2016 TRECVID Multimedia Event Detection Report
Team INF

Junwei Liang, Poyao Huang, Lu Jiang, Zhenzhong Lan, Jia Chen and Alexander Hauptmann
Outline

• System Overview – (10Ex, 100Ex)
  – Feature Representations
• Selected Topics
  – Learning with Miss Videos
• Final Results (MED16EvalSub)
• 0Ex System
• Conclusions
Outline

• System Overview – (10Ex, 100Ex)
  – Feature Representations
• Selected Topics
  – Learning with Miss Videos
• Final Results (MED16EvalSub)
• 0Ex System
• Conclusions
MED-System (10Ex, 100Ex)

- mfcc4k
- VGG19-fc6fc7
- ResNet-pool5prob
- VGG19-Semantic (FCVID+actNet)
- Improved Dense Trajectories
- IDT-Semantic (YFCC+SIN+Sports)

Feature Representations

Self-Paced Curriculum Training (with Miss Videos)

Model Transformation

Weighted Fusion

Carnegie Mellon University
MED-System (10Ex,100Ex)

4 low-level features + 2 semantic features

- mfcc4k
- VGG19-fc6fc7
- ResNet-pool5prob
- VGG19-Semantic (FCVID+actNet)
- Improved Dense Trajectories
- IDT-Semantic (YFCC+SIN+Sports)

Self-Paced Curriculum Training (with Miss Videos)

Model Transformation

Weighted Fusion
MED-System (10Ex,100Ex)
MED-System (10Ex, 100Ex)

- mfcc4k
- VGG19-fc6fc7
- ResNet-pool5prob
- VGG19-Semantic (FCVID+actNet)
- Improved Dense Trajectories
- IDT-Semantic (YFCC+SIN+Sports)

Feature Representations

Self-Paced Curriculum Training (with Miss Videos)

Model Transformation

Weighted Fusion

WX+b
MED-System (10Ex,100Ex)

Feature Representations
- mfcc4k
- VGG19-fc6fc7
- ResNet-pool5prob
- VGG19-Semantic (FCVID+actNet)
- Improved Dense Trajectories
- IDT-Semantic (YFCC+SIN+Sports)

Self-Paced Curriculum Training (with Miss Videos)

Weighted Fusion

Cross-validated weights
MED-System (10Ex, 100Ex)
MED-System (10Ex, 100Ex)

Simple Bag-of-Audio-Word method

Feature Representations

- mfcc4k: 4096x3
- VGG19-fc6fc7: 8192x7
- ResNet-pool5prob: 3048x7
- Explicit Feature Map (order 3 - Intersection Kernel)
- VGG19-Semantic (FCVID+actNet): 200+239
- Improved Dense Trajectories: 110k
- IDT-Semantic (YFCC+SIN+Sports): 1433
MED-System (10Ex,100Ex)

- Low-level CNN features with Explicit Feature Map
- Residual Net - 152
MED-System (10Ex,100Ex)

Improved Dense Trajectories
**MED-System (10Ex,100Ex)**

Semantic feature trained on existing video dataset

<table>
<thead>
<tr>
<th>Feature Representation</th>
<th>Dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>mfcc4k</td>
<td>4096x3</td>
</tr>
<tr>
<td>VGG19-fc6fc7</td>
<td>8192x7</td>
</tr>
<tr>
<td>ResNet-pool5prob</td>
<td>3048x7</td>
</tr>
<tr>
<td><strong>VGG19-Semantic (FCVID+actNet)</strong></td>
<td><strong>200+239</strong></td>
</tr>
<tr>
<td>Improved Dense Trajectories</td>
<td>110k</td>
</tr>
<tr>
<td><strong>IDT-Semantic (YFCC+SIN+Sports)</strong></td>
<td><strong>1433</strong></td>
</tr>
</tbody>
</table>

Carnegie Mellon University
MED-System (10Ex,100Ex)

- **Representations**
  - DCNN
    - ResNet > VGG
  - Kernel
    - Intersection > Chi-square (for CNN features)

* Based on experiments on MED11 TEST
Outline

• System Overview – (10Ex, 100Ex)
  – Feature Representations
• Selected Topics
  – Learning with Miss Videos
• Final Results (MED16EvalSub)
• 0Ex System
• Conclusions
MED – Learning with Miss Videos

• Model Training
  – Batch Train
  – Self-Paced Curriculum Learning
    • Including Miss Videos
      – 10Ex: 10+5; 100Ex: 100+50
Self-Paced Curriculum Learning

• Curriculum Learning (Bengio et al. 2009) or self-paced learning (Kumar et al 2010) is a recently proposed learning paradigm that is inspired by the learning process of humans and animals.

• The samples are not learned randomly but organized in a meaningful order which illustrates from easy to gradually more complex ones.
Self-Paced Curriculum Learning

- Easy samples to complex samples.
  - Easy sample $\Rightarrow$ smaller loss to the already learned model.
  - Complex sample $\Rightarrow$ bigger loss to the already learned model.

\[
\frac{1}{g - kv} \frac{dv}{dt} = 1 \\
\int_{0}^{T} \frac{1}{g - kv} \frac{dv}{dt} dt = \int_{0}^{T} dt \\
\int_{v_0}^{v(T)} \frac{1}{g - kv} dv = T \\
\frac{1}{k} \ln |g - kv(T)| \mid_{v_0}^{v(T)} = T \\
\ln \left| \frac{g - kv(T)}{g - kv_0} \right| = -kT \\
\frac{g - kv(T)}{g - kv_0} = e^{-kT}
\]
Self-Paced Curriculum Learning

• Easy samples to complex samples.
  – Easy sample ➔ Positive Videos
  – Complex sample ➔ Miss Videos

\[
\frac{1}{g - kv} \, dv = 1
\]

\[
\int_0^T \frac{1}{g - kv} \, dv = \int_0^T dt
\]

\[
\int_{v_0}^{v(T)} \frac{1}{g - kv} \, dv = T
\]

\[
\frac{1}{k} \ln |g - kv| \bigg|_{v_0}^{v(T)} = T
\]

\[
\ln \left( \frac{g - kv(T)}{g - kv_0} \right) = -kT
\]

\[
\frac{g - kv(T)}{g - kv_0} = e^{-kT}
\]
Self-Paced Curriculum Learning

Latent weight variable: \( \mathbf{v} = [v_1, \cdots, v_n]^T \)
Model Age: \( \lambda \)
Curriculum Region: \( \Psi \)

\[
\min_{\mathbf{w}, \mathbf{v} \in [0,1]^n} \mathbb{E}(\mathbf{w}, \mathbf{v}, \lambda, \Psi) = \sum_{i=1}^{n} v_i L(y_i, g(x_i, \mathbf{w})) + f(\mathbf{v}; \lambda),
\]
subject to \( \mathbf{v} \in \Psi \)

Loss Function
Regularizer
Biconvex Optimization Problem – Alternate Convex Search
Prior Knowledge
## Model Training - Experiments

### Batch train model

<table>
<thead>
<tr>
<th>5/10 fold mean MAP</th>
<th>mfcc4k</th>
<th>VGG-fc6fc7</th>
<th>resNet-pool5prob</th>
<th>s-vgg</th>
<th>IDT</th>
<th>s-IDT</th>
</tr>
</thead>
<tbody>
<tr>
<td>10Ex-noMiss</td>
<td>0.392</td>
<td>0.380</td>
<td>0.243</td>
<td>0.314</td>
<td>0.279</td>
<td>0.282</td>
</tr>
<tr>
<td>10Ex-incldueMiss</td>
<td>0.377</td>
<td>0.404</td>
<td>0.286</td>
<td>0.191</td>
<td>0.335</td>
<td>0.330</td>
</tr>
<tr>
<td>100Ex-noMiss</td>
<td>0.644</td>
<td>0.772</td>
<td>0.793</td>
<td>0.708</td>
<td>0.593</td>
<td>0.579</td>
</tr>
<tr>
<td>100Ex-incldueMiss</td>
<td>0.556</td>
<td>0.761</td>
<td>0.753</td>
<td>0.692</td>
<td>0.586</td>
<td>0.561</td>
</tr>
</tbody>
</table>

