83% precision 58% recall for CNN

The graph shows the log-likelihood after each feature election, with iterations ranging from 1 to 8. The lines represent different combinations:

- AV + Speech
- AV
- Speech
- S+C+A+V
- S+A+V
- A+V+C
- A+V
- S

The graph highlights the improvement in log-likelihood with each iteration.
Video Pattern Mining

- So far, we know what we want to detect.
- We train the model we choose.
- But ...
- How to deal with new domains, locations, collections?
Event mining in a rapidly deployed sensor networks

- **Goal:** automatic discovery of new events and patterns in *rapidly* deployed sensor networks

- **Issues:**
  - mining of events, spatio-temporal patterns
  - Normalcy definition, alert detection
  - distributed processing/communication
Challenge: Unsupervised Pattern Discovery

- Given a new domain/data, discover patterns automatically
  - E.g., Consumer, surveillance, and personal life log
- Technical Objectives:
  - Find appropriate spatio-temporal statistical models
  - Locate segments that match such models

Issues
- What’s the adequate class of models?
- How to determine model structures?
- What are “good” features?
In Selecting Models – Analyze Characteristics of Features & Dynamics

- Distinctive patterns are characterized by features and temporal transitions
- HMM has been used in many successful cases. Demo
Unsupervised Pattern Discovery using Hierarchical Hidden Markov Model

- Intuitive Representation for Videos
  - High-level states represent distinct events
  - Presence of each event produces observations modeled by low-level HMMs

Baseball Example

- Top-level states: running, pitching, break
- Bottom-level states: field view, 1st base, bench, close up, audience, pitcher, batter
Hierarchical HMM

Flexible Control Structure (Bottom-up control with exit state)
- Extensible to multiple levels and distributions
- Efficient inference technique available
  - Complexity $O(D \cdot T \cdot Q^\alpha D)$, $\alpha = 1.5$ to 2

[S. Fine, S. Singer, N. Tishby '98]
[K. Murphy, '01]
Hard Issues Emerge ...

- No knowledge about the model structure and complexity
- Perhaps no knowledge about adequate feature set
- No supervised labels available for checking feature correlation
- Use data-driven approach

--> find consistent & compact hypotheses \{(model, feature set)\}

1. Start with a large feature pool and a generic model
2. Partition features into groups that support consistent model structures and segmentation results
3. Within each feature-model pair, use MCMC stochastic method for structure perturbation and convergence
4. Bayesian quality criteria for ranking hypotheses
MCMC-Bayesian Adaptation → Finding the Right Model Structure

Initial HHMM

Random Proposal

EM
Split
Merge
Swap

(next iteration)

Accept proposal?

α = min{ 1, (e^{BIC ratio}) · (proposal ratio) · J } u ∈ U[0, 1]

u < α ?

Stop

Acceptance probability

\[ \alpha = \min\{1, \frac{P(x|\hat{m})}{P(x|m)} \cdot \frac{P(\hat{m})}{P(m)} \cdot \frac{P(\hat{m},\hat{u})}{P(m,u)} \cdot \frac{\partial(\hat{m},\hat{u})}{\partial(m,u)} \} \]
Feature Selection

generate feature seed

Feature pool
(1 color, 3 motion, 5 audio)

reference feature

EM+MCMC

wrapper around EM+MCMC

Information gain
\( M(p|p_0) \)

State partitions

 Relevant Feature sets

Markov blanket filtering

BIC-based model ranking

\[
BIC = \bar{L} \cdot \lambda - \frac{1}{2} |\Theta| \log(T)
\]

Remove Redundant Features

\( Q_{M_f \cup f} \approx Q_{M_f} \)

Multiple Feature-Model Pairs

S.-F. Chang, Columbia U.
Mining Patterns in Structured Video

9 av features, zero supervision

{dominant color ratio, horizontal motion}

{Volume, Spec-rolloff}

BIC score

Feature set

Learned Models

(82.3%)

(52.4%)

When compare with play/break labels

What do they code?

S.-F. Chang, Columbia U.
Promise and Open Issues

- Very encouraging results
  - Completely unsupervised discovery of patterns in (features + dynamics)
  - Perhaps we are lucky due to the highly constrained domain and production rules

- Open Issues
  - How to address granularity and sparseness of patterns?
  - How to evaluate the discovery results?
  - How to annotate discovered patterns?
  - How to fuse low-level features vs. mid-level objects (ball kick, cheers, fouls, etc)
  - How do we know whether/what we miss?
Part 2. Perceptual Level Adaptation
Perceptual-level adaptation

- Match videos to different resource conditions and user preferences, e.g., bandwidth, resolution, power, time
- A goal is to choose optimal operation to maximize perceptual quality
- Many dimensions of adaptation exist
- Video coding has successfully used Rate-Distortion theory – but not real-time also hard to go beyond signal-level distortion
- New Theme → content-based prediction of Utility Function
Utility Function defined based on Adaptation-Resource-Utility relations

Each point represents an adaptation operator

Required resources

Resulting utility

Utility Function: distribution of the op points in the R-U space

Each point represents an adaptation operator
Example:
Content-Based Object-Level Encoding/Transcoding

- Time-Lapsed Digital Video Recorder for large archive
  
  (A. Vetro et al, Mitsubishi, ICME 03)

- Encode foreground moving objects with higher temporal rate
  \( \rightarrow \) 80% bit rate reduction at comparable comprehension quality

- Potential Issues: object segmentation, background refresh rate, miss of important events

S.-F. Chang, Columbia U.
Content-based Coding Example

Courtesy of A. Vetro of MERL
Another Example: MPEG-4 spatio-temporal transcoding

- Tradeoff between spatial and temporal quality
  - FD: frame dropping. B or P frames in each GOP
  - CD: coefficient dropping in each frame using Lagrange optimization

- Other Examples:
  - MPEG-4 Fine Grained Scalability – trade-off spatial and temporal
  - 3D Wavelet Spatio-Temporal-Resolution Scalability
  - MPEG-4 Object Profile
The bitrate range and utility ranking of different operations vary with content types.
Hypothesize that distinctive UF classes exist and can be predicted by content features.

