Discovery of Shared Semantic Spaces for Multi-Scene Video Query

Xun Xu, Timothy Hospedales and Shaogang Gong
Problem

Tasks:
(1) Behaviour Profiling; (2) Behaviour Query; (3) Classification; (4) Summarization
Conventional Approaches

**Approaches**
- Exhaustively annotate each scene
- Train independent models

**Limitations**
- Discover related scenes
- Discover similar activities
- Cross-scene query
- Multi-scene summarization

Wang, Xiaogang, Xiaoxu Ma, and Eric Grimson. "Unsupervised activity perception by hierarchical bayesian models." *CVPR07*

Hospedales, Timothy, Shaogang Gong, and Tao Xiang. "Video behaviour mining using a dynamic topic model." *IJCV12*

Varadarajan, Jagannadan, Rémi Emonet, and Jean-Marc Odobez. "A sequential topic model for mining recurrent activities from long term video logs." *IJCV13*

Multi-Scene Approach

- Challenges
  1. Compute Scene Relatedness
  2. Selective Sharing Information
  3. Construct a Shared Representation
Local Activities

Learning Local Activities

- Local Representations
- Scene Level Clustering
- Scenes in one Cluster

Addressed Problems

- Cross-Scene Query by Example
- Cross-Scene Classification
- Multi-Scene Video Summarization

Behaviour as Shared Profile

Shared Representation
Local Activities

- Feature Construction

  Quantize Optical Flow into 8 directions

Accumulated Optical Flow
Local Activities

- Latent Dirichlet Allocation (LDA)

\[ \alpha^s \] Dirichlet Prior

\[ \beta^s \] Topics/Activities

\[ \theta_j^s \sim \text{Dir}(\alpha^s) \] Activity Distribution in a Video Clip/Document

\[ y_{ij}^s \sim \text{Multinomial}(\theta_j^s) \] Activity indicator

\[ x_{ij}^s \sim \text{Multinomial}(\beta^s ; y_{ij}^s) \] Quantized Optical Flow Vector

Variational Inference to Estimate \( \alpha \) and \( \beta \) given lots of observed video clips
Local Activities

- Examples of Local Activities $\beta$
Multi-Layer Clustering

Cluster Scenes and learn Shared Topic Basis

Local Representations

Scene Level Clustering

Scenes in one Cluster

Addressed Problems

Behaviour as Shared Profile

Shared Representation

Cross-Scene Query by Example

Cross-Scene Classification

Multi-Scene Video Summarization
Multi-Layer Clustering

I. Surveillance video scenes $s = 1 \ldots S$

II. Scene level Clusters $c = 1 \ldots C$

Local topics for each scene
$\{\beta^s_c\}_{l=1}^L$

III. Activity Clusters $t = 1 \ldots T$

STB topics
$\{\beta^{t,b}_k\}_{l=1}^L$
Scene Alignment

- Scaling and Translation to align two scenes to remove cross-scene variance

Scene A

\[ T^A_{\text{norm}} \]

Zscore Normalize Optical Flow Vectors

Scene B

\[ T^B_{\text{norm}} \]

Align Scene B to Scene A

Scene B aligned to Scene A

\[ \left( T^A_{\text{norm}} \right)^{-1} \cdot T^B_{\text{norm}} \]
Scene Level Clustering

- Scene Relatedness Measurement
  - #Activities=6
  - #Activities=7
  - #Activities=5
  - #Activities=4
Scene Level Clustering

Scene Relatedness Measurement

- #Activities=6
  - Relatedness: $(6+6)/(6+7)=0.92$

- #Activities=7
  - Relatedness: $(3+3)/(7+5)=0.5$

- #Activities=5
  - Relatedness: $(4+4)/(5+4)=0.89$
Scene Level Clustering

- Scene Relatedness Measure
  
<table>
<thead>
<tr>
<th>#Activities</th>
<th>Relatedness</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>(6+6)/(6+7)=0.92</td>
</tr>
<tr>
<td>7</td>
<td>(3+3)/(7+5)=0.5</td>
</tr>
<tr>
<td>5</td>
<td>(4+4)/(5+4)=0.89</td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

- Scene Level Clustering
  - Spectral clustering is used to cluster scenes
Learning A Shared Topic Basis

- A single Shared Topic Basis is learned per scene cluster
I. Surveillance video scenes
   \( s = 1 \ldots S \)

II. Scene level Clusters
    \( c = 1 \ldots C \)

   Local topics for each scene
   \( \{ \beta_k^s \}_{k=1}^{K^s} \)

III. STB Topics 1-20
Behaviour as Shared Profile

Each clip is represented as a multinomial distribution.

Video Clip (Behaviour) Profiled by STB

STB Profile $\gamma$ as Bar Chart

STB Topics 1-20
Addressed Problems

Local Representations

Scene Level Clustering

Scenes in one Cluster

Addressed Tasks

Cross-Scene Query by Example

Cross-Scene Classification

Multi-Scene Video Summarization

Behaviour as Shared Profile

Shared Representation
Cross-Scene Query

Retrieval relevant video clips from other scenes by providing a query clip. L2 or cosine distance is computed on STB profile.
Cross-Scene Classification

- Predict the label of a clip in a new scene given training data from other scenes
Multi-Scene Summarization

- Select K clips to cover as many unique behaviours as possible

**Kcenter Clustering:**

\[ J = \max_{j,s \in \mathcal{C}} \left( \min_{j' \in \mathcal{S}} D_{\gamma} (\gamma_{j'}, \gamma_{js}) \right) \]

Select K clips that minimize the farthest distance from any candidate clip to the closest selected clip. Kcenter is good at keeping outliers.