### SPCL train model

<table>
<thead>
<tr>
<th>5/10 fold mean MaxMAP</th>
<th>mfcc4k</th>
<th>VGG-fc6fc7</th>
<th>resNet-pool5prob</th>
<th>s-vgg</th>
<th>IDT</th>
<th>s-IDT</th>
</tr>
</thead>
<tbody>
<tr>
<td>10Ex-noMiss</td>
<td>0.465</td>
<td>0.506</td>
<td>0.391</td>
<td>0.487</td>
<td>0.392</td>
<td>0.483</td>
</tr>
<tr>
<td>10Ex-incldueMiss</td>
<td>0.464</td>
<td>0.575</td>
<td>0.505</td>
<td>0.446</td>
<td>0.437</td>
<td>0.494</td>
</tr>
<tr>
<td>100Ex-noMiss</td>
<td>0.704</td>
<td>0.809</td>
<td>0.829</td>
<td>0.744</td>
<td>0.622</td>
<td>0.618</td>
</tr>
<tr>
<td>100Ex-incldueMiss</td>
<td>0.694</td>
<td>0.818</td>
<td>0.834</td>
<td>0.759</td>
<td>0.630</td>
<td>0.620</td>
</tr>
</tbody>
</table>
# Model Training - Experiments

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>mfcc4k</th>
<th>VGG-fc6fc7</th>
<th>resNet-pool5prob</th>
<th>s-vgg</th>
<th>IDT</th>
<th>s-IDT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>5/10 fold mean MAP</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10Ex-noMiss</td>
<td>10Ex-incldueMiss</td>
<td>0.392</td>
<td>0.380</td>
<td>0.243</td>
<td>0.314</td>
<td>0.279</td>
<td>0.282</td>
</tr>
<tr>
<td>100Ex-noMiss</td>
<td>100Ex-incldueMiss</td>
<td>0.644</td>
<td>0.772</td>
<td>0.793</td>
<td>0.708</td>
<td>0.593</td>
<td>0.579</td>
</tr>
<tr>
<td>100Ex-incldueMiss</td>
<td></td>
<td>0.556</td>
<td>0.761</td>
<td>0.753</td>
<td>0.692</td>
<td>0.586</td>
<td>0.561</td>
</tr>
</tbody>
</table>

**Cross-validated results on training set**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>mfcc4k</th>
<th>VGG-fc6fc7</th>
<th>resNet-pool5prob</th>
<th>s-vgg</th>
<th>IDT</th>
<th>s-IDT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>5/10 fold mean MaxMAP</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10Ex-noMiss</td>
<td>10Ex-incldueMiss</td>
<td>0.465</td>
<td>0.506</td>
<td>0.391</td>
<td>0.487</td>
<td>0.392</td>
<td>0.483</td>
</tr>
<tr>
<td>100Ex-noMiss</td>
<td>100Ex-incldueMiss</td>
<td>0.464</td>
<td>0.575</td>
<td>0.505</td>
<td>0.446</td>
<td>0.437</td>
<td>0.494</td>
</tr>
<tr>
<td>100Ex-incldueMiss</td>
<td></td>
<td>0.704</td>
<td>0.809</td>
<td>0.829</td>
<td>0.744</td>
<td>0.622</td>
<td>0.618</td>
</tr>
<tr>
<td>100Ex-incldueMiss</td>
<td></td>
<td>0.694</td>
<td>0.818</td>
<td>0.834</td>
<td>0.759</td>
<td>0.630</td>
<td>0.620</td>
</tr>
</tbody>
</table>
### Model Training - Experiments

#### Batch train model

<table>
<thead>
<tr>
<th>5/10 fold mean MAP</th>
<th>mfcc4k</th>
<th>VGG-fc6fc7</th>
<th>resNet-pool5prob</th>
<th>s-vgg</th>
<th>IDT</th>
<th>s-IDT</th>
</tr>
</thead>
<tbody>
<tr>
<td>10Ex-noMiss</td>
<td>0.392</td>
<td>0.380</td>
<td>0.243</td>
<td>0.314</td>
<td>0.279</td>
<td>0.282</td>
</tr>
<tr>
<td>10Ex-incldueMiss</td>
<td>0.377</td>
<td>0.404</td>
<td>0.286</td>
<td>0.191</td>
<td>0.335</td>
<td>0.330</td>
</tr>
<tr>
<td>100Ex-noMiss</td>
<td>0.644</td>
<td>0.772</td>
<td>0.793</td>
<td>0.708</td>
<td>0.593</td>
<td>0.579</td>
</tr>
<tr>
<td>100Ex-incldueMiss</td>
<td>0.556</td>
<td>0.761</td>
<td>0.753</td>
<td>0.692</td>
<td>0.586</td>
<td>0.561</td>
</tr>
</tbody>
</table>

Max MAP: the best MAP each run can achieve if we can find the best iteration

#### SPCL train model

<table>
<thead>
<tr>
<th>5/10 fold mean MaxMAP</th>
<th>mfcc4k</th>
<th>VGG-fc6fc7</th>
<th>resNet-pool5prob</th>
<th>s-vgg</th>
<th>IDT</th>
<th>s-IDT</th>
</tr>
</thead>
<tbody>
<tr>
<td>10Ex-noMiss</td>
<td>0.465</td>
<td>0.506</td>
<td>0.391</td>
<td>0.487</td>
<td>0.392</td>
<td>0.483</td>
</tr>
<tr>
<td>10Ex-incldueMiss</td>
<td>0.464</td>
<td>0.575</td>
<td>0.505</td>
<td>0.446</td>
<td>0.437</td>
<td>0.494</td>
</tr>
<tr>
<td>100Ex-noMiss</td>
<td>0.704</td>
<td>0.809</td>
<td>0.829</td>
<td>0.744</td>
<td>0.622</td>
<td>0.618</td>
</tr>
<tr>
<td>100Ex-incldueMiss</td>
<td>0.694</td>
<td>0.818</td>
<td>0.834</td>
<td>0.759</td>
<td>0.630</td>
<td>0.620</td>
</tr>
<tr>
<td></td>
<td>5/10 fold mean MAP</td>
<td></td>
<td></td>
<td>s-vgg</td>
<td>IDT</td>
<td>s-IDT</td>
</tr>
<tr>
<td>----------------------</td>
<td>--------------------</td>
<td>----------------------</td>
<td>----------------------</td>
<td>-------</td>
<td>-----</td>
<td>-------</td>
</tr>
<tr>
<td></td>
<td>mfcc4k</td>
<td>VGG-fc6fc7</td>
<td>resNet-pool5prob</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10Ex-noMiss</td>
<td>0.392</td>
<td>0.380</td>
<td>0.243</td>
<td>0.314</td>
<td>0.279</td>
<td>0.282</td>
</tr>
<tr>
<td>10Ex-incldueMiss</td>
<td>0.377</td>
<td>0.404</td>
<td>0.286</td>
<td>0.191</td>
<td>0.335</td>
<td>0.330</td>
</tr>
<tr>
<td>100Ex-noMiss</td>
<td>0.644</td>
<td>0.772</td>
<td>0.793</td>
<td>0.708</td>
<td>0.593</td>
<td>0.579</td>
</tr>
<tr>
<td>100Ex-incldueMiss</td>
<td>0.556</td>
<td>0.761</td>
<td>0.753</td>
<td>0.692</td>
<td>0.586</td>
<td>0.561</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>5/10 fold mean MaxMAP</th>
<th></th>
<th></th>
<th>s-vgg</th>
<th>IDT</th>
<th>s-IDT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mfcc4k</td>
<td>VGG-fc6fc7</td>
<td>resNet-pool5prob</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10Ex-noMiss</td>
<td>0.465</td>
<td>0.506</td>
<td>0.391</td>
<td>0.487</td>
<td>0.392</td>
<td>0.483</td>
</tr>
<tr>
<td>10Ex-incldueMiss</td>
<td>0.464</td>
<td>0.575</td>
<td>0.505</td>
<td>0.446</td>
<td>0.437</td>
<td>0.494</td>
</tr>
<tr>
<td>100Ex-noMiss</td>
<td>0.704</td>
<td>0.809</td>
<td>0.829</td>
<td>0.744</td>
<td>0.622</td>
<td>0.618</td>
</tr>
<tr>
<td>100Ex-incldueMiss</td>
<td>0.694</td>
<td>0.818</td>
<td>0.834</td>
<td>0.759</td>
<td>0.630</td>
<td>0.620</td>
</tr>
</tbody>
</table>

