ISIMP 2004

Video Pattern Mining

Prof. Shih-Fu Chang

Digital Video and Multimedia Lab
Columbia University
October, 2004
http://www.ee.columbia.edu/dvmm

Joint work
- Lexing Xie, Lyndon Kennedy, Winston Hsu, Dong-Qing Zhang (Columbia U.)
- Ajay Divakaren and Huifan Sun (MERL)
- John R. Smith and Ching-Yung Lin (IBM)

Supported in part by
- MERL
- ARDA VACE phase II
- IBM-Columbia Semantrix Project
Temporal Patterns Everywhere ...

- There are interesting patterns representing useful information in different domains.

Example Patterns in Video

- financial news, CNN
- anchor interview text/graphics footage ...
- 98-05-20
- 98-06-02
- 98-06-07
- soccer video
- play start pass interception attempts attempt at the goal break
- baseball
- time

S.-F. Chang, Columbia U.
Patterns Across News Sources

- Story/event often re-occur within or across channels
- Related stories share intrinsic audio-visual patterns -- forming news thread
- Tendency how a story is repeated at different times and by different channels -- forming higher-level patterns

Types of Temporal Patterns

- Dense and stochastic
- Sparse and deterministic
- Temporal association of events
  - If A occurs in channel 1, then B occurs within time T in channel 2, e.g. recurrent news topics
  - Others ...
(Grand) Challenges for Pattern Mining

- There are many patterns in video at different levels.
  - How do we discover them (semi)-automatically?
  - Do they correspond to any semantic meanings?
  - What are the underlying features/structures of each pattern?
  - Which patterns are more predictable and thus detectable?
- Why do these matter?
  - Explore structures and knowledge in a domain without a priori manual definition
  - → discover interesting, meaningful concepts & events
  - → define normal states and outliers
  - Facilitate scalability to new content and domains

Challenge: Unsupervised Pattern Discovery

- Given a new domain/corpus, discover patterns automatically
  - E.g., News, consumer, surveillance, and personal life log
- Technical Issues:
  - Find appropriate spatio-temporal statistical models to represent and explain patterns
  - Find locations in the video stream that fit such models
- Issues
  - What’s the adequate class of models?
  - How to learn suitable model structures?
  - How to select good features?
Finding the right class of models ...
- Lessons from prior work
  supervised + unsupervised

Many video temporal patterns are characterized by
Dynamics and Features

- Distinctive patterns are characterized by state-dependent transitions and features.
- Like speech recognition, HMM and variants may serve as promising candidates.
A simple test: HMM Parsing of Soccer Video

(training phase)

Labeled Segments (color, motion, audio, objects) → Train HMM families for soccer plays/breaks → learn high-level transition prob. (training phase)

(testing phase)

Test Data Feature Sequence → Likelihood Estimation for each Model → High-Level Play-Break Optimal sequence (Dynamic Programming)

Supervised HMM models are effective for parsing structural patterns in sports video

Sports video from various countries
Evaluation by Cross Validation

<table>
<thead>
<tr>
<th>Test set</th>
<th>Training Set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Argentina</td>
</tr>
<tr>
<td>Argentina</td>
<td>87.2%</td>
</tr>
<tr>
<td>KoreaA</td>
<td>78.1%</td>
</tr>
<tr>
<td>KoreaB</td>
<td>79.9%</td>
</tr>
<tr>
<td>Espana</td>
<td>79.9%</td>
</tr>
</tbody>
</table>

- The good classification provides preliminary evidence supporting HMM model and the A-V features
- Avg. Play-Break Classification Rate: 83.5% vs. 60% of blind guessing
- Boundary timing accuracy: 62% within 3 seconds
Temporal Pattern Mining with a Single Model: Hierarchical HMM

- Intuitive Representation for Video Patterns
  - Patterns occur at different levels following different transition models
  - States in each level may correspond to different semantic concepts

Applying HMM to unsupervised mining

- A simplified structure without higher level state transition
  - [Clarkson et al '99]
  - Long personal videos captured by wearable device
  - Color histogram and MFCC features at 10Hz
  - discovered pattern sequence has high correlation with ground truth events

- [Naphade et al '02]
  - Films and talk shows
  - Color and edge histogram, MFCC and energy features at 30Hz
  - discover recurrent patterns of explosion and applause
Hierarchical HMM

- Flexible control structure (bottom-up control with exit state)
- Extensible to multiple levels and distributions
- Efficient inference technique available
  - Complexity $O(D \cdot T \cdot Q^{\alpha})$, $\alpha = 1.5$ to $2$
- Application in unsupervised discovery has not been explored
- Questions: how to find right model structures and feature sets?

