GENIE MED 2012 System

Multimedia Archive

Feature Computation
- HOG3D
- Object Bank
- GIST
- MFCC
- ASM

Codebook generation

Base Classifiers
- HIK SVM
- NGD SVM
- Latent SVM

Learn classifiers

Score Fusion
- MFoM
- Expert Forest

Result List

Event Name: Assembling a shelter (Query Event Kit)

Event Kits
- Testing
- Training

500 core Linux PC cluster, 4 GB RAM per core

Single quad-core PC, 8GB RAM
## MED 12 Feature List

<table>
<thead>
<tr>
<th>Feature</th>
<th>Type</th>
<th>Temporal</th>
<th>Spatial</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOG3D</td>
<td>Video</td>
<td>Every 5(^{th}) fr</td>
<td>Max 160 pixels</td>
</tr>
<tr>
<td>Gist</td>
<td>Video</td>
<td>Every 20(^{th}) fr</td>
<td>Full</td>
</tr>
<tr>
<td>Object Bank</td>
<td>Video</td>
<td>1 fr / 2 secs</td>
<td>Full</td>
</tr>
<tr>
<td>MFCC</td>
<td>Audio</td>
<td>10ms</td>
<td>N/A</td>
</tr>
<tr>
<td>ASM</td>
<td>Audio</td>
<td>100-300ms</td>
<td>N/A</td>
</tr>
<tr>
<td>Color-SIFT</td>
<td>Video</td>
<td>1 fr / 2 secs</td>
<td>Full</td>
</tr>
<tr>
<td>Transformed Color Histogram</td>
<td>Video</td>
<td>1 fr / 2 secs</td>
<td>Full</td>
</tr>
<tr>
<td>ISA (Le et al. CVPR 2011)</td>
<td>Video</td>
<td>Full</td>
<td>Max 160</td>
</tr>
<tr>
<td>SUN 09</td>
<td>Video</td>
<td>1 fr / 4 secs</td>
<td>Max 400</td>
</tr>
</tbody>
</table>

Each feature can be used by more than one event agent.
MED As DET Optimization

- DET curves can improve just by fusing more things
  - But does this “solve” MED?

Fusion using Geometric Mean

- 13 multimedia classifiers
- 3 “gradient” classifiers (HOG3D, ISA, CSIFT)
With better fusion algorithms, go on for ever?

Again, does this “solve” MED?

(Average $P_{\text{miss}}@\text{TER}$ over 10 events)
“Solving” MED

- Scene Types Model
  - Begin to “understand” the constituent elements of the video

- MED <-> text
  - Begin to “understand” semantics (of low-level features, black box classifiers, etc.)
Video Representation

- Bag-of-words model?
  - Simple model, lose all temporal information

- Temporal model, e.g. HMM?
  - Relatively temporal rigid structure, often model every frame
We use a key frame representation

- Describe event class by a small set of discriminative sub-events

How to describe a key frame?
Scene Types

- “Scene types” discrete quantization of individual frames

“This is a scene in a kitchen with a person at a table” (scene type X)
Scene types are automatically learned by clustering training video frames.
Learning Scene Type Model for an Event

- Scene types contain some useful clusters
  - And lots of slag

- Which are useful for discriminating an event?

- Develop a Latent SVM to automatically learn which scene types are discriminative for the event

  - Parameters describe which scene types occur in which events
  - Learning only needs single video-level event label
    - All other information is latent, automatically inferred during training/testing
DET Curve (Event 8: Flash Mob)
DET Curves (Event 13: Parkour)
DET Curves (Event 14: Repairing an Appliance)
MED Results

- Probability of missed detection at 5% false positive rate (lower is better)