**Include Miss Videos or Not**

**Batch train model**

**SPCL train model**

**AP** : Better  **AP** : Worse
# Include Miss Videos or Not

## Batch train model

<table>
<thead>
<tr>
<th>5/10 fold mean MAP</th>
<th>mfcc4k</th>
<th>VGG-fc6fc7</th>
<th>resNet-pool5prob</th>
<th>s-vgg</th>
<th>IDT</th>
<th>s-IDT</th>
</tr>
</thead>
<tbody>
<tr>
<td>10Ex-noMiss</td>
<td>0.392</td>
<td>0.380</td>
<td>0.243</td>
<td>0.314</td>
<td>0.279</td>
<td>0.282</td>
</tr>
<tr>
<td>10Ex-incldueMiss</td>
<td>0.377</td>
<td>0.404</td>
<td>0.286</td>
<td>0.191</td>
<td>0.335</td>
<td>0.330</td>
</tr>
<tr>
<td>100Ex-noMiss</td>
<td>0.644</td>
<td>0.772</td>
<td>0.793</td>
<td>0.708</td>
<td>0.593</td>
<td>0.579</td>
</tr>
<tr>
<td>100Ex-incldueMiss</td>
<td>0.556</td>
<td>0.761</td>
<td>0.753</td>
<td>0.692</td>
<td>0.586</td>
<td>0.561</td>
</tr>
</tbody>
</table>

**10Ex with Batch Train: Still Better to include Miss Videos**

## SPCL train model

<table>
<thead>
<tr>
<th>5/10 fold mean MaxMAP</th>
<th>mfcc4k</th>
<th>VGG-fc6fc7</th>
<th>resNet-pool5prob</th>
<th>s-vgg</th>
<th>IDT</th>
<th>s-IDT</th>
</tr>
</thead>
<tbody>
<tr>
<td>10Ex-noMiss</td>
<td>0.465</td>
<td>0.506</td>
<td>0.391</td>
<td>0.487</td>
<td>0.392</td>
<td>0.483</td>
</tr>
<tr>
<td>10Ex-incldueMiss</td>
<td>0.464</td>
<td>0.575</td>
<td>0.505</td>
<td>0.446</td>
<td>0.437</td>
<td>0.494</td>
</tr>
<tr>
<td>100Ex-noMiss</td>
<td>0.704</td>
<td>0.809</td>
<td>0.829</td>
<td>0.744</td>
<td>0.622</td>
<td>0.618</td>
</tr>
<tr>
<td>100Ex-incldueMiss</td>
<td>0.694</td>
<td>0.818</td>
<td>0.834</td>
<td>0.759</td>
<td>0.630</td>
<td>0.620</td>
</tr>
</tbody>
</table>

**AP**: Better  **AP**: Worse
## Include Miss Videos or Not

### Batch train model

<table>
<thead>
<tr>
<th>5/10 fold mean MAP</th>
<th>mfcc4k</th>
<th>VGG-fc6fc7</th>
<th>resNet-pool5prob</th>
<th>s-vgg</th>
<th>IDT</th>
<th>s-IDT</th>
</tr>
</thead>
<tbody>
<tr>
<td>10Ex-noMiss</td>
<td>0.392</td>
<td>0.380</td>
<td>0.243</td>
<td>0.314</td>
<td>0.279</td>
<td>0.282</td>
</tr>
<tr>
<td>10Ex-incldueMiss</td>
<td>0.377</td>
<td>0.404</td>
<td>0.286</td>
<td>0.191</td>
<td>0.335</td>
<td>0.330</td>
</tr>
<tr>
<td>100Ex-noMiss</td>
<td>0.644</td>
<td>0.772</td>
<td>0.793</td>
<td>0.708</td>
<td>0.593</td>
<td>0.579</td>
</tr>
<tr>
<td>100Ex-incldueMiss</td>
<td>0.556</td>
<td>0.761</td>
<td>0.753</td>
<td>0.692</td>
<td>0.586</td>
<td>0.561</td>
</tr>
</tbody>
</table>

**100Ex with Batch Train:** Miss videos confuses the classifiers

### SPCL train model

<table>
<thead>
<tr>
<th>5/10 fold mean MaxMAP</th>
<th>mfcc4k</th>
<th>VGG-fc6fc7</th>
<th>resNet-pool5prob</th>
<th>s-vgg</th>
<th>IDT</th>
<th>s-IDT</th>
</tr>
</thead>
<tbody>
<tr>
<td>10Ex-noMiss</td>
<td>0.465</td>
<td>0.506</td>
<td>0.391</td>
<td>0.487</td>
<td>0.392</td>
<td>0.483</td>
</tr>
<tr>
<td>10Ex-incldueMiss</td>
<td>0.464</td>
<td>0.575</td>
<td>0.505</td>
<td>0.446</td>
<td>0.437</td>
<td>0.494</td>
</tr>
<tr>
<td>100Ex-noMiss</td>
<td>0.704</td>
<td>0.809</td>
<td>0.829</td>
<td>0.744</td>
<td>0.622</td>
<td>0.618</td>
</tr>
<tr>
<td>100Ex-incldueMiss</td>
<td>0.694</td>
<td>0.818</td>
<td>0.834</td>
<td>0.759</td>
<td>0.630</td>
<td>0.620</td>
</tr>
</tbody>
</table>

**AP** : Better  **AP** : Worse
# Include Miss Videos or Not

## Batch train model

<table>
<thead>
<tr>
<th>5/10 fold mean MAP</th>
<th>mfcc4k</th>
<th>VGG-fc6fc7</th>
<th>resNet-pool5prob</th>
<th>s-vgg</th>
<th>IDT</th>
<th>s-IDT</th>
</tr>
</thead>
<tbody>
<tr>
<td>10Ex-noMiss</td>
<td>0.392</td>
<td>0.380</td>
<td>0.243</td>
<td>0.314</td>
<td>0.279</td>
<td>0.282</td>
</tr>
<tr>
<td>10Ex-incldueMiss</td>
<td>0.377</td>
<td>0.404</td>
<td>0.286</td>
<td>0.191</td>
<td>0.335</td>
<td>0.330</td>
</tr>
<tr>
<td>100Ex-noMiss</td>
<td>0.644</td>
<td>0.772</td>
<td>0.793</td>
<td>0.708</td>
<td>0.593</td>
<td>0.579</td>
</tr>
<tr>
<td>100Ex-incldueMiss</td>
<td>0.556</td>
<td>0.761</td>
<td>0.753</td>
<td>0.692</td>
<td>0.586</td>
<td>0.561</td>
</tr>
</tbody>
</table>

**SPCL Train:** Including miss videos is almost always better

## SPCL train model

<table>
<thead>
<tr>
<th>5/10 fold mean MaxMAP</th>
<th>mfcc4k</th>
<th>VGG-fc6fc7</th>
<th>resNet-pool5prob</th>
<th>s-vgg</th>
<th>IDT</th>
<th>s-IDT</th>
</tr>
</thead>
<tbody>
<tr>
<td>10Ex-noMiss</td>
<td>0.465</td>
<td>0.506</td>
<td>0.391</td>
<td>0.487</td>
<td>0.392</td>
<td>0.483</td>
</tr>
<tr>
<td>10Ex-incldueMiss</td>
<td>0.464</td>
<td>0.575</td>
<td>0.505</td>
<td>0.446</td>
<td>0.437</td>
<td>0.494</td>
</tr>
<tr>
<td>100Ex-noMiss</td>
<td>0.704</td>
<td>0.809</td>
<td>0.829</td>
<td>0.744</td>
<td>0.622</td>
<td>0.618</td>
</tr>
<tr>
<td>100Ex-incldueMiss</td>
<td>0.694</td>
<td>0.818</td>
<td>0.834</td>
<td>0.759</td>
<td>0.630</td>
<td>0.620</td>
</tr>
</tbody>
</table>

AP : Better

AP : Worse
### Include Miss Videos or Not

#### Batch train model

<table>
<thead>
<tr>
<th>5/10 fold mean MAP</th>
<th>mfcc4k</th>
<th>VGG-fc6fc7</th>
<th>resNet-pool5prob</th>
<th>s-vgg</th>
<th>IDT</th>
<th>s-IDT</th>
</tr>
</thead>
<tbody>
<tr>
<td>10Ex-noMiss</td>
<td>0.392</td>
<td>0.380</td>
<td>0.243</td>
<td>0.314</td>
<td>0.279</td>
<td>0.282</td>
</tr>
<tr>
<td>10Ex-incldueMiss</td>
<td>0.377</td>
<td>0.404</td>
<td>0.286</td>
<td>0.191</td>
<td>0.335</td>
<td>0.330</td>
</tr>
<tr>
<td>100Ex-noMiss</td>
<td>0.644</td>
<td>0.772</td>
<td>0.793</td>
<td>0.708</td>
<td>0.593</td>
<td>0.579</td>
</tr>
<tr>
<td>100Ex-incldueMiss</td>
<td>0.556</td>
<td>0.761</td>
<td>0.753</td>
<td>0.692</td>
<td>0.586</td>
<td>0.561</td>
</tr>
</tbody>
</table>