The Need for Model Selection

- Different domains have different descriptive complexities.
Model Selection with RJ-MCMC

Possible Model Operations:
- EM
- Split
- Merge
- Swap

Operations:
- (move, state) = (split, 2-2)

Optimum points:
- balance (data fitness) + (model complexity)

Stop criteria:
- Prob. Thresh. for accepting change = \( (\text{eBIC ratio}) \times \text{(proposal ratio)} \times J \)

Which Features Shall We Use?

Features:
- color histogram
- edge histogram
- MFCC
- zero-crossing rate
- delta energy
- Spectral rolloff
- pitch
- LPC coeff.
- delta energy
- logtf-entropy
- keywords
- tf-idf
- face?
- outdoors?
- people?
- vehicle?
- motion estimates
- Gabor wavelet descriptors
- zernike moments
- outdoors?
- vehicle?
- text

S.-F. Chang, Columbia U.
Issues of Feature Selection

Standard process for Supervised learning:
Identify a good subset of observations to improve model
generalization performance and reduce computation.

Criteria: (1) find the feature-feature or feature-concept relevance
(2) eliminate any redundancy

But new problems for temporal sequence mining:
- Unsupervised Temporal sequence
- No target concept defined a priori
- Temporal samples not independent

A Divide and Conquer Approach:
Feature Selection for Temporal Pattern Mining

Feature pool Multiple consistent feature sets Ranked feature sets with redundancy eliminated

Feature sequences wrapper filter

State sequences

Mutual information

| q1="abaaabbb" |
| q2="BABBBAAA" |
| I(q1,q2)=1 |

Markov Blanket

\[
X' \perp Q | X_1
\]

\[
Q \Rightarrow \{X', X_2\}
\Rightarrow q1="abaaabbb"
\Rightarrow q1'="abaaabbb"= q1
\Rightarrow \text{Eliminate } X'
\]
Results: on Sports Videos

- videos: baseball, soccer
- features: visual: dominant color ratio, camera translation motion; audio: energies, zero-crossing rate, spectral rolloff
- HHMM: top-level label sequence
- vs. play/break?

A Simple Test: Mining Baseball Videos

- videos: baseball
- HHMM + feature selection:
  1. dominant color ratio, horizontal motion (vertical motion)
  2. audio volume, low-band energy
  3. ...
- Ranking based on BIC score
- Semantics of patterns?
  Correspond to play/break with 82.3% accuracy
  (demo)
Unsupervised Mining not less Effective than Supervised Learning

Fixed features \{DCR, MI\}, MPEG-7 Korean Soccer video

<table>
<thead>
<tr>
<th>Model</th>
<th>Supervised?</th>
<th>Model Selection</th>
<th>Correspondence w. Play/Break</th>
</tr>
</thead>
<tbody>
<tr>
<td>HHMM</td>
<td>N</td>
<td>Y</td>
<td>75.2 ± 1.3%</td>
</tr>
<tr>
<td>HHMM</td>
<td>N</td>
<td>N</td>
<td>75.0 ± 1.2%</td>
</tr>
<tr>
<td>HMM</td>
<td>Y</td>
<td>N</td>
<td>75.5 ± 1.8%</td>
</tr>
<tr>
<td>LR-HHMM</td>
<td>N</td>
<td>N</td>
<td>73.1 ± 1.1%</td>
</tr>
<tr>
<td>K-Means</td>
<td>N</td>
<td>N</td>
<td>64.0 ± 10.0%</td>
</tr>
</tbody>
</table>

Automatic selection of both model and features

<table>
<thead>
<tr>
<th>Test clip</th>
<th>Feature Set</th>
<th># “events”</th>
<th>Correspondence w. Play/Break</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>

* DCR=“dominant-color-ratio’, MI=’motion-intensity’, Mx=’horizontal-camera-pan’

Other examples of video pattern discovery (e.g., Jojic and Frey)

- Similar principle in using video generative models
- Assume latent spatial-temporal models and transformations
- AV features are observations of the generation process
- Unsupervised inferencing

Video mining methods seem promising in finding hidden patterns in video.