Utility Function Classes – customized for different codecs

Content Feature – codec independent;
UF Based Clustering

- Clustering to define UF classes
- SVM classification to map content features to UF class
- Local regression for predicting UF values
Content-Based Prediction Performance

Content-Based Prediction achieves significant gain in accuracy, especially at large bitrate reduction.

Open issues:
- Subjective utility measure is needed for spatio-temporal transcoding
- Extend UF to model other resources – e.g., power, CPU
Demo: Content-based UF Prediction

- A real-time visualization interface for studying the relations between video, UF, features, resource, and quality
Part 3: Video Skimming Based on Perceptual-level Analysis & Syntax
Scenarios for video skimming

active (e.g. search)

Shots, scenes & structural syntax

task + client

resources

UI

skim

cpu speed

User’s time

bandwidth

passive (e.g. PVR)

S.-F. Chang, Columbia U.
Video Skim Generation

Skim: Drastically condensed audio-video clips

1. What’s the right level of entity for manipulation – shot, syntax, scene?
2. Possible operations: dropping and trimming.
3. How will skimming affect audio-video relations?
4. How is the “quality” affected?
   Aesthetic affects, information comprehension

→ Need **Content-based Analysis**

[Sundaram, Chang, ’01 ’02]

Shot removal

dropped frames

Generalized utility framework for optimal video skimming

Film original

30% film Skim-

News original

17% news skim
Modeling Utility of Shots

- Thesis – skimming effect on quality depends on content
- Conduct subjective experiments to explore content-utility relationship
- Human subjects answer how much time is required for generic comprehension (who, what, where, when)?

Results suggest that visual complexity can approximately predict the required viewing time.

S.-F. Chang, Columbia U.
**Need other content information:**
syntactic structures and audio-visual interaction

- **Important Content Factors:**
  - ordering and structure of the shots
    (e.g., open-close, dialog, point of view, close up-long-close up)
  - Relative durations of the shots to direct viewer attention
    (e.g., long-short-short ...)
  - Audio-visual enforcement and synchronization is important in making emphasis
Content-based skim generation framework

Content analysis:
- video shot/syntax detection
- auditory analysis

Audio/video duration constraints
- target skim time
- AV tie constraints
- visual syntax constraints

Content-based video utility model
Content-based audio utility model

Objective function

Constraints

Iterative maximization

Skim generation
Subjective Quality Evaluation of CB Skims

- user study to validate the content-based video skims
  - 12 users
  - three skim generation mechanisms
  - three compression rates (90%, 80%, 50%)

- The user study indicates:
  - the optimal skim, has a superior raw score, in all cases.
  - the optimal skim is perceptually superior, in a statistically significant sense, at the high reduction rates.
5. Example Application: medical
Remote patients may not have access to clinical specialists

Lossy video compression and transmission may not be acceptable

Semantic/syntactic summary provides an effective solution.
Analyze spatio-temporal structures

- Deterministic patterns following AAC standard + statistical orders in actual on production → need statistical modeling and detection
- Not every view is needed
- Content-adaptive transmission → Transmit selective views/beats/frames only, details on demand
Echo Video Digital Library & Remote Medicine

Echo Video Acquisition

- Structure Parsing
- Domain Knowledge
- Video Clinic Summary
- View / event Recognizer
- Diagnosis Reports
- Multi-Modal linking
  - Adaptive transmission
  - Content search

S.-F. Chang, Columbia U.
DEVL Medical Echo Library Interfaces (demo)

Disease Taxonomy Interface

View Browsing Interface

Table of Contents showing list of views

Representative frames of modes under selected view

3D model showing transducer angle

3D Heart Model courtesy of New York University School of Medicine
Conclusions

Theme: Content-Aware Media Adaptation

- Content analysis has important impact on video adaptation applications
  - Ubiquitous Media, Remote Medicine, Distance Learning etc.
- Promising results shown
  - Domain-specific event detection and filtering
  - Real-time adaptive video streaming
  - Perceptual-level utility function prediction
  - Syntax preserving video skimming
- Open Issues
  - Automatic pattern discovery
  - Multi-modal fusion for complex events
  - Modeling of user preferences
Acknowledgements

- Unsupervised Video Mining:
  *Lexing Xie, Peng Xu, Ajay Divakaran, Huifang Sun*

- Utility Function Based Video Adaptation/Transcoding:
  *Yong Wang, Di Zhong, Raj Kumar, Jaegon Kim*

- Object-Based Video Coding
  A. Vetro, H. Sun, T. Hago, and K. Sumi of Mitsibushi Research

- Sports Event Filtering
  *DongQing Zhang*

- News Video Story Segmentation:
  *Winston Hsu*

- Syntax Preserving Video Skimming:
  *Hari Sundaram*

- Medical Video Indexing:
  *Shahram Ebadollahi, Henry Wu*

- 3D Heart Model courtesy of New York University School of Medicine
More Information

- Columbia DVMM Lab
  http://www.ee.columbia.edu/dvmm

- Prof. Shih-Fu Chang
  http://www.ee.columbia.edu/~sfchang

- Publications
  http://www.ee.columbia.edu/dvmm/publications.htm