Multi-Cue Fusion for Semantic Video Indexing

Ming-Fang Weng and Yung-Yu Chuang
National Taiwan University
mfueng@cmlab.csie.ntu.edu.tw

Motivation

• A rapidly growing amount of videos derives a need in semantic video search
Semantic Concepts

• Comprehensively characterize the meaning of the video content, e.g.,

Goal

• To **improve the accuracy** of semantic video indexing

A ranking list according to confidence measure
A Typical Approach

• Supervised learning

Feature Extraction → Training Concept Classifiers → Semantic Concept Prediction

Video Segmentation → Lexicon Annotation

Video Archive

Training path: —— Testing path: ——

Main problems:

• The annotation data is not fully utilized
• The label for all concepts in all shots are predicted independently
**Ground Truth Annotation**

A sequence of video shots (training set)

<table>
<thead>
<tr>
<th>Concept</th>
<th>car</th>
<th>outdoor</th>
<th>urban</th>
<th>building</th>
<th>sky</th>
<th>people</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>1 1 1 1 1 1 1 1 1</td>
<td>1 1 1 1 1 1 1 1 1</td>
<td>1 1 0 1 1 1 1 1 1</td>
<td>1 0 0 1 1 1 1 1 0</td>
<td>0 1 1 1 0</td>
<td>0 0 0 1 1 1 0</td>
</tr>
</tbody>
</table>

A lexicon of concepts

Temporal dependency

Contextual correlation

---

**Semantic Concept Prediction**

A sequence of video shots (test set)

<table>
<thead>
<tr>
<th>Concept</th>
<th>car</th>
<th>outdoor</th>
<th>urban</th>
<th>building</th>
<th>sky</th>
<th>people</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score</td>
<td>0.92</td>
<td>0.66</td>
<td>0.81</td>
<td>0.62</td>
<td>0.74</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>0.91</td>
<td>0.72</td>
<td>0.37</td>
<td>0.68</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>0.84</td>
<td>0.21</td>
<td>0.86</td>
<td>0.88</td>
<td>0.85</td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td>0.33</td>
<td>0.87</td>
<td>0.73</td>
<td>0.61</td>
<td>0.75</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>0.84</td>
<td>0.94</td>
<td>0.62</td>
<td>0.83</td>
<td>0.26</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>0.96</td>
<td>0.83</td>
<td>0.74</td>
<td>0.90</td>
<td>0.80</td>
<td>0.58</td>
</tr>
</tbody>
</table>

A lexicon of concepts
Our Views

- Detectors’ predictions form a “noisy image” in the contextual-temporal domain

A sequence of video shots (test set)

A lexicon of concepts

Image Denoising

Input Image  
Image denoising  
Output image
Image Denoising

\[
\min_{p_{ij}} \sum_{ij} \left( p_{ij} - q_{ij} \right)^2 \\
+ \sum_{ij} \left( p_{ij} - \frac{p_{i+1,j} + p_{i-1,j} + p_{i,j+1} + p_{i,j-1}}{4} \right)^2
\]

- **Observation**
- **Prior relationship**
- **Energy minimization**
- **Enhanced image**
Main Ideas

- **Denoising**: Exploit prior relationships among nodes to reduce the noise

A sequence of video shots (test set)

- A lexicon of concepts

- Contextual correlation
- Temporal dependency
Main Ideas

• **Denoising**: Exploit prior relationships among nodes to reduce the noise

A sequence of video shots (test set)

```
  car  outdoor  urban  building  sky  people
```

A lexicon of concepts

---

For Semantic Video Indexing

• **Observation** = Detectors’ prediction

• **Prior relationships** = ?

• **Energy function** = ?
Outline

• Multi-Cue Fusion Framework
• Modeling High-Order Relationship
• Inference using High-Order Relationship
• Experiments and Results
• Conclusions

System Framework

Feature Extraction \(\rightarrow\) Training Concept Classifiers \(\rightarrow\) Semantic Concept Prediction

Video Segmentation \(\rightarrow\) Lexicon Annotation \(\rightarrow\) Modeling High-Order Relationships

Video Archive \(\rightarrow\) Shot Ranking

Training path:
Testing path:
Relationship Modeling

• Two issues
  – Relationship discovering?
  – Relationship representation?