But how to find the meanings of the patterns?

Approach

- fuse the metadata streams when available.
- Multi-modal fusion

Outline

- The problem
- Unsupervised pattern discovery with HHMM
- Finding meaningful patterns
  - With text association
  - By multi-modal fusion
- Summary
Towards Meaningful Patterns

- What are the meanings of the hidden patterns?
- Manual association feasible only if meanings are few and known.
- Metadata come to the rescue.

Associating Patterns with Text

- videos news
- AV features and concepts
- HHMM
- Patterns/tokens
- label-word association

S.-F. Chang, Columbia U.
Co-Occurrence of HHMM Labels & Words

**Conditional prob.**

\[
C(q \mid w) = \frac{C(q, w)}{C(q, \cdot)}
\]

\[
C(w \mid q) = \frac{C(q, w)}{C(\cdot, w)}
\]

**Likelihood ratio:**

\[
L'(q, w) = \frac{C(q, w)}{C(q, \cdot)C(\cdot, w)}
\]

"correlation" between HHMM labels and words
→ co-occurrence counts.

Problems with the Co-occurrence Statistics

<table>
<thead>
<tr>
<th>Story#</th>
<th>News Video</th>
<th>HHMM label</th>
<th>ASR token</th>
<th>&quot;true mapping prob.&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>q₁</td>
<td>w₁</td>
<td>2 0</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>q₂</td>
<td>w₂</td>
<td>1 2</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>q₂</td>
<td>w₁</td>
<td>2 1</td>
</tr>
</tbody>
</table>

Machine Translation

[Brown’93]

Son chien dactylographie sur mon ordinateur.

Her dog is typing on my computer.

f ∈‘computer’

dactylograph* mon/ma

ordinateur son/sa

sur Co-occurrence True mapping

Applied to Image Annotation

[Dyugulu et. al. 2002]

image ~ blobs \{b₁, ..., bₙ\}

~ words \{w₁, ..., wₙ\}
Inferring the translation probability between visual tokens and words

\[ \text{find optimal } \theta^* = \argmax \log p(w \mid b, \theta) \]

\[ \log p(w \mid b, \theta) = \sum_{n=1}^{N} \sum_{j=1}^{M_j} \sum_{i=1}^{L_n} p(a_{nj} = i) t(w \mid w_{nj}, b_{nj}, \theta) \]

**Auxiliary Function for E-M**

\[ Q(\theta, \theta^{\text{old}}) = \sum_{n=1}^{N} \sum_{j=1}^{M_j} \sum_{i=1}^{L_n} p(a_{nj} = i \mid w_{nj}, b_{nj}, \theta^{\text{old}}) \log[p(a_{nj} = i \mid w = w_{nj}, b = b_{nj})] \]

**Prob. of the hidden alignments**

**Sum over the true joint prob.**

---

Translation between AV Tokens & Words

The problem: Co-occurrence "un-smoothing".

know: \( C(q, w) \);

seek: \( t(w \mid q), t(q \mid w) \)

Solve with EM [Brown'93]

\[ L_n^q(q, w^c) \] (assume q's are ind.)

\[ L_n^w(q, w^c) \] (assume w's are ind.)

\[ \gamma = \frac{C(q, w)}{C(q \cdot C'(-1), w)} \]

Mappings become much more selective
Experiments

- TRECVID2003 news
  - 44 half-hour videos, ABC/CNN
  - Raw audio-visual features: color, motion, speech
  - 12 semantic concepts for each shot [IBM-TREC'03]
    (weather, people, sports, non-studio, nature-vegetation, outdoors, news-subject-face, female speech, airplane, vehicle, building, road)
  - ASR transcript
- HHMM on concept confidence scores
  - Construct 10 HHMM models with automatically determined size and features
  - Machine translation methods to find association between HHMM patterns and words