![Original Data](image1)

![Kmeans Result](image2)

![Kcenter Result](image3)
Dataset
- 27 real traffic surveillance scenes
- Each with 18000 frames in 10 fps. 9000 frames for training and rest for testing

LDA settings:
- Optical flow quantize into 8 directions
- 25 frames per clip/document (360 clips per scene)
- $\#\ \text{topics} = 15$

Application Settings:
- 80 frames per clip/document (112 clips per scene)

Annotations:
- 6 scenes from two clusters are annotated into 31/59 categories of behaviours
Multi-Scene Profiling

Multi-Scene Profiling based on Shared Activity Basis

Profilers:
- Vertical Vehicle and Tram (VVT)
- Vertical Vehicle (VV)
- Tram Down (TD)
- Horizontal Pedestrian (HP)
- Horizontal Vehicle (HV)
- Left to Up Turn (LUT)
- Up to Right Turn (URT)
- Vertical Pedestrian (VP)
- Pedestrian and Vehicle Up (PVU)
- Horizontal Pedestrian and Vehicle Up (HPVU)

Profilers based on Local Topics

Active Activities in Scene 1:
- Vehicle Right
- Horizontal Vehicles
- Vertical Vehicles
- Vertical Vehicles & Tram Down
- Vehicle Left

Active Activities in Scene 2:
- Vehicle Down
- Vertical and Down Pedestrian
- Vertical Vehicles
- Vehicle Down & Tram Down
- Vehicle Up & Pedestrian Up

Active Activities in Scene 3:
- Vertical Vehicles
- Left Vehicles & Right to Down Turn
- Vertical Vehicles
- Tram Down
- Vehicle Left
- Vehicle Right & Left to Up Turn

Active Activities in Scene 4:
- Up to Right Turn
- Vehicle Left
- Vertical Vehicle & Tram Down
- Vertical Vehicles
- Vehicle Down
# Cross-Scene Query

<table>
<thead>
<tr>
<th>Query Videos</th>
<th>Cross-Domain Retrieved Videos</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="query from scene 1" /></td>
<td></td>
</tr>
</tbody>
</table>
Cross-Scene Query

- **Comparison of Models:**
  - Flat Model (FM): without multi-layer clustering.
  - Our Scene Cluster Model (SCM): with multi-layer clustering.

- **Evaluation:** Mean Average Precision for first T retrievals
# Cross-Scene Classification

- **Settings**: Leave-One-Out Cross-Validation
- **Evaluation**: Average Accuracy
- **Comparison of Models**:
  - Flat Model (FM): without multi-layer clustering.
  - Our Scene Cluster Model (SCM): with multi-layer clustering.

<table>
<thead>
<tr>
<th>Category</th>
<th>SCM 31</th>
<th>FM 31</th>
<th>SCM 59</th>
<th>FM 59</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scene 1</td>
<td>55.36%</td>
<td>50.89%</td>
<td>42.86%</td>
<td>40.18%</td>
</tr>
<tr>
<td>Scene 2</td>
<td>27.68%</td>
<td>39.29%</td>
<td>18.75%</td>
<td>16.96%</td>
</tr>
<tr>
<td>Scene 3</td>
<td>49.11%</td>
<td>41.96%</td>
<td>39.29%</td>
<td>37.50%</td>
</tr>
<tr>
<td>Scene 4</td>
<td>54.46%</td>
<td>46.43%</td>
<td>37.50%</td>
<td>36.61%</td>
</tr>
<tr>
<td>Scene 5</td>
<td>30.36%</td>
<td>26.79%</td>
<td>17.86%</td>
<td>17.86%</td>
</tr>
<tr>
<td>Scene 6</td>
<td>38.39%</td>
<td>25.00%</td>
<td>20.54%</td>
<td>12.50%</td>
</tr>
<tr>
<td>Average</td>
<td>42.56%</td>
<td>38.39%</td>
<td>29.47%</td>
<td>26.94%</td>
</tr>
</tbody>
</table>
Multi-Scene Summarization

- **Settings**: Select K clips from all video clip across 6 scenes
- **Evaluation**: The percentage of covered unique behaviours in summary
- **Comparison of Scene Model**:
  - Single Scene: concatenate summary from each single scene
  - Flat Model (FM): without multi-layer clustering.
  - Our Scene Cluster Model (SCM): with multi-layer clustering.
- **Comparison of Summarization Models**:
  - Random
  - User Attention
  - Graph Cut
Multi-Scene Summarization

- Scene Cluster 3 (4 scenes in total)
Multi-Scene Summarization

- Scene Cluster 7 (2 scenes in total)
Multi-Scene Summarization

- Across Scene Cluster 3 and 7 (6 scenes in total)
# Multi-Scene Summarization

<table>
<thead>
<tr>
<th>Original Videos</th>
<th>Multi-Scene Summary Videos</th>
</tr>
</thead>
<tbody>
<tr>
<td>origin video elapsed time</td>
<td>2.0 sec</td>
</tr>
<tr>
<td>summary video elapsed time</td>
<td>0.0 sec</td>
</tr>
</tbody>
</table>

[Frame 1]

---

[Images of original and summarized videos with highlighted areas]
Conclusions

- Proposed to model multiple scenes jointly
- Discover scene relatedness by matched topic pairs
- Discover shared activities across scenes
- Multi-scene Activity Profiling
- Cross-scene Query
- Cross-scene Classification
- Multi-scene Summarization
Thank You