#### SPCL train model

<table>
<thead>
<tr>
<th>5/10 fold mean MaxMAP</th>
<th>mfcc4k</th>
<th>VGG-fc6fc7</th>
<th>resNet-pool5prob</th>
<th>s-vgg</th>
<th>IDT</th>
<th>s-IDT</th>
</tr>
</thead>
<tbody>
<tr>
<td>10Ex-noMiss</td>
<td>0.465</td>
<td>0.506</td>
<td>0.391</td>
<td>0.487</td>
<td>0.392</td>
<td>0.483</td>
</tr>
<tr>
<td>10Ex-incldueMiss</td>
<td>0.464</td>
<td>0.575</td>
<td>0.505</td>
<td>0.446</td>
<td>0.437</td>
<td>0.494</td>
</tr>
<tr>
<td>100Ex-noMiss</td>
<td>0.704</td>
<td>0.809</td>
<td>0.829</td>
<td>0.744</td>
<td>0.622</td>
<td>0.618</td>
</tr>
<tr>
<td>100Ex-incldueMiss</td>
<td>0.694</td>
<td>0.818</td>
<td>0.834</td>
<td>0.759</td>
<td>0.630</td>
<td>0.620</td>
</tr>
</tbody>
</table>

10Ex SPCL Train with low-level features: improved over 25%

**AP**: Better  **AP**: Worse
## Comparing BatchTrain and SPCL

<table>
<thead>
<tr>
<th>5/10 fold mean MAP</th>
<th>Batch train model</th>
<th></th>
<th>MaxMAP</th>
<th>SPCL train model</th>
<th></th>
<th>MaxMAP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mfcc4k</td>
<td>VGG-fc6fc7</td>
<td>resNet-pool5prob</td>
<td>s-vgg</td>
<td>IDT</td>
<td>s-IDT</td>
</tr>
<tr>
<td>10Ex-noMiss</td>
<td>0.392</td>
<td>0.380</td>
<td>0.243</td>
<td>0.314</td>
<td>0.279</td>
<td>0.282</td>
</tr>
<tr>
<td>10Ex-incl dueMiss</td>
<td>0.377</td>
<td>0.404</td>
<td>0.286</td>
<td>0.191</td>
<td>0.335</td>
<td>0.330</td>
</tr>
<tr>
<td>100Ex-noMiss</td>
<td>0.644</td>
<td>0.772</td>
<td>0.793</td>
<td>0.708</td>
<td>0.593</td>
<td>0.579</td>
</tr>
<tr>
<td>100Ex-incl dueMiss</td>
<td>0.556</td>
<td>0.761</td>
<td>0.753</td>
<td>0.692</td>
<td>0.586</td>
<td>0.561</td>
</tr>
</tbody>
</table>

Carnegie Mellon University
# Comparing BatchTrain and SPCL

## Batch train model

<table>
<thead>
<tr>
<th>5/10 fold mean MAP</th>
<th>mfcc4k</th>
<th>VGG-fc6fc7</th>
<th>resNet-pool5prob</th>
<th>s-vgg</th>
<th>IDT</th>
<th>s-IDT</th>
</tr>
</thead>
<tbody>
<tr>
<td>10Ex-noMiss</td>
<td>0.392</td>
<td>0.380</td>
<td>0.243</td>
<td>0.314</td>
<td>0.279</td>
<td>0.282</td>
</tr>
<tr>
<td>10Ex-incldueMiss</td>
<td>0.377</td>
<td>0.404</td>
<td>0.286</td>
<td>0.191</td>
<td>0.335</td>
<td>0.330</td>
</tr>
<tr>
<td>100Ex-noMiss</td>
<td>0.644</td>
<td>0.772</td>
<td>0.793</td>
<td>0.708</td>
<td>0.593</td>
<td>0.579</td>
</tr>
<tr>
<td>100Ex-incldueMiss</td>
<td>0.556</td>
<td>0.761</td>
<td>0.753</td>
<td>0.692</td>
<td>0.586</td>
<td>0.561</td>
</tr>
</tbody>
</table>

**SPCL outperforms BatchTrain on all features - 10Ex**

## SPCL train model

<table>
<thead>
<tr>
<th>5/10 fold mean MaxMAP</th>
<th>mfcc4k</th>
<th>VGG-fc6fc7</th>
<th>resNet-pool5prob</th>
<th>s-vgg</th>
<th>IDT</th>
<th>s-IDT</th>
</tr>
</thead>
<tbody>
<tr>
<td>10Ex-noMiss</td>
<td>0.465</td>
<td>0.506</td>
<td>0.391</td>
<td>0.487</td>
<td>0.392</td>
<td>0.483</td>
</tr>
<tr>
<td>10Ex-incldueMiss</td>
<td>0.464</td>
<td>0.575</td>
<td>0.505</td>
<td>0.446</td>
<td>0.437</td>
<td>0.494</td>
</tr>
<tr>
<td>100Ex-noMiss</td>
<td>0.704</td>
<td>0.809</td>
<td>0.829</td>
<td>0.744</td>
<td>0.622</td>
<td>0.618</td>
</tr>
<tr>
<td>100Ex-incldueMiss</td>
<td>0.694</td>
<td>0.818</td>
<td>0.834</td>
<td>0.759</td>
<td>0.630</td>
<td>0.620</td>
</tr>
</tbody>
</table>

Carnegie Mellon University
## Comparing BatchTrain and SPCL

### Batch train model

<table>
<thead>
<tr>
<th>5/10 fold mean MAP</th>
<th>mfcc4k</th>
<th>VGG-fc6fc7</th>
<th>resNet-pool5prob</th>
<th>s-vgg</th>
<th>IDT</th>
<th>s-IDT</th>
</tr>
</thead>
<tbody>
<tr>
<td>10Ex-noMiss</td>
<td>0.392</td>
<td>0.380</td>
<td>0.243</td>
<td>0.314</td>
<td>0.279</td>
<td>0.282</td>
</tr>
<tr>
<td>10Ex-incldueMiss</td>
<td>0.377</td>
<td>0.404</td>
<td>0.286</td>
<td>0.191</td>
<td>0.335</td>
<td>0.330</td>
</tr>
<tr>
<td><strong>100Ex-noMiss</strong></td>
<td><strong>0.644</strong></td>
<td><strong>0.772</strong></td>
<td><strong>0.793</strong></td>
<td><strong>0.708</strong></td>
<td><strong>0.593</strong></td>
<td><strong>0.579</strong></td>
</tr>
<tr>
<td>100Ex-incldueMiss</td>
<td>0.556</td>
<td>0.761</td>
<td>0.753</td>
<td>0.692</td>
<td>0.586</td>
<td>0.561</td>
</tr>
</tbody>
</table>

**SPCL outperforms BatchTrain on all features - 100Ex**

### SPCL train model

<table>
<thead>
<tr>
<th>5/10 fold mean MaxMAP</th>
<th>mfcc4k</th>
<th>VGG-fc6fc7</th>
<th>resNet-pool5prob</th>
<th>s-vgg</th>
<th>IDT</th>
<th>s-IDT</th>
</tr>
</thead>
<tbody>
<tr>
<td>10Ex-noMiss</td>
<td>0.465</td>
<td>0.506</td>
<td>0.391</td>
<td>0.487</td>
<td>0.392</td>
<td>0.483</td>
</tr>
<tr>
<td>10Ex-incldueMiss</td>
<td>0.464</td>
<td>0.575</td>
<td>0.505</td>
<td>0.446</td>
<td>0.437</td>
<td>0.494</td>
</tr>
<tr>
<td><strong>100Ex-noMiss</strong></td>
<td><strong>0.704</strong></td>
<td><strong>0.809</strong></td>
<td><strong>0.829</strong></td>
<td><strong>0.744</strong></td>
<td><strong>0.622</strong></td>
<td><strong>0.618</strong></td>
</tr>
<tr>
<td><strong>100Ex-incldueMiss</strong></td>
<td><strong>0.694</strong></td>
<td><strong>0.818</strong></td>
<td><strong>0.834</strong></td>
<td><strong>0.759</strong></td>
<td><strong>0.630</strong></td>
<td><strong>0.620</strong></td>
</tr>
</tbody>
</table>
## Comparing BatchTrain and SPCL