Example Correspondences

<table>
<thead>
<tr>
<th>HHMM label</th>
<th>Visual Concept</th>
<th>Translated Words</th>
<th>Manual Inspection</th>
</tr>
</thead>
<tbody>
<tr>
<td>(6,3)</td>
<td>people, non-studio-setting</td>
<td>storm, rain, forecast, flood, coastal, el, nino, administer, water, cost, weather, protect, Starr, north, plane, …</td>
<td>El-nino Storm '98 (recall 80%)</td>
</tr>
<tr>
<td>(9,1)</td>
<td>indoor, news-subject-face, building</td>
<td>murder, Lewinski, congress, allege, jury, judge, clinton, preside, politics, saddam, lawyer, accuse, independent, monica, charge, …</td>
<td>Clinton-Jones (Recall 45%, Precision 15%)</td>
</tr>
<tr>
<td>(m, q): model # m state # q</td>
<td>Obtained with SVM classifiers [IBM'03]</td>
<td>Lexicon obtained by shallow parsing of keywords from speech recognition output.</td>
<td>Manual Inspection</td>
</tr>
</tbody>
</table>

[Xie et al. ICIP’04]

[IBM'03] (m, q): model # m state # q S.F. Chang
Can’t we find such patterns using conventional clustering? (HHMM vs. K-means comparison)

HHMM provides unambiguous association between discovered patterns and words

Outline

- The problem
- Unsupervised pattern discovery with HHMM
- Finding meaningful patterns
  - With text association
  - By multi-modal fusion
- Summary
Is MT the right paradigm?

Potential Problems:
- Words ≠ topical meanings
- Temporal correspondence between words and pattern tokens may not exist.
- Audio-visual features and Text are processed separately and then fused at a late stage.

Change to Joint AVT Pattern Mining

- Instead, both the AV pattern tokens and words should be used to jointly define hidden semantic states.
- A multi-modal data fusion problem [AVSR, multi-modal interaction]
- Aiming at recovering the semantics [TDT 1998-2004]
Layer Dynamic Mixture Model for Produced Videos

Extracting High-level Concepts from Text – pLSA

- Use data-driven analysis to find concept association and latent semantic topics
- Use the syntactic story structures of video to define co-occurrence
- Use graphics model statistical inferencing to discover latent semantics
  - Not just counting occurrences
Some semantic clusters from text pLSA

<table>
<thead>
<tr>
<th>&quot;financial&quot;</th>
<th>&quot;iraq&quot;</th>
<th>&quot;weather&quot;</th>
<th>&quot;bad&quot; clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>dow</td>
<td>saddam</td>
<td>temperature</td>
<td>cancer</td>
</tr>
<tr>
<td>nasdaq</td>
<td>iraq</td>
<td>rain</td>
<td>increase</td>
</tr>
<tr>
<td>industrial</td>
<td>baghdad</td>
<td>coast</td>
<td>secure</td>
</tr>
<tr>
<td>average</td>
<td>weapon</td>
<td>snow</td>
<td>temperature</td>
</tr>
<tr>
<td>wall</td>
<td>hussein</td>
<td>el</td>
<td>texas</td>
</tr>
<tr>
<td>jones</td>
<td>strike</td>
<td>heavy</td>
<td>accusation</td>
</tr>
<tr>
<td>gain</td>
<td>secure</td>
<td>northern</td>
<td>chance</td>
</tr>
<tr>
<td>trade</td>
<td></td>
<td>storm</td>
<td>nasdaq</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td>forecast</td>
<td>pressure</td>
</tr>
<tr>
<td>&quot;olympics&quot;</td>
<td>&quot;investigation&quot;</td>
<td>tornado</td>
<td>center</td>
</tr>
<tr>
<td>gold</td>
<td>jury</td>
<td>pressure</td>
<td>...</td>
</tr>
<tr>
<td>olympics</td>
<td>lewinski</td>
<td>east</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>starr</td>
<td>florida</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>grand</td>
<td>nino</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>accusation</td>
<td>gulf</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>sexual</td>
<td>weather</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>independent</td>
<td></td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>water</td>
<td></td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>monica</td>
<td></td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>investigation</td>
<td></td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>president</td>
<td></td>
<td>...</td>
</tr>
</tbody>
</table>

Layered Dynamic Mixture Model (LDMM)

high-level clusters Z
mid-level labels Y
observations X

What do the highest-level patterns represent?
Conjecture: News Topics with salient multimedia and temporal cues!