<table>
<thead>
<tr>
<th>System</th>
<th>Event Class</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>L0:MFCC</td>
<td>52.9</td>
<td>75.0</td>
<td>55.6</td>
<td>67.1</td>
<td>85.0</td>
<td>77.8</td>
<td>64.7</td>
<td>69.3</td>
<td>26.1</td>
<td>74.7</td>
<td>64.8</td>
<td></td>
</tr>
<tr>
<td>L0:OB</td>
<td>70.9</td>
<td>48.2</td>
<td>23.7</td>
<td>30.5</td>
<td>66.3</td>
<td>62.2</td>
<td>49.2</td>
<td>51.5</td>
<td>55.7</td>
<td>54.4</td>
<td>51.3</td>
<td></td>
</tr>
<tr>
<td>L0:MFCC+OB</td>
<td>50.6</td>
<td>39.3</td>
<td>21.5</td>
<td>26.8</td>
<td>65.0</td>
<td>62.2</td>
<td>39.0</td>
<td>43.6</td>
<td>23.9</td>
<td>50.6</td>
<td>42.3</td>
<td></td>
</tr>
<tr>
<td>L1:MFCC</td>
<td>40.7</td>
<td>67.9</td>
<td>41.5</td>
<td>64.6</td>
<td>76.3</td>
<td>71.1</td>
<td>59.9</td>
<td>60.4</td>
<td>25.0</td>
<td>67.1</td>
<td>57.5</td>
<td></td>
</tr>
<tr>
<td>L1:OB</td>
<td>50.0</td>
<td>38.4</td>
<td>14.1</td>
<td>32.9</td>
<td>60.0</td>
<td>45.2</td>
<td>34.2</td>
<td>40.6</td>
<td>34.1</td>
<td>49.4</td>
<td>39.9</td>
<td></td>
</tr>
<tr>
<td>L1:MFCC+OB (proposed)</td>
<td>34.3</td>
<td>33.9</td>
<td>12.6</td>
<td>25.6</td>
<td>48.8</td>
<td>54.1</td>
<td>28.3</td>
<td>30.7</td>
<td>19.3</td>
<td>50.6</td>
<td>33.8</td>
<td></td>
</tr>
</tbody>
</table>
Analysis

- Following slides show examples from MED11 data

- For each video, set of 5 latent key frames are shown
  - Scores for all key frames
  - Corresponding scene-type cluster for each latent key frame
High-scoring Positives
Making a Sandwich

Latent Frame Scores

Score

Frame

Latent Frames

Corresponding Scene-Type Cluster
Making a Sandwich

Latent Frame Scores

Score

Frame

Latent Frames

Corresponding Scene-Type Cluster
Making a Sandwich

Latent Frame Scores

Score

Frame

Latent Frames

Corresponding Scene-Type Cluster

bettyskitchen
Making a Sandwich

Latent Frame Scores

Score

Frame

Latent Frames

Corresponding Scene-Type Cluster
Making a Sandwich

Latent Frame Scores

Score

Latent Frames

Corresponding Scene-Type Cluster

bettyskitchen
Hard Negatives
Making a Sandwich, Hard Negative

Latent Frame Scores

Score

Latent Frames

Corresponding Scene-Type Cluster
Making a Sandwich, Hard Negative

Latent Frame Scores

Score

Frame

Latent Frames

Corresponding Scene-Type Cluster
Making a Sandwich, Hard Negative

Latent Frame Scores

Score

Frame

Latent Frames

Corresponding Scene-Type Cluster
Making a Sandwich, Hard Negative

Latent Frame Scores

Score

Frame

Latent Frames

Corresponding Scene-Type Cluster
Making a Sandwich, Hard Negative

Latent Frame Scores

Score

Frame

Latent Frames

Corresponding Scene-Type Cluster
Visualized MER output for HVC585090

Object Evidence: microphone-with-upper-body, microphone-on-stand, upright-camera-man, mob-5-sitting, mob-5-standing, mob-10-standing, board-on-wall, person

Scene Evidence: crowded indoor

Inferred Evidence Descriptions: Labels from topic and Part-of-speech models

meeting/VERB town/NOUN hall/OBJ microphone/OBJ man/SUBJ-HUMAN people/OBJ speaks/VERB woman/SUBJ-HUMAN chairs/NOUN talking/VERB standing/VERB cameras/OBJ politician/SUBJ-HUMAN podium/OBJ speaking/VERB

(Human Summary - the president answers questions at a town hall meeting in New Hampshire)
Object Evidence: mob-5-sitting, board-on-wall, mob-5-standing, mob-10-sitting, mob-10-standing, person

Scene Evidence: crowded indoor

Inferred Evidence Descriptions: Labels from topic and Part-of-speech models
meeting/VERB hall/OBJ town/NOUN woman/SUBJ-HUMAN people/OBJ speaks/VERB
question/VERB microphone/OBJ audience/SUBJ representative/SUBJ-HUMAN man/SUBJ-
HUMAN talking/VERB asks/VERB podium/OBJ chairs/NOUN
(Human Summary - amateur ad for an institute that instructs and hosts town hall meetings)
Beyond DET Curves

- Demonstration of exploration tool
Thanks!

This work is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20069 and by the Defence Advanced Research Projects Agency (DARPA) under contract number HR0011-08-C-0135. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, DARPA, or the U.S. Government.