<table>
<thead>
<tr>
<th></th>
<th>Batch train model</th>
<th>SPCL train model</th>
</tr>
</thead>
<tbody>
<tr>
<td>5/10 fold mean</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mfcc4k</td>
<td>0.392</td>
<td>0.465</td>
</tr>
<tr>
<td>VGG-fc6fc7</td>
<td>0.380</td>
<td>0.506</td>
</tr>
<tr>
<td>resNet-pool5prob</td>
<td>0.243</td>
<td>0.391</td>
</tr>
<tr>
<td>s-vgg</td>
<td>0.314</td>
<td>0.487</td>
</tr>
<tr>
<td>IDT</td>
<td>0.279</td>
<td>0.392</td>
</tr>
<tr>
<td>s-IDT</td>
<td>0.282</td>
<td>0.483</td>
</tr>
<tr>
<td>10Ex-noMiss</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MaxMAP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mfcc4k</td>
<td>0.377</td>
<td>0.464</td>
</tr>
<tr>
<td>VGG-fc6fc7</td>
<td>0.404</td>
<td>0.575</td>
</tr>
<tr>
<td>resNet-pool5prob</td>
<td>0.286</td>
<td>0.505</td>
</tr>
<tr>
<td>s-vgg</td>
<td>0.191</td>
<td>0.446</td>
</tr>
<tr>
<td>IDT</td>
<td>0.335</td>
<td>0.437</td>
</tr>
<tr>
<td>s-IDT</td>
<td>0.330</td>
<td>0.494</td>
</tr>
<tr>
<td>10Ex-incldueMiss</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100Ex-noMiss</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100Ex-incldueMiss</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The weights of late fusion are calculated from cross-validation result (this table)
Outline

• System Overview – (10Ex, 100Ex)
  – Feature Representations
• Selected Topics
  – Learning with Miss Videos
• Final Results (MED16EvalSub)
• 0Ex System
• Conclusions
Final Results – MED16EvalSub

• Test Set
  – Pre-specified Events
  – MED16EvalSub
  – 32000 (16000 HAVIC + 16000 YFCC100M)
YFCC Resources

- YFCC100M video collection:
  - raw and resized videos
  - key-frames
  - video-level and shot-level DCNN features
  - Extracted concepts
  - API to content-based video engine.

https://sites.google.com/site/videosearch100m/
# Final Results – MED16EvalSub

<table>
<thead>
<tr>
<th></th>
<th>MeanxInfAP</th>
<th>E024</th>
<th>E037</th>
</tr>
</thead>
<tbody>
<tr>
<td>BatchTrain_010Ex</td>
<td>33.6</td>
<td>8.8</td>
<td>20.5</td>
</tr>
<tr>
<td>SPCL_010Ex</td>
<td>33.9</td>
<td>13.0</td>
<td>21.7</td>
</tr>
<tr>
<td>BestRun_010Ex*</td>
<td>38.5</td>
<td>19.2</td>
<td>24.5</td>
</tr>
<tr>
<td>BatchTrain_100Ex</td>
<td>46.4</td>
<td>20.0</td>
<td>33.0</td>
</tr>
<tr>
<td>SPCL_100Ex</td>
<td>47.3</td>
<td>24.8</td>
<td>36.6</td>
</tr>
<tr>
<td>BestRun_100Ex*</td>
<td>47.5</td>
<td>16.4</td>
<td>31.2</td>
</tr>
</tbody>
</table>

* Excluding our runs
## Final Results – MED16EvalSub

<table>
<thead>
<tr>
<th></th>
<th>MeanxInfAP</th>
<th>E024</th>
<th>E037</th>
</tr>
</thead>
<tbody>
<tr>
<td>BatchTrain_010Ex</td>
<td>33.6</td>
<td>8.8</td>
<td>20.5</td>
</tr>
<tr>
<td><strong>SPCL_010Ex</strong></td>
<td><strong>33.9</strong></td>
<td>13.0</td>
<td>21.7</td>
</tr>
<tr>
<td>BestRun_010Ex*</td>
<td>38.5</td>
<td>19.2</td>
<td>24.5</td>
</tr>
<tr>
<td>BatchTrain_100Ex</td>
<td>46.4</td>
<td>20.0</td>
<td>33.0</td>
</tr>
<tr>
<td><strong>SPCL_100Ex</strong></td>
<td><strong>47.3</strong></td>
<td>24.8</td>
<td>36.6</td>
</tr>
<tr>
<td>BestRun_100Ex*</td>
<td>47.5</td>
<td>16.4</td>
<td>31.2</td>
</tr>
</tbody>
</table>

*Excluding our runs

SPCL performs OK on 100Ex, badly on 10Ex
## Final Results – MED16EvalSub

<table>
<thead>
<tr>
<th></th>
<th>MeanxInfAP</th>
<th>E024</th>
<th>E037</th>
</tr>
</thead>
<tbody>
<tr>
<td>BatchTrain_010Ex</td>
<td>33.6</td>
<td>8.8</td>
<td>20.5</td>
</tr>
<tr>
<td><strong>SPCL_010Ex</strong></td>
<td><strong>33.9</strong></td>
<td><strong>13.0</strong></td>
<td><strong>21.7</strong></td>
</tr>
<tr>
<td>BestRun_010Ex*</td>
<td>38.5</td>
<td>19.2</td>
<td>24.5</td>
</tr>
<tr>
<td>BatchTrain_100Ex</td>
<td>46.4</td>
<td>20.0</td>
<td>33.0</td>
</tr>
<tr>
<td><strong>SPCL_100Ex</strong></td>
<td><strong>47.3</strong></td>
<td><strong>24.8</strong></td>
<td><strong>36.6</strong></td>
</tr>
<tr>
<td>BestRun_100Ex*</td>
<td>47.5</td>
<td>16.4</td>
<td>31.2</td>
</tr>
</tbody>
</table>

* Excluding our runs

SPCL performs slightly better than BatchTrain
(How to find the best iteration model?)
(Now we use Iteration 10/30 model)
Final Results – MED16EvalSub

<table>
<thead>
<tr>
<th></th>
<th>MeanxInfAP</th>
<th>E024</th>
<th>E037</th>
</tr>
</thead>
<tbody>
<tr>
<td>BatchTrain_010Ex</td>
<td>33.6</td>
<td>8.8</td>
<td>20.5</td>
</tr>
<tr>
<td>SPCL_010Ex</td>
<td>33.9</td>
<td>13.0</td>
<td>21.7</td>
</tr>
<tr>
<td>BestRun_010Ex*</td>
<td>38.5</td>
<td>19.2</td>
<td>24.5</td>
</tr>
<tr>
<td>BatchTrain_100Ex</td>
<td>46.4</td>
<td>20.0</td>
<td>33.0</td>
</tr>
<tr>
<td><strong>SPCL_100Ex</strong></td>
<td><strong>47.3</strong></td>
<td><strong>24.8</strong></td>
<td><strong>36.6</strong></td>
</tr>
<tr>
<td>BestRun 100Ex*</td>
<td>47.5</td>
<td>16.4</td>
<td>31.2</td>
</tr>
</tbody>
</table>

* Excluding our runs

Selected Events where SPCL is better than the other runs
# Final Results – MED16EvalSub

<table>
<thead>
<tr>
<th></th>
<th>MeanxInfAP</th>
<th>E024</th>
<th>E037</th>
</tr>
</thead>
<tbody>
<tr>
<td>BatchTrain_010Ex</td>
<td>33.6</td>
<td>8.8</td>
<td>20.5</td>
</tr>
<tr>
<td>SPCL_010Ex</td>
<td>33.9</td>
<td>13.0</td>
<td>21.7</td>
</tr>
<tr>
<td>BestRun_010Ex*</td>
<td>38.5</td>
<td>19.2</td>
<td>24.5</td>
</tr>
<tr>
<td>BatchTrain_100Ex</td>
<td>46.4</td>
<td>20.0</td>
<td>33.0</td>
</tr>
<tr>
<td>SPCL_100Ex</td>
<td>47.3</td>
<td>24.8</td>
<td>36.6</td>
</tr>
<tr>
<td>BestRun_100Ex*</td>
<td>47.5</td>
<td>16.4</td>
<td>31.2</td>
</tr>
</tbody>
</table>