Inference: EM

[S-Xie’TR04]
[Oliver’02]
[Stork’96]
[Nefian’02]
Question: does the discovered thread correspond to distinct semantics?
- no ground truth
- tentatively use NIST TDT topics

Fusion Results Evaluated by Text TDT Topics & Metrics

http://www.nist.gov/speech/tests/tdt/

- "NIST TDT research develops algorithms for discovering and threading together topically related material in streams of data such as newswire and broadcast news in both English and Mandarin Chinese."

- Current TDT use text or ASR only.

TDT Tasks
1. **Story Segmentation** - Detect changes between topically cohesive sections
2. **Topic Tracking** - Keep track of stories similar to a set of example stories
3. **Topic Detection** - Build clusters of stories that discuss the same topic
4. **First Story Detection** - Detect if a story is the first story of a new, unknown topic
5. **Link Detection** - Detect whether or not two stories are topically linked
Topic Detection and Tracking


- The tracking of known topics -- classification;
- The detection of unknown topics -- clustering;
- The detection of pairs of stories on the same topic (links).

```
S.-F. Chang

# of documents in the TDT text corpus over different topics

# of stories in trecvid2003 (151 videos)

Videos have different topical distribution from text documents
```

S.-F. Chang, Columbia U. 23
523 video stories in 30 frequent topics

<table>
<thead>
<tr>
<th>Topic #</th>
<th>#stories</th>
<th>topic title</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30</td>
<td>Asian Economic Crisis</td>
</tr>
<tr>
<td>2</td>
<td>87</td>
<td>Monica Lewinsky Case</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>Casey Martin Sues PGA</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>Pope visits Cuba</td>
</tr>
<tr>
<td>13</td>
<td>16</td>
<td>1998 Winter Olympics</td>
</tr>
<tr>
<td>15</td>
<td>67</td>
<td>Current Conflict with Iraq</td>
</tr>
<tr>
<td>18</td>
<td>12</td>
<td>Bombing AL Clinic</td>
</tr>
<tr>
<td>19</td>
<td>9</td>
<td>Cable Car Crash</td>
</tr>
<tr>
<td>20</td>
<td>4</td>
<td>China Airlines Crash</td>
</tr>
<tr>
<td>21</td>
<td>9</td>
<td>Tornado in Florida</td>
</tr>
<tr>
<td>31</td>
<td>5</td>
<td>John Glenn</td>
</tr>
<tr>
<td>32</td>
<td>13</td>
<td>Sgt. Gene McKinney</td>
</tr>
<tr>
<td>41</td>
<td>7</td>
<td>Grossberg baby murder</td>
</tr>
<tr>
<td>43</td>
<td>6</td>
<td>Dr. Spock Dies</td>
</tr>
<tr>
<td>44</td>
<td>38</td>
<td>National Tobacco Settlement</td>
</tr>
<tr>
<td>47</td>
<td>28</td>
<td>Viagra Approval</td>
</tr>
<tr>
<td>48</td>
<td>20</td>
<td>Jonesboro shooting</td>
</tr>
<tr>
<td>56</td>
<td>7</td>
<td>James Earl Ray's Retrial?</td>
</tr>
<tr>
<td>65</td>
<td>15</td>
<td>Rats in Space!</td>
</tr>
<tr>
<td>70</td>
<td>39</td>
<td>&quot;India, A Nuclear Power?&quot;</td>
</tr>
<tr>
<td>71</td>
<td>24</td>
<td>Israeli-Palestinian Talks (London)</td>
</tr>
<tr>
<td>76</td>
<td>20</td>
<td>Anti-Suharto Violence</td>
</tr>
<tr>
<td>77</td>
<td>7</td>
<td>Unabomber</td>
</tr>
<tr>
<td>83</td>
<td>4</td>
<td>World AIDS Conference</td>
</tr>
<tr>
<td>84</td>
<td>6</td>
<td>Job incentives</td>
</tr>
<tr>
<td>86</td>
<td>24</td>
<td>GM Strike</td>
</tr>
<tr>
<td>87</td>
<td>18</td>
<td>NBA finals</td>
</tr>
<tr>
<td>89</td>
<td>4</td>
<td>Afghan Earthquake</td>
</tr>
<tr>
<td>91</td>
<td>9</td>
<td>German Train derails</td>
</tr>
<tr>
<td>96</td>
<td>12</td>
<td>Clinton-Jiang Debate</td>
</tr>
</tbody>
</table>

Out of 151 videos (V, F1, F2, Test), 4241 segments, 3066 stories.