Relationship Representation

• The probability of presence of $X$

\[ P(X) = \text{a constant} \]
The probability of presence of $X$

- Given a binary variable $Y$

\[
P(X) = P(X|Y=1)P(Y=1) + P(X|Y=0)P(Y=0)
\]

- Given two binary variables $Y$ and $Z$

\[
P(X) = P(X|Y=1, Z=1)P(Y=1, Z=1) + P(X|Y=0, Z=1)P(Y=0, Z=1) + P(X|Y=1, Z=0)P(Y=1, Z=0) + P(X|Y=0, Z=0)P(Y=0, Z=0)
\]
Relationship Representation

• The probability of presence of $X$
  • Given a partition of data

\[
P(X) = \sum_k P(X|S_k)P(S_k)
\]

Relationship Discovering

• A recursive algorithm selects the variables which are
  – Highly correlated to the target variable
  – Independent of other selected variables

• Chi-square test
  – Discovers the hidden associations
  – Judges whether a correlation is reliable
Toy Example

- Concept lexicon:
  - \{Mountain(M), Sky(S), Tree(T), River(R)\}
- Annotation data \(D\):

- To discover the contextual relationship for Mountain

Toy Example

- Correlation measuring

  \(D\)

- Assume that Sky is the most correlative concept to Mountain
Toy Example

• Data partition according to **Sky**

\[
P(M) = P(M|S = 1)P(S = 1) + P(M|S = 0)P(S = 0)
\]

• Correlation measuring

– Assume that there is **no concept** with significant correlation to **Mountain**
Toy Example

• Correlation measuring

- Assume that River is the most correlative concept to Mountain

To yExam ple

• Data partition according to River

\[
P(M) = P(M|S=1)P(S=1) \\
\quad + P(M|S=0, R=1)P(S=0, R=1) \\
\quad + P(M|S=0, R=0)P(S=0, R=0)
\]

\[S = 1, \quad S = 0, R = 1, \quad S = 0, R = 0\]
Toy Example

- Conditional probability estimation
  \[
  P(M) = P(M|S=1)P(S=1) \\
  + P(M|S=0,R=1)P(S=0,R=1) \\
  + P(M|S=0,R=0)P(S=0,R=0)
  \]
  \[
  \text{E.g.,}
  P(M|S=1) = \frac{7}{8} = 0.875
  \]
  \[
  P(M) = 0.875P(S=1) \\
  + 0.75P(S=0,R=1) \\
  + 0.333P(S=0,R=0)
  \]

- The high-order relationship
  \[
  P(M) = 0.875P(S=1) \\
  + 0.75P(S=0,R=1) \\
  + 0.333P(S=0,R=0)
  \]
  \[
  \text{Independence assumption}
  P(S=0,R=1) \approx P(S=0)P(R=1) \\
  = (1 - P(S))P(R)
  \]
  \[
  P(M) = 0.875P(S) \\
  + 0.75(1 - P(S))P(R) \\
  + 0.333(1 - P(S))(1 - P(R))
  \]
An Example from Real Data

- Concept lexicon: **Columbia374**
- Annotation data: **TRECVid 2005 devel. set**
- The discovered contextual relationship of concept **Mountain**

H: hill  
P: military_personnel  
S: sky  
G: group  
L: landscape  
V: valleys  
C: commercial_advertisement  
R: river  
F: forest  
K: rocky_ground  
W: waterways  
T: trees

Discover temporal dependence among neighboring shots  
Similar to the way discovering the contextual relationships  
Tests the correlation between neighboring shots in the temporal order
Inference using Relationships