* Excluding our runs

Selected Events where SPCL performs better than BatchTrain
Final Results – MED16EvalSub

<table>
<thead>
<tr>
<th></th>
<th>MeanxInfAP</th>
<th>E022</th>
<th>E028</th>
<th>E036</th>
</tr>
</thead>
<tbody>
<tr>
<td>BatchTrain_010Ex</td>
<td>33.6</td>
<td>15.8</td>
<td>40.0</td>
<td>48.2</td>
</tr>
<tr>
<td>SPCL_010Ex</td>
<td>33.9</td>
<td>13.7</td>
<td>47.3</td>
<td>52.5</td>
</tr>
<tr>
<td>BestRun_010Ex*</td>
<td>38.5</td>
<td>18.3</td>
<td>47.0</td>
<td>38.1</td>
</tr>
<tr>
<td>BatchTrain_100Ex</td>
<td>46.4</td>
<td>40.1</td>
<td>58.7</td>
<td>54.2</td>
</tr>
<tr>
<td>SPCL_100Ex</td>
<td>47.3</td>
<td>41.0</td>
<td>57.5</td>
<td>50.9</td>
</tr>
<tr>
<td>BestRun_100Ex*</td>
<td>47.5</td>
<td>39.4</td>
<td>52.0</td>
<td>47.1</td>
</tr>
</tbody>
</table>

* Excluding our runs

But sometimes SPCL is worse than BatchTrain (Important to find the best model in SPCL)
Outline

• System Overview – (10Ex, 100Ex)
  – Feature Representations
• Selected Topics
  – Learning with Miss Videos
• Final Results (MED16EvalSub)
• 0Ex System
• Conclusions
MED-pipeline (0Ex)

Test Videos

VGG19-Semantic (FCVID+actNet)
IDT-Semantic (YFCC+SIN+Sports)

Feature Representations

Event Kits

Semantic Feature Dictionary

Word Matching
Exact/Stemmed

FCVID+actNet
YFCC+SIN+Sports

Event Model W1
Event Model W2

Weighted Fusion

\[ y = Wx \]
Simple word matching to get regression models (No SQG)

It performs well if the event kit text is in the dictionary
(E037 Parking a vehicle -> ParkingCars FCVID)
Outline

• System Overview – (10Ex, 100Ex)
  – Feature Representations
• Selected Topics
  – Learning with Miss Videos
• Final Results (MED16EvalSub)
• 0Ex System
• Conclusions
Conclusions

• We present a 10/100 Ex system trained with miss video using self-paced curriculum learning.

• In the future, we will find better way to get model from SPCL iterations (the model before overfitting to noise)
2016 TRECVID
Ad-hoc Video Search - Report
Team INF

Junwei Liang, Poyao Huang, Lu Jiang, Zhenzhong Lan, Jia Chen and Alexander Hauptmann
Outline

• System Overview
• Selected Topics
  – Webly-Labeled Learning
  – Experimental Results
    • FCVID and YFCC
    • AVS Extra
• Conclusions
Outline

• System Overview

• Selected Topics
  – Webly-Labeled Learning
  – Experimental Results
    • FCVID and YFCC
    • AVS Extra

• Conclusions
System Overview

• Task
  – Given a text query, find relevant video shots in 116,097 shots (> 3sec)
  – Queries:
    01 Find shots of a person playing guitar outdoors
        ...
    03 Find shots of a person playing drums indoors
        ...
    28 Find shots of a person wearing a helmet
    29 Find shots of a person lighting a candle
        ...
System Overview

• System Type
  – F: Fully Automatic
  – E: Used only training data collected automatically using only the official query textual description. (No annotation Run)
System Overview

Ad-hoc Video Query
A person drinking from a cup, mug, bottle, or other container

Video Detector

Target Video Collection

Webly Learning System

Internet Videos

Query Result

1  2  3
System Overview

Ad-hoc Video Query
A person drinking from a cup, mug, bottle, or other container

Ad-hoc Query Text

Video Detector

Target Video Collection

Webly Learning System

Internet Videos

Query Result

1
2
3

Carnegie Mellon University
System Overview

Ad-hoc Video Query
A person drinking from a cup, mug, bottle, or other container.

Video Detector

Webly Learning System

Target Video Collection

Internet Videos e.g. Youtube

Query Result
1 2 3...
Outline

• System Overview
• Selected Topics
  – Webly-Labeled Learning
  – Experimental Results
    • FCVID and YFCC
    • AVS Extra
• Conclusions
Webly Labeled Learning

• Learn from webly-labeled* video data
  – Virtually unlimited data
  – No need for manual annotation
  – But very noisy

*Webly stands for typically useful but often unreliable information in web content
Webly Labeled Video:

**Title:** How I walk with my dog

**Description:** Every morning my dog jumps and pulls hard on his leash, because he is excited. So my dog needs training. When I take him out for a walk, I tell him to sit many times. When he pulls the leash, I stop, turn around, and make a "clicking" sound or tap my thigh. Then he will follow me. When he behaves while we’re walking, I praise him for 2 to 10 seconds and then give him a treat. Praising is very important.

**Tags:** How, to, walk, with, my, dog, (along, road)

**Categories:** “Pets & Animals”

**Comments:**
- ArcticElite: “Let me guess, hes a good boy, eh? ;-)”
- rucksluvr: “nice snood”

“"All right. Good Boy. Good Boy. Wait. Sit, good.”

- cocker: 0.402, bullterrier: 0.338, shipperke: 0.163, schnauzer: 0.121, greyhound: 0.118, labrador: 0.109, bulldog: 0.088...
Webly Labeled Video:

Title: How I walk with my dog
Description: Every morning my dog jumps and pulls hard on his leash, because he is excited. So my dog needs training. When I take him out for a walk, I tell him to sit many times. When he pulls the leash, I stop, turn around, and make a "clicking" sound or tap my thigh. Then he will follow me. When he behaves while we’re walking, I praise him for 2 to 10 seconds and then give him a treat. Praising is very important.

Tags: How, to, walk, with, my, dog, (along, read)
Categories: “Pets & Animals”
Comments:
- ArcticElite: “Let me guess, hes a good boy, eh? ;)”
- rucksluvr: “nice snood”


cocker:0.402, bullterrier:0.338, shipperke:0.163, schnauzer:0.121, greyhound:0.118, labrador:0.109, bulldog:0.088…
AVS Webly Learning Pipeline

Video Webistes (Youtube)

Phrase-Table (GoogleNews)

BM25 + Word2vec

Query Words

VGG19-fc6fc7
Explicit Feature Map (order 3 - Intersection Kernel)

C3D

8192x7
4096x7

Feature Representations

Curriculum

Self-Paced Curriculum Training

Test Videos

Carnegie Mellon University
AVS Webly Learning Pipeline

Collect Videos & Design Curriculum (i.e. How Confident the videos are related to the query)

Video Webistes (Youtube)

Phrase-Table (GoogleNews)

Query Words

Prior knowledge

Text Query

BM25 + Word2vec

Curriculum

Self-Paced Curriculum Training

Model Transformation

Average Fusion

Test Videos

Feature Representations

VGG19-fc6fc7

C3D

Explicit Feature Map (order 3 - Intersection Kernel)

8192x7

4096x7
AVS Webly Learning Pipeline

Video Webistes (Youtube)

Phrase-Table (GoogleNews)

Query Words

Video

Metadata

BM25 + Word2vec

VGG19-fc6fc7

C3D

Explicit Feature Map (order 3 - Intersection Kernel)

Feature Representations

8192x7

4096x7

Curriculum

Self-Paced Curriculum Training

Model Transformation

Average Fusion

Test Videos

Test Videos
AVS Webly Learning Pipeline

- Video Webistes (Youtube)
- Pharse-Table (GoogleNews)
- Query Words
- BM25 + Word2vec
- Text Query

- Videos + Metadata
- Test Videos

- Feature Representations
  - VGG19-fc6fc7
  - C3D
  - Explicit Feature Map (order 3 - Intersection Kernel)

- Curriculum
- Self-Paced Curriculum Training
- Model Transformation
- Average Fusion

Test Videos
WEbly-Labeled Learning

• Curriculum Learning (Bengio et al. 2009) or self-paced learning (Kumar et al 2010) is a recently proposed learning paradigm that is inspired by the learning process of humans and animals.

