Mining of Statistical Temporal Patterns in Video

Shih-Fu Chang

Digital Video and Multimedia Lab
ADVENT University-Industry Consortium
Columbia University
June 20th 2003

http://www.ee.columbia.edu/dvmm
email: sfchang@ee.columbia.edu

Joint Work

- with Lexing Xie, Ajay Divakaran, and Huifang Sun


Patterns Abound in Video

**Patterns exist at multiple scales**
(e.g., recurrent views, plays in sports)

View level
- Play level
- Break

More Video Pattern Examples

**News: Story**
- Filmm: Sporadic Patterns like Dialog, Point of View

S.-F. Chang, Columbia U. 06/03
Spatio-temporal patterns in distributed sensor networks

- In a distributed sensor network, patterns exist at
  - individual sensors: passing persons, footstep sounds
  - across sensors: “door opening likely followed by passing person in T seconds”
- Event breaking normal patterns → alerts

Patterns related to semantics

- Patterns correspond to semantics in specific domains

Multi-modal features + temporal transitions → Salient patterns → Semantic Events

domain
Patterns – conceptually defined

- Patterns are **recurrent** temporal segments in one or more sequences with predictable syntactic characteristics – MM feature and temporal transition.
  - Can be statistical or rule-based
  - Can be spare or dense, regular or irregular

- Other pattern-rich data:
  - Network traffics and transactions (for security, e-commerce)
  - Speech, language, and music
  - Biological Data (DNA or protein sequences)

Conventional Video Indexing Tasks

- **Supervised methods**
  - Expert-developer collaborative solutions
  - Given a domain
    - identify important events, concepts, and structures
    - Develop best detectors and classifiers
  - **Problems**
    - Lack of scalability
    - Cannot address data variations (e.g., personal collection)

- **Unsupervised clustering methods**
  - Explore feature similarity and time proximity of scenes
  - Rich temporal transitional characteristics unexplored
Challenge: Unsupervised Pattern Discovery

- Given a new domain/data, discover patterns automatically
  - E.g., Consumer, surveillance, and personal media 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 complexity
- What’s a “good” feature set

Simple Observations of Patterns

Distinctive patterns are characterized by feature distribution and temporal transitions
Unsurprising evidence for Using HMM

- Prior works: HMM for temporal structures in video (Huang, Liu, & Wang news, Wolf Basketball, Liu & Kender documentary, Naphade & Huang films)
- Supervised learning of model parameters and high-level transitions
- ML classification of new videos plus temporal smoothing

(labeled & chunked Plays & Breaks)

Train HMM families For plays/breaks

Estimate high-level transition prob.

Play-Break sequence temporal smoothing (Dynamic Programming)

ML Classification + Model Refinement

Test Data Segment

Test set

<table>
<thead>
<tr>
<th>Test set</th>
<th>Argentina</th>
<th>KoreaA</th>
<th>KoreaB</th>
<th>Espana</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argentina</td>
<td>87.2%</td>
<td>82.5%</td>
<td>82.5%</td>
<td>80.6%</td>
</tr>
<tr>
<td>KoreaA</td>
<td>78.1%</td>
<td>84.3%</td>
<td>84.3%</td>
<td>79.8%</td>
</tr>
<tr>
<td>KoreaB</td>
<td>79.9%</td>
<td>85.3%</td>
<td>85.3%</td>
<td>89.6%</td>
</tr>
<tr>
<td>Espana</td>
<td>79.9%</td>
<td>89.6%</td>
<td>89.6%</td>
<td>81.7%</td>
</tr>
</tbody>
</table>

HMM effective for parsing soccer structures

[Xie, Chang, Divakaren, Sun 02]

- 4 test clips, 15~25 min., various countries
- Cross Validation

• Avg. Play-Break Classification Rate: 83.5% vs. 60% of blind classification
• Boundary timing accuracy: 62% within 3 seconds
• The good classification provides preliminary evidence supporting HMM model and the selected features
Generalize to Unsupervised Discovery: Hierarchical Hidden Markov Model

