CMU @ TRECVID Event Detection

Ming-yu Chen & Alex Hauptmann
School of Computer Science
Carnegie Mellon University
CMU @ TRECVID 2009 Event Detection

- CMU submitted all 10 event detection tasks

- Part-based generic approach
  - Local features extracted from videos
    - Local features describe both appearance and motion
    - Bag of word features represent video content
  - Robust to action deformation, occlusion and illumination

- Sliding window detection approach
  - Extend part-based method to detection tasks
  - False alarm reduction is a critical task
System overview

Training

Testing

Interest point detection

Feature descriptors

Feature Extraction

Clustering

Video codebook

Present each sequence as a bag of video words

Learn a model for each action

Model 1

Model n

Sliding windows on test video

Detection/Recognition
MoSIFT – feature detection

- MoSIFT detects spatial interest points in multiple scales
  - Local maximum of Difference of Gaussian (DoG)
- MoSIFT computes optical flow to detect moving areas
- MoSIFT detects video interest areas by local maximum of DoG and optical flows
MoSIFT – feature description

- Descriptor of shape
  - Histogram of Gradient (HoG)
  - Aggregate neighbor areas as 4x4 grids; each grid is described as 8 orientations
  - $4 \times 4 \times 8 = 128$ dimensional vector to describe shape of interest areas

- Descriptor of motion
  - Histogram of Optical Flow (HoF); the same format as HoG
  - 128 dimensional vector to describe motion of interest areas

- 256 dimensional vectors as feature descriptors
Event detection

- K-mean cluster algorithm is applied to quantize feature points extracted from videos
  - K is chosen by cross-validation
- A video codebook is built by clustering result
  - A visual code is a category of similar video interest points
- Bag of word (BoW) feature is constructed for each video sequence
  - Soft weight is used to construct BoW feature
- Event models are trained by Support Vector Machine (SVM)
  - $X^2$ kernel is applied
- Sliding window approach creates video sequence in both training and testing sets
**Evaluation metric - DCR**

- Normalized Detection Cost Rate (NDCR) is used to evaluate performances.

\[
\text{DetectionCost}(S, E) = \text{Cost}_{\text{Miss}} \cdot P_{\text{Miss}}(S, E) + \text{Cost}_{\text{FA}} \cdot R_{\text{FA}}(S, E)
\]

- Strongly penalize false alarms
  - NDCR doesn’t encourage to detect more positive examples as much as reducing false alarms
  - Reducing false alarms is then extremely important to improve NDCR scores
False alarm reduction

- Cascade architecture is highly used to reduce false alarm in detect tasks
- We applied the idea of cascade algorithm in test phase to reduce false alarm
  - Two positive biased classifiers are built (due to computation, it can extend to more layers)
  - Windows pass both classifiers will be predicted as positive

![Diagram of cascade algorithm]

All windows ➔ M1 ➔ M2 ➔ Detected windows

Rejected windows
False alarm reduction (Cont.)

- Lesson from last year, multi-scale sliding window approach has a lot of false alarm
- We do not apply multi-scale this year
- Instead of several short positive predictions, we aggregated consecutive positive predictions as a long positive segment
  - Reduce number of positive predictions
- Performance improves 80% by cascade algorithm
- Performance improves 40% by concatenating short predictions to long predictions
**System set up**

- MoSIFT features are extracted via 3 different scales every 5 frames
  - approximate 2160 hours for a single core to extract MoSIFT features
- A sliding window (25 frames) slides every 5 frames
- 1000 video codes
- Soft weighted BoW feature representation (4 nearest clusters)
- One against all SVM model for each action of each camera view
  - 50 models are built (10 actions * 5 camera views)
Performance comparison

![Graph showing performance comparison for different actions.](image)
Correct detection comparison

![Bar chart showing comparison of correct detection across various actions.]
# Performance (2008 v.s. 2009)