• The samples are not learned randomly but organized in a meaningful order which illustrates from easy to gradually more complex ones.
WEbly-Labeled Learning

• Easy samples to complex samples.
  – Easy sample $\Rightarrow$ smaller loss to the already learned model.
  – Complex sample $\Rightarrow$ bigger loss to the already learned model.

\[
\frac{1}{g - kv} \frac{dv}{dt} = 1 \\
\int_{t_0}^{T} \frac{1}{g - kv} \frac{dv}{dt} dt = \int_{0}^{T} dt \\
\int_{t_0}^{v(T)} \frac{1}{g - kv} dv = T \\
\frac{1}{k} \ln |g - kv|_{t_0}^{v(T)} = T \\
\ln \left( \frac{g - kv(T)}{g - kv_0} \right) = -kT \\
\frac{g - kv(T)}{g - kv_0} = e^{-kT}
\]
WEbly-Labeled Learning

$$\min_{w,v \in [0,1]^n} \mathbb{E}(w,v,\lambda,\Psi) = \sum_{i=1}^{n} v_i L(y_i,g(x_i,w)) + f(v;\lambda),$$
subject to $v \in \Psi$

Latent weight variable: $v = [v_1, \cdots, v_n]^T$
Model Age: $\lambda$
Curriculum Region: $\Psi$
WEbly-Labeled Learning

\[
\min_{\mathbf{w}, \mathbf{v} \in [0,1]^n} \mathbb{E}(\mathbf{w}, \mathbf{v}; \lambda, \Psi) = \sum_{i=1}^{n} v_i L(y_i, g(x_i, \mathbf{w})) + f(\mathbf{v}; \lambda),
\]

subject to \( \mathbf{v} \in \Psi \)

Latent weight variable: \( \mathbf{v} = [v_1, \cdots, v_n]^T \)
Model Age: \( \lambda \)
Curriculum Region: \( \Psi \)
WEbly-Labeled Learning

Latent weight variable: \( \mathbf{v} = [v_1, \ldots, v_n]^T \)
Model Age: \( \lambda \)
Curriculum Region: \( \Psi \)

\[
\min_{\mathbf{w}, \mathbf{v} \in [0,1]^n} \mathbb{E}(\mathbf{w}, \mathbf{v}, \lambda, \Psi) = \sum_{i=1}^{n} v_i L(y_i, g(x_i, \mathbf{w})) + f(\mathbf{v}; \lambda),
\]
subject to \( \mathbf{v} \in \Psi \)

Biconvex Optimization Problem – Alternate Convex Search

Loss Function
Regularizer

Webly Labeled Prior Knowledge
Algorithm

\[
\min_{w, v \in [0, 1]^n} \mathbb{E}(w, v, \lambda, \Psi) = \sum_{i=1}^{n} v_i L(y_i, g(x_i, w)) + f(v; \lambda),
\]

subject to \( v \in \Psi \)

---

**Algorithm 1: Webly-labeled Learning (WELL).**

**input:** Input dataset \( \mathcal{D} \), curriculum region \( \Psi \), self-paced function \( f \) and a step size \( \mu \)

**output:** Model parameter \( w \)

1. Initialize \( v^* \), \( \lambda \) in the curriculum region;
2. while not converged do
3. \quad Update \( w^* = \arg\min_w \mathbb{E}(w, v^*; \lambda, \Psi); \)
4. \quad Update \( v^* = \arg\min_v \mathbb{E}(w^*, v; \lambda, \Psi); \)
5. \quad if \( \lambda \) is small then increase \( \lambda \) by the step size \( \mu \);
6. end
7. return \( w^* \)
Algorithm

\[
\begin{align*}
\min_{w,v} \mathbb{E}(w,v,\lambda,\Psi) = & \sum_{i=1}^{n} v_i L(y_i, g(x_i, w)) + f(v; \lambda), \\
\text{subject to } v \in \Psi
\end{align*}
\]

Algorithm 1: Webly-labeled Learning (WELL).

**input**: Input dataset \( \mathcal{D} \), curriculum region \( \Psi \), self-paced function \( f \) and a step size \( \mu \)

**output**: Model parameter \( w \)

1. Initialize \( v^*, \lambda \) in the curriculum region;
2. while not converged do
3. \hspace{1cm} Update \( w^* = \arg\min_w \mathbb{E}(w, v^*; \lambda, \Psi) \);
4. \hspace{1cm} Update \( v^* = \arg\min_v \mathbb{E}(w^*, v; \lambda, \Psi) \);
5. \hspace{1cm} if \( \lambda \) is small then increase \( \lambda \) by the step size \( \mu \);
6. end
7. return \( w^* \)
Algorithm 24

\[
\min_{w, v \in [0,1]^n} \mathbb{E}(w, v, \lambda, \Psi) = \sum_{i=1}^{n} v_i L(y_i, g(x_i, w)) + f(v; \lambda),
\]

subject to \( v \in \Psi \)

**Algorithm 1: Webly-labeled Learning (WELL).**

**input:** Input dataset \( D \), curriculum region \( \Psi \), self-paced function \( f \) and a step size \( \mu \)

**output:** Model parameter \( w \)

1. Initialize \( v^* \), \( \lambda \) in the curriculum region;
2. **while not converged** do
3. \( \text{Update } w^* = \arg \min_w \mathbb{E}(w, v^*; \lambda, \Psi); \)
4. \( \text{Update } v^* = \arg \min_v \mathbb{E}(w^*, v; \lambda, \Psi); \)
5. **if** \( \lambda \text{ is small} \) **then** increase \( \lambda \) by the step size \( \mu \);
6. **end**
7. **return** \( w^* \)
Outline

• System Overview

• Selected Topics
  – Webly-Labeled Learning
  – Experimental Results
    • FCVID and YFCC (*)
    • AVS Extra

• Conclusions

Outline

• System Overview
• Selected Topics
  – Webly-Labeled Learning
  – Experimental Results
    • FCVID and YFCC (-)
    • AVS Extra
• Conclusions
## AVS – Extra Experiments

<table>
<thead>
<tr>
<th>Model</th>
<th>MeanxInfAP</th>
<th>505</th>
<th>509</th>
<th>511</th>
</tr>
</thead>
<tbody>
<tr>
<td>IACC.3_VGG</td>
<td>0.003</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BatchTrain_VGG_top1000</td>
<td>0.016</td>
<td>0.002</td>
<td>0.099</td>
<td>0.033</td>
</tr>
<tr>
<td>C3D_top1000</td>
<td>0.024</td>
<td>0.003</td>
<td><strong>0.123</strong></td>
<td>0.040</td>
</tr>
<tr>
<td>VGG_top1000*</td>
<td>0.024</td>
<td>0.020</td>
<td>0.030</td>
<td>0.080</td>
</tr>
<tr>
<td>VGG_top500</td>
<td>0.029</td>
<td><strong>0.021</strong></td>
<td>0.044</td>
<td>0.088</td>
</tr>
<tr>
<td>C3D+VGG_top1000*</td>
<td>0.040</td>
<td>0.013</td>
<td>0.117</td>
<td><strong>0.109</strong></td>
</tr>
<tr>
<td>Best System (F)**</td>
<td>0.054</td>
<td><strong>0.002</strong></td>
<td><strong>0.036</strong></td>
<td><strong>0.025</strong></td>
</tr>
</tbody>
</table>

* The system runs that we submitted
** Excluding our system runs
## AVS – Extra Experiments

Only learning from IACC.3 metadata - failed

<table>
<thead>
<tr>
<th></th>
<th>Mean x InfAP</th>
<th>505</th>
<th>509</th>
<th>511</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>IACC.3_VGG</strong></td>
<td>0.003</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BatchTrain_VGG_top1000</td>
<td>0.016</td>
<td>0.002</td>
<td>0.099</td>
<td>0.033</td>
</tr>
<tr>
<td>C3D_top1000</td>
<td>0.024</td>
<td>0.003</td>
<td><strong>0.123</strong></td>
<td>0.040</td>
</tr>
<tr>
<td>VGG_top1000*</td>
<td>0.024</td>
<td>0.020</td>
<td>0.030</td>
<td>0.080</td>
</tr>
<tr>
<td>VGG_top500</td>
<td>0.029</td>
<td><strong>0.021</strong></td>
<td>0.044</td>
<td>0.088</td>
</tr>
<tr>
<td>C3D+VGG_top1000*</td>
<td>0.040</td>
<td>0.013</td>
<td>0.117</td>
<td><strong>0.109</strong></td>
</tr>
<tr>
<td>Best System (F)**</td>
<td>0.054</td>
<td><strong>0.002</strong></td>
<td>0.036</td>
<td>0.025</td>
</tr>
</tbody>
</table>