A sequence of shots

A lexicon of concepts

\[ \text{Energy Function} \]

\[ \min_{p_{ij}} \sum_{i} \sum_{j} \| p_{ij} - O \mathbf{L}_{ij} \|^2 + \lambda_i \| p_{ij} - C \mathbf{R}_{ij} \|^2 + \kappa_i \| p_{ij} - T \mathbf{R}_{ij} \|^2 \]

- Observed Likelihood
- Contextual Relationship
- Temporal Relationship
Optimization

- **Parameter estimation**
  - Obtain the concept-dependent parameters from training corpus with cross validation

- **Energy minimization**
  - Use *Conjugate Gradient Methods* to solve this non-linear function
  - Adopt prediction scores from detectors as an initial guess

---

Experimental Settings$^{1/2}$

- **TRECVid benchmark**

<table>
<thead>
<tr>
<th>Dataset</th>
<th># of Videos</th>
<th># of Shots</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training data</td>
<td>TV05 devel set</td>
<td>137</td>
</tr>
<tr>
<td>Test data</td>
<td>TV06 test set</td>
<td>259</td>
</tr>
</tbody>
</table>

- **Performance evaluation**
  - 20 officially selected concepts in TRECVid 2006
  - *Inferred average precision* (infAP) for individual concept performance
  - *Mean infAP* for overall system performance
Experimental Settings

- Baselines in our experiments

<table>
<thead>
<tr>
<th></th>
<th>VIREO-374</th>
<th>Columbia374</th>
</tr>
</thead>
<tbody>
<tr>
<td>Provider</td>
<td>City U. of H. K.</td>
<td>Columbia University</td>
</tr>
<tr>
<td>Feature</td>
<td>Color moment, Wavelet texture, Keypoint features</td>
<td>Edge direction histogram, Garbor, Grid color moment</td>
</tr>
<tr>
<td>Learning</td>
<td>SVMs</td>
<td>SVMs</td>
</tr>
<tr>
<td>Fusion</td>
<td>lately average</td>
<td>lately average</td>
</tr>
<tr>
<td><strong>Accuracy</strong></td>
<td><strong>high</strong></td>
<td><strong>medium</strong></td>
</tr>
</tbody>
</table>

Overall performance

<table>
<thead>
<tr>
<th>Baseline</th>
<th>VIREO-374</th>
<th>Columbia374</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean infAP</td>
<td>0.1542</td>
<td>0.0948</td>
</tr>
<tr>
<td>Contextual cues only</td>
<td>Liu et al.</td>
<td>0.2%</td>
</tr>
<tr>
<td>MCF</td>
<td><strong>16.7%</strong></td>
<td><strong>19.6%</strong></td>
</tr>
<tr>
<td>Temporal cues only</td>
<td>Liu et al.</td>
<td>10.6%</td>
</tr>
<tr>
<td>MCF</td>
<td><strong>14.6%</strong></td>
<td><strong>17.3%</strong></td>
</tr>
<tr>
<td>Both cues</td>
<td>Liu et al.</td>
<td>11.2%</td>
</tr>
<tr>
<td>MCF-AC</td>
<td>19.7%</td>
<td>23.3%</td>
</tr>
<tr>
<td>MCF-EM</td>
<td><strong>27.3%</strong></td>
<td><strong>32.1%</strong></td>
</tr>
</tbody>
</table>

**MCF-AC**: MCF with average combination
**MCF-EM**: MCF with energy minimization
Performance of Individual Concepts

Comparison with TRECVid 2006 Submissions
Observations

- MCF improved each of the 20 concepts with ranges varying from 5.9% to 88.1%
- 15 of 20 concepts showed more than 20% improvement
- We achieved ~30% performance gain for two detectors with different levels of accuracy

Concluding Remarks

- Proposed a general framework which improved the accuracy of semantic concept detection in videos
- **MCF has 4 advantages:**
  - Salable
  - Annotation data is reused
  - Temporal and contextual information are used simultaneously
  - Independent to classifiers
Thank You for Your Attention