Experiments

- **News videos [TRECVID2003]**
  - ABC, CNN : 30 min x 151 video programs
  - Training/testing on same channels

<table>
<thead>
<tr>
<th>modality</th>
<th>feature elements</th>
<th>granularity</th>
<th>bottom-level model</th>
</tr>
</thead>
<tbody>
<tr>
<td>audio</td>
<td>pitch, pause, audio-class</td>
<td>.5 sec</td>
<td>HHMM</td>
</tr>
<tr>
<td>color</td>
<td>histogram (15-d)</td>
<td>1 sec</td>
<td></td>
</tr>
<tr>
<td>motion</td>
<td>camera translation (2-d)</td>
<td>1 sec</td>
<td></td>
</tr>
<tr>
<td>visual</td>
<td>22 concepts</td>
<td>every shot</td>
<td></td>
</tr>
<tr>
<td>text</td>
<td>ASR word-stems tf-idf</td>
<td>every story</td>
<td>PLSA</td>
</tr>
</tbody>
</table>

S.-F. Chang, Columbia U.
Evaluate the MM patterns against TDT topics

- Auto-story boundary, 32 clusters
- Minimum $C_{det}$ among the CNN videos for each topic

- Multi-modal fusion able to detect certain topics more accurately than text alone
- Such topics tend to be rich in non-textual cues

Example: Cluster #28

- Correspond to “Dollars & Sense, CNN” (demo)

Unique audio-visual + text terms + temporal transition
- AV: graphics, music, anchor speech, subject etc
- Textual terms: most related to financial
- Temporal structure: statistically predictable transition patterns
Example: Sports

Consistent audio-visual (color, motion, graphics), words, and temporal transitions.

Example: Advertisements

- Audio-visual dominate in this thread.
- Text not useful
Application in Multi-Source News Threading

Joint Work with IBM T.J. Watson Research
Supported by ARDA VACE II program

Reconstructing Semantic Threads Across Multiple Video Broadcast News Sources Using Multi-Level Concept Modeling

**Objectives**
- Extraction of semantic threads through multi-level video content analysis (core- and semantic-level features, scene markers & similarities, commonality of content and structure)
- Establishment of viability of automatic concept detection as basis for video semantics understanding and higher-level analysis functions and user interactions
- Development of framework and technologies supporting extraction of thousands of concepts from news video

**Novel Approach**
- Hierarchical segmentation and threading at multiple granularities (key-frames, shots, scenes, episodes and stories)
- Dynamic probabilistic graphical modeling for detection of concepts to form “semantic basis” of objects, sites, actions
- Statistical modeling & mining of spatio-temporal & semantic structures across multiple broadcast video news sources

**PIs:** IBM (John R. Smith); Columbia Univ. (Prof. Shih-Fu Chang)
Multi-level News Video Content Analysis

Multi-level feature extraction and systematic development of higher-level access functions

PIs: IBM (John R. Smith); Columbia Univ. (Prof. Shih-Fu Chang)

Detection of Image Near Duplicate (IND)

With Dongqing Zhang
How to Link Video Threads with external Sources (e.g., web pages)?

**News Video Sources**

- Search by example
- Threading News Stories

<table>
<thead>
<tr>
<th>Source #1</th>
<th>Source #2</th>
<th>Source #3</th>
</tr>
</thead>
<tbody>
<tr>
<td>News Web Site 1</td>
<td>News Web Site 2</td>
<td></td>
</tr>
</tbody>
</table>

**Image Near-Duplicate Detection**

**Image Near-Duplicate (IND):** A pair of images in which one is close to the exact duplicate of the other, but differs slightly due to variations of content and camera parameters.