<table>
<thead>
<tr>
<th>Event</th>
<th>#Ref</th>
<th>#Sys</th>
<th>#CorDet</th>
<th>#FA</th>
<th>#Miss</th>
<th>Act. RFA</th>
<th>Act. PMiss</th>
<th>Act. DCR</th>
<th>Min RFA</th>
<th>Min PMiss</th>
<th>Min DCR</th>
</tr>
</thead>
<tbody>
<tr>
<td>CellToEar Event</td>
<td>194</td>
<td>22658</td>
<td>100</td>
<td>22558</td>
<td>94</td>
<td>1479.483</td>
<td>0.484</td>
<td>7.882</td>
<td>20.660</td>
<td>0.995</td>
<td>1.098</td>
</tr>
<tr>
<td>ElevatorNoEntry Event</td>
<td>3</td>
<td>1041</td>
<td>3</td>
<td>1038</td>
<td>0</td>
<td>68.078</td>
<td>0.000</td>
<td>0.340</td>
<td>7.739</td>
<td>0.000</td>
<td>0.039</td>
</tr>
<tr>
<td>Embrace</td>
<td>175</td>
<td>20080</td>
<td>146</td>
<td>19934</td>
<td>29</td>
<td>1307.386</td>
<td>0.166</td>
<td>6.703</td>
<td>1.377</td>
<td>0.989</td>
<td>0.996</td>
</tr>
<tr>
<td>ObjectPut</td>
<td>621</td>
<td>2353</td>
<td>42</td>
<td>2311</td>
<td>579</td>
<td>151.569</td>
<td>0.932</td>
<td>1.690</td>
<td>3.017</td>
<td>0.998</td>
<td>1.014</td>
</tr>
<tr>
<td>OpposingFlow Event</td>
<td>1</td>
<td>2195</td>
<td>1</td>
<td>2194</td>
<td>0</td>
<td>143.895</td>
<td>0.000</td>
<td>0.720</td>
<td>30.956</td>
<td>0.000</td>
<td>0.155</td>
</tr>
<tr>
<td>PeopleMeet Event</td>
<td>449</td>
<td>2130</td>
<td>58</td>
<td>2072</td>
<td>391</td>
<td>135.894</td>
<td>0.871</td>
<td>1.550</td>
<td>36.466</td>
<td>0.998</td>
<td>1.180</td>
</tr>
<tr>
<td>PeopleSplitUp Event</td>
<td>187</td>
<td>10184</td>
<td>28</td>
<td>10156</td>
<td>159</td>
<td>666.088</td>
<td>0.850</td>
<td>4.181</td>
<td>0.721</td>
<td>0.995</td>
<td>0.998</td>
</tr>
<tr>
<td>PersonRuns Event</td>
<td>107</td>
<td>23721</td>
<td>87</td>
<td>23634</td>
<td>20</td>
<td>1530.053</td>
<td>0.187</td>
<td>7.937</td>
<td>2.427</td>
<td>0.991</td>
<td>1.003</td>
</tr>
<tr>
<td>Pointing</td>
<td>1063</td>
<td>7941</td>
<td>234</td>
<td>7707</td>
<td>829</td>
<td>505.469</td>
<td>0.780</td>
<td>3.307</td>
<td>0.066</td>
<td>0.999</td>
<td>0.999</td>
</tr>
<tr>
<td>TakePicture Event</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>0.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>
High level feature extraction

- Motion related high level features
  - 7 motion related concepts
  - Airplane flying, Person playing soccer, Hand, Person playing a musical instrument, Person riding a bicycle, Person eating, People dancing

<table>
<thead>
<tr>
<th></th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>MM</td>
<td>0.24</td>
</tr>
<tr>
<td>PKU</td>
<td>0.21</td>
</tr>
<tr>
<td>TITG</td>
<td>0.20</td>
</tr>
<tr>
<td>CMU</td>
<td>0.18</td>
</tr>
<tr>
<td>FTRD</td>
<td>0.18</td>
</tr>
<tr>
<td>VIREO</td>
<td>0.18</td>
</tr>
<tr>
<td>Eurecom</td>
<td>0.18</td>
</tr>
</tbody>
</table>
Conclusion & future work

- **Conclusion:**
  - A generic approach to detect events
  - MoSIFT features captures both shape and motion information
  - Perform robust over all tasks
  - False alarm reduction is critical to improve DCR

- **Future work:**
  - The approach can’t localize where the action is
  - The approach can further fuse with people tracking and global features
  - Bag of word representation is lack of spatial constraints