- HHMM successfully used in tracking and recognition
- Intuitive Representation for Videos
  - High-level states represent distinct events
  - Presence of each event produces observations modeled by low-level HMMs

**Baseball Example**

Hierarchical HMM

- Flexible Control Structure (Bottom-up control with exit state)
- Extensible to multiple levels and distributions
- Applications to video event discovery [Clarkson & Pentland '99, Naphade & Huang '02]
  - left-right models with fixed model and features
- Given HHMM, efficient inference technique available
  - Complexity O(D·T·Q^{α}), α=1.5 to 2 [Murphy '01, Xie et al '02]

- [Fine, Singer, Tishby '98]
- [K. Murphy, '01]
But Still Some Critical Issues

- No knowledge about the model structure and complexity
- No knowledge about optimal feature set
- data-driven approach
  1. MCMC random model search for model adaptation
  2. Hybrid wrapper/filtering and Bayesian criteria for feature selection

MCMC-Bayesian Adaptation → Finding the Right Model Complexity

![Diagram showing the process of MCMC-Bayesian Adaptation with initial HHMM proposal, random proposal, EM, Split, Merge, Swap, and acceptance probability equation.](image)
Feature Selection [Xie, Chang, Divakaren, Sun ICIP 03]

- Feature pool (1 color, 3 motion, 5 audio)
- Multiple Feature-Model Pairs
- BIC-based model ranking
- $BIC = \bar{L} \cdot \lambda - \frac{3}{2} |\Theta| \log(T)$
- Remove Redundant Features
- Markov blanket filtering
- Relevant Feature sets

Information gain $M(p|p_0)$

Wrapper around EM+MCMC

EM+MCMC

Generate feature seed

Test Case: Baseball

- Model learning + feature selection
- BIC score
- Feature set
- HHMM
- Correspondence (with play/break)

(dominate color ratio, horizontal motion)

(Volume, Spec-rolloff)

82.3%

52.4%
Comparing with supervised methods

Unsupervised discovery of models and features

<table>
<thead>
<tr>
<th>Test clip</th>
<th>Feature Set</th>
<th># states</th>
<th>Correspondence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Korea</td>
<td>DCR, Mx</td>
<td>2~4</td>
<td>75.2%</td>
</tr>
<tr>
<td>Spain</td>
<td>DCR, Volume</td>
<td>2~3</td>
<td>74.8%</td>
</tr>
<tr>
<td>Baseball</td>
<td>DCR, Mx</td>
<td>2</td>
<td>82.3%</td>
</tr>
</tbody>
</table>

Comparison with supervised approaches (soccer)

<table>
<thead>
<tr>
<th>Model</th>
<th>Supervised?</th>
<th>Adaptation?</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>HHMM</td>
<td>N</td>
<td>Y</td>
<td>75.2%</td>
</tr>
<tr>
<td>HHMM</td>
<td>Y</td>
<td>N</td>
<td>75.0%</td>
</tr>
<tr>
<td>HMM**</td>
<td>Y</td>
<td>N</td>
<td>75.5%</td>
</tr>
</tbody>
</table>

* Korean soccer, MPEG-7 soccer dataset, 25 minutes long
** Basic feature set: dominant color ratio (DCR), motion intensity (MI)
** Trained HMM family + post smoothing with dynamic programming

- Automatic approaches find meaningful clusters and features
- Unsupervised solutions achieve comparable performance in play/break classification

Open Issues

- Evaluation
  - Is classification based on manually selected semantic classes adequate?
    - No predefined semantic targets used in mining
  - Miss/False rates when compared to salient patterns identified manually
- How to annotate discovered patterns?
- How to address granularity and sparseness of patterns?
- Low-level features vs. mid-level features (e.g., audio class, object detector)
How to assign meanings to discovered patterns?

- Idea: learning statistical correspondence between patterns and synchronized text or metadata streams

Conclusions

- The spatio-temporal dimensions in videos offer rich patterns to be mined
- Video pattern mining facilitates scalability and personalization
- Data-driven approaches using models like HHMM and feature selection methods show interesting results
- Many issues remain
  - evaluation, semantic tagging, granularity, fusing with domain knowledge
More Information

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

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

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