**Application of IND Detection:**

Threading news videos across sources.
Part-based Approach for IND Detection

- Visual scene is composed of objects (parts) with spatial/attribute relations
- Part-based representation by Attributed Relational Graph (ARG)

Part representation: we use interest point based representation
Part detection: SUSAN corner detector – detecting corners using local geometries, such as curvature. Efficient algorithm compared with others, such as Harris

Features
- Part: Color, spatial location, Gabor wavelet coefficients
- Inter-part relation: Difference of spatial locations

IND Detection becomes ARG similarity problem

Graphical Model for ARG similarity

ARG similarity is the likelihood ratio of the stochastic transformation process

$\text{ARG similarity} = \frac{p(Y' \mid \theta', H = 1)}{p(Y' \mid \theta', H = 0)}$
Inference and Learning

Likelihood computation

\[ p(Y^t | Y^s, H) = \sum_{X \in \mathcal{X}} p(Y^t | Y^s, X, H) p(X | Y^s, H) \]

Computing exact likelihood is intractable, so approximate it using Bethe approximation, leading to Loopy Belief Propagation in the product graph.

Learning

- Learning using node-level annotation
- Learning using image-level annotation

Learning is realized by Variational E-M

\[ \text{Learning} = \text{Parameter Computation} \]

Inference results (after thresholding beliefs)

\[ \text{Inference} \rightarrow \text{Parameter Estimation} \]

Statistical Approach to Graph Matching

- Stochastic Graph Editing Process
- IND vs. non-IND Likelihood Ratio → a new similarity measure

Similarity by Likelihood Ratio

\[ \text{Similarity} = \frac{p(\text{Graph}_t \mid \text{Graph}_s, \text{Two_graph_is_NID})}{p(\text{Graph}_t \mid \text{Graph}_s, \text{Two_graph_is_not_NID})} \]

- Compute Likelihood
  \[ p(Y^t | Y^s, H) = \sum_{x \in \mathcal{X}} p(Y^t | Y^s, x, H) p(x | Y^s, H) \quad x \in \{0, 1\}^{y \times M} \]

Intractable! So approximate it by using Jensen’s lower bound

\[ \ln p(Y^t | Y^s, H = h) \geq \sum_{\tilde{Y} \in \mathcal{Y}_s} \psi_{\tilde{Y}^s} \left( \psi_{\tilde{Y}^t} + \psi_{\tilde{Y}^s} \right) + \text{constant} \]

1. Inference: Compute the approximate distribution \( \hat{q} \) by Loopy Belief Propagation

2. Learning: Estimate \( \psi_{\tilde{Y}^t} \) and \( \phi_i \) by using E-M

S.-F. Chang, Columbia U.
IND Detection Performance (demo)

Statistics of Image Near-Duplicate (IND) in TREC News Video Database

Dataset for IND detection
- 150 IND pairs, 300 non-duplicate images
- No multiple duplicate

Training set
- 30 IND pairs, 60 non-duplicate images
- Node-level learning: 5 IND pairs, 5 non-duplicate images

Detection Results

Retrieval Results

Conclusions

Theme: Media Pattern Mining for Semantics Discovery

- Patterns abound in multimedia data
- Mining facilitates auto discovery of salient or novel concepts → scalable automatic knowledge discovery
- Current results seen in
  - Multi-level temporal patterns mining through HHMM
  - Video generation models
  - Fusion between pattern labels and ASR metadata
- Challenging Issues
  - Mining of patterns of different types or complex patterns at higher levels
  - Detection of alerts and novel events
  - Evaluation and Visualization
Conclusions (2)

Threading News Videos Across Sources Using Multi-modal Cues

- Multi-source news video analysis and content exploitation offers technical challenges and application opportunities
  - Multimedia Video Mining
  - Cross-source linking using image near duplicate
  - Multi-modal fusion for story segmentation
  - Current work:
    - Large scale AV concept detection
    - Story topical threading
- Closely related to evaluation tasks in TRECVID
- A workshop for developing large-scale concept ontology (LSCOM)

More Information

- Digital Video| Multimedia Lab
  http://www.ee.columbia.edu/dvmm
- Papers