CCNY-SRI: An interactive visual event detection system


*The City College of New York,
** SRI International Sarnoff
About Us

• Media lab, The City College of New York (CCNY)
• SRI International

We participated last year’s SED task as “MediaCCNY” for the 1st year
Overview of Our System

- Human tracking is involved
- User is involved as the final decision maker
Outline

• Feature Extraction
• Feature Purification
• Representation
• Event Inference (Classification)
• User Interaction
Feature Extraction

• 2 feature channels are used:
  – 1. STIP-HOG/HOF
  – 2. SURF/MHI – HOG

  Two detectors extract complementary interest feature points
  – Frames are downsampled: 720x576 -> 360x288
Feature Extraction

• Descriptor Channels:
  – Histogram of Gradients (HOG)  Spatial Feature
  – Histogram of Flows (HOF)      Temporal
Outline

• Feature Extraction
• Feature Purification
• Representation
• Event Inference (Classification)
• User Interaction
Feature Purification

• Two issues with extracted feature points:
  – Huge number
  – Too much Noise

• Feature purification is conducted on:
  – Objective Saliency Capture (moving people)
  – Semantic Saliency Capture (event frequency prior)
Human Tracking Mask

- Multiple human tracking bounding boxes are used as filtering masks
• Event specific event belief region is used to capture semantic saliency
Outline

• Feature Extraction
• Feature Purification
• Representation
• Event Inference (Classification)
• User Interaction
Feature Representation

- Local features *(short strings)* inside a ``window'' are aggregated using Bag-of-words model
- Dimension Augmentation using feature mapping *(long strings)*
Feature Aggregation

- Feature dimension:
  - STIP-HOG/HOF: 162   SURF/MHI-HOG: 256

- Code book is built on K-means clustering

- Spatial pooling uses a 3-layer pyramid:

  \[ \text{Feature dimension: 7K} \]
Feature Mapping

• “XOR” problem:

<table>
<thead>
<tr>
<th>label</th>
<th>Original feature (x,y)</th>
<th>Mapped feature (x, y, xy)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>(1,1)</td>
<td>(1,1,1)</td>
</tr>
<tr>
<td>-1</td>
<td>(-1, -1)</td>
<td>(-1,-1,1)</td>
</tr>
<tr>
<td>1</td>
<td>(1, -1)</td>
<td>(1, -1, -1)</td>
</tr>
<tr>
<td>1</td>
<td>(-1, 1)</td>
<td>(1, -1, -1)</td>
</tr>
</tbody>
</table>

• Feature mapping: map original feature to some high dimensional feature space
Outline

• Feature Extraction
• Feature Purification
• Representation
• Event Inference (Classification)
• User Interaction
Event Inference

• Cascade SVMs are used as classifier
• Each unit sample is a temporal window of 60 frames
A Demo iter 1
A Demo iter 2
A Demo iter 3
A Demo iter 4
A Demo iter 4
Outline

• Feature Extraction
• Feature Purification
• Representation
• Event Inference (Classification)
• User Interaction
Human Interaction

• Motivation
  – Let an expert user be the final decision maker
Some Facts about our UI

- “Reject” is the basic move
- “<=“ or “=>” are seldom used
- More than 5 basic moves can be distracting
What did a user do?

Ground Truth

Automatic Detections

After Interaction
What did a user do?

Ground Truth

Automatic Detections

After Interaction
Results

– With 25 mins limit: (rejecting all others)

<table>
<thead>
<tr>
<th>Event</th>
<th>2013 Best</th>
<th>Ours</th>
<th>Cor./FA/Mis.</th>
<th>2013 Best</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>CellToEar</td>
<td>0.902</td>
<td>1.0024</td>
<td>1/23/193</td>
<td>0.9057</td>
<td>0.9991</td>
</tr>
<tr>
<td>Embrace</td>
<td>0.623</td>
<td>0.8573</td>
<td>26/18/149</td>
<td>0.6514</td>
<td>0.8573</td>
</tr>
<tr>
<td>ObjectPut</td>
<td>0.9806</td>
<td>0.9936</td>
<td>6/10/615</td>
<td>0.9803</td>
<td>0.9916</td>
</tr>
<tr>
<td>PeopleMeet</td>
<td>0.8704</td>
<td>0.9534</td>
<td>33/82/416</td>
<td>0.8684</td>
<td>0.9527</td>
</tr>
<tr>
<td>PeopleSplitUp</td>
<td>0.7781</td>
<td>0.9029</td>
<td>20/30/167</td>
<td>0.7771</td>
<td>0.9016</td>
</tr>
<tr>
<td>PersonRuns</td>
<td>0.5850</td>
<td>0.8596</td>
<td>16/28/91</td>
<td>0.5844</td>
<td>0.8590</td>
</tr>
<tr>
<td>Pointing</td>
<td>0.9564</td>
<td>1.0006</td>
<td>13/39/1050</td>
<td>0.9655</td>
<td>0.9959</td>
</tr>
</tbody>
</table>

– Remove 25 mins limit:

<table>
<thead>
<tr>
<th>Event</th>
<th>2013 Best</th>
<th>Ours</th>
<th>Cor./FA/Mis.</th>
<th>2013 Best</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>CellToEar</td>
<td>0.902</td>
<td>1.0027</td>
<td>1/24/193</td>
<td>0.9057</td>
<td>0.9991</td>
</tr>
<tr>
<td>Embrace</td>
<td>0.623</td>
<td>0.7919</td>
<td>39/45/136</td>
<td>0.6514</td>
<td>0.7909</td>
</tr>
<tr>
<td>ObjectPut</td>
<td>0.9806</td>
<td>0.9934</td>
<td>10/29/611</td>
<td>0.9803</td>
<td>0.9924</td>
</tr>
<tr>
<td>PeopleMeet</td>
<td>0.8704</td>
<td>0.9195</td>
<td>65/196/384</td>
<td>0.8684</td>
<td>0.9177</td>
</tr>
<tr>
<td>PeopleSplitUp</td>
<td>0.7781</td>
<td>0.8053</td>
<td>43/75/146</td>
<td>0.7771</td>
<td>0.8050</td>
</tr>
<tr>
<td>PersonRuns</td>
<td>0.5850</td>
<td>0.8596</td>
<td>16/28/91</td>
<td>0.5844</td>
<td>0.8590</td>
</tr>
<tr>
<td>Pointing</td>
<td>0.9564</td>
<td>1.0079</td>
<td>70/225/993</td>
<td>0.9655</td>
<td>0.9952</td>
</tr>
</tbody>
</table>
Observations

- Significant bias is observed between user judgment and ground truth
  - E.g. in PeopleMeet, user brought in 146 clips, while 114 of them is false alarm.

- Improvement is observed in those events with reasonable number of detections
  - weighted fraction of total time for different events?
Acknowledgement

• Our team members:

Xiaodong Yang
Chucai Yi
Prof. Yingli Tian

Dr. Qian Yu
Dr. Amir Tamrakar
Dr. Ajay Divakaran

CCNY
SRI International
Q&A