* The system runs that we submitted
** Excluding our system runs
## AVS – Extra Experiments

**Better than simple batch train 50%**

<table>
<thead>
<tr>
<th>Model Name</th>
<th>MeanInfAP</th>
<th>505</th>
<th>509</th>
<th>511</th>
</tr>
</thead>
<tbody>
<tr>
<td>IACC.3_VGG</td>
<td>0.003</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BatchTrain_VGG_top1000</td>
<td>0.016</td>
<td>0.002</td>
<td>0.099</td>
<td>0.033</td>
</tr>
<tr>
<td>C3D_top1000</td>
<td>0.024</td>
<td>0.003</td>
<td><strong>0.123</strong></td>
<td>0.040</td>
</tr>
<tr>
<td><strong>VGG_top1000</strong>*</td>
<td>0.024</td>
<td>0.020</td>
<td>0.030</td>
<td>0.080</td>
</tr>
<tr>
<td>VGG_top500</td>
<td>0.029</td>
<td><strong>0.021</strong></td>
<td>0.044</td>
<td>0.088</td>
</tr>
<tr>
<td><strong>C3D+VGG_top1000</strong>*</td>
<td>0.040</td>
<td>0.013</td>
<td>0.117</td>
<td><strong>0.109</strong></td>
</tr>
<tr>
<td>Best System (F)**</td>
<td>0.054</td>
<td><strong>0.002</strong></td>
<td><strong>0.036</strong></td>
<td><strong>0.025</strong></td>
</tr>
</tbody>
</table>

* The system runs that we submitted
** Excluding our system runs
AVS – Extra Experiments

Combining C3D and VGG improved 67%

<table>
<thead>
<tr>
<th></th>
<th>MeanxInfAP</th>
<th>505</th>
<th>509</th>
<th>511</th>
</tr>
</thead>
<tbody>
<tr>
<td>IACC.3_VGG</td>
<td>0.003</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BatchTrain_VGG_top1000</td>
<td>0.016</td>
<td>0.002</td>
<td>0.099</td>
<td>0.033</td>
</tr>
<tr>
<td>C3D_top1000</td>
<td>0.024</td>
<td>0.003</td>
<td><strong>0.123</strong></td>
<td>0.040</td>
</tr>
<tr>
<td>VGG_top1000*</td>
<td>0.024</td>
<td>0.020</td>
<td>0.030</td>
<td>0.080</td>
</tr>
<tr>
<td>VGG_top500</td>
<td>0.029</td>
<td><strong>0.021</strong></td>
<td>0.044</td>
<td>0.088</td>
</tr>
<tr>
<td>C3D+VGG_top1000*</td>
<td>0.040</td>
<td>0.013</td>
<td>0.117</td>
<td><strong>0.109</strong></td>
</tr>
<tr>
<td>Best System (F)**</td>
<td>0.054</td>
<td><strong>0.002</strong></td>
<td>0.036</td>
<td>0.025</td>
</tr>
</tbody>
</table>

* The system runs that we submitted
** Excluding our system runs
## AVS – Extra Experiments

<table>
<thead>
<tr>
<th>Model</th>
<th>MeanInfAP</th>
<th>505</th>
<th>509</th>
<th>511</th>
</tr>
</thead>
<tbody>
<tr>
<td>IACC.3_VGG</td>
<td>0.003</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>C3D_top1000</td>
<td>0.024</td>
<td>0.003</td>
<td>0.123</td>
<td>0.040</td>
</tr>
<tr>
<td>VGG_top1000*</td>
<td>0.024</td>
<td>0.020</td>
<td>0.030</td>
<td>0.080</td>
</tr>
<tr>
<td>VGG_top500</td>
<td>0.029</td>
<td>0.021</td>
<td>0.044</td>
<td>0.088</td>
</tr>
<tr>
<td>C3D+VGG_top1000*</td>
<td>0.040</td>
<td>0.013</td>
<td>0.117</td>
<td>0.109</td>
</tr>
<tr>
<td>Best System (F)**</td>
<td>0.054</td>
<td>0.002</td>
<td>0.036</td>
<td>0.025</td>
</tr>
</tbody>
</table>

* The system runs that we submitted
** Excluding our system runs

Selected queries where our system significantly outperforms the rest
AVS – Extra Experiments

Selected queries where our system performs very badly (about 14 out of 30 are under 0.01)

<table>
<thead>
<tr>
<th>Model Combination</th>
<th>MeanInfAP</th>
<th>506</th>
<th>513</th>
<th>522</th>
</tr>
</thead>
<tbody>
<tr>
<td>IACC.3_VGG</td>
<td>0.003</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>C3D_top1000</td>
<td>0.024</td>
<td>0.002</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>VGG_top1000*</td>
<td>0.024</td>
<td>0.016</td>
<td>0.000</td>
<td>0.006</td>
</tr>
<tr>
<td>VGG_top500</td>
<td>0.029</td>
<td><strong>0.032</strong></td>
<td>0.000</td>
<td><strong>0.010</strong></td>
</tr>
<tr>
<td>C3D+VGG_top1000*</td>
<td>0.040</td>
<td>0.017</td>
<td>0.000</td>
<td>0.002</td>
</tr>
<tr>
<td>Best System (F)**</td>
<td>0.054</td>
<td><strong>0.435</strong></td>
<td><strong>0.176</strong></td>
<td><strong>0.229</strong></td>
</tr>
</tbody>
</table>

* The system runs that we submitted
** Excluding our system runs
## AVS – Extra Experiments

506 Find shots of the 43rd president George W. Bush sitting down talking with people indoors
- Not enough data

<table>
<thead>
<tr>
<th></th>
<th>MeanxInfAP</th>
<th>506</th>
<th>513</th>
<th>522</th>
</tr>
</thead>
<tbody>
<tr>
<td>IACC.3_VGG</td>
<td>0.003</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>C3D_top1000</td>
<td>0.024</td>
<td>0.002</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>VGG_top1000*</td>
<td>0.024</td>
<td>0.016</td>
<td>0.000</td>
<td>0.006</td>
</tr>
<tr>
<td>VGG_top500</td>
<td>0.029</td>
<td><strong>0.032</strong></td>
<td>0.000</td>
<td><strong>0.010</strong></td>
</tr>
<tr>
<td>C3D+VGG_top1000*</td>
<td>0.040</td>
<td>0.017</td>
<td>0.000</td>
<td>0.002</td>
</tr>
<tr>
<td>Best System (F)**</td>
<td>0.054</td>
<td><strong>0.435</strong></td>
<td><strong>0.176</strong></td>
<td><strong>0.229</strong></td>
</tr>
</tbody>
</table>

* The system runs that we submitted
** Excluding our system runs
AVS – Extra Experiments

513 Find shots of military personnel interacting with protesters

<table>
<thead>
<tr>
<th></th>
<th>MeanxInfAP</th>
<th>506</th>
<th>513</th>
<th>522</th>
</tr>
</thead>
<tbody>
<tr>
<td>IACC.3_VGG</td>
<td>0.003</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C3D_top1000</td>
<td>0.024</td>
<td>0.002</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>VGG_top1000*</td>
<td>0.024</td>
<td>0.016</td>
<td>0.000</td>
<td>0.006</td>
</tr>
<tr>
<td>VGG_top500</td>
<td>0.029</td>
<td><strong>0.032</strong></td>
<td>0.000</td>
<td><strong>0.010</strong></td>
</tr>
<tr>
<td>C3D+VGG_top1000*</td>
<td>0.040</td>
<td>0.017</td>
<td>0.000</td>
<td>0.002</td>
</tr>
<tr>
<td>Best System (F)**</td>
<td>0.054</td>
<td><strong>0.435</strong></td>
<td><strong>0.176</strong></td>
<td><strong>0.229</strong></td>
</tr>
</tbody>
</table>

* The system runs that we submitted
** Excluding our system runs
## AVS – Extra Experiments

522 Find shots of a person sitting down with a laptop visible
- Not good for retrieval based on textual metadata

<table>
<thead>
<tr>
<th></th>
<th>MeanxInfAP</th>
<th>506</th>
<th>513</th>
<th>522</th>
</tr>
</thead>
<tbody>
<tr>
<td>IACC.3_VGG</td>
<td>0.003</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>C3D_top1000</td>
<td>0.024</td>
<td>0.002</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>VGG_top1000*</td>
<td>0.024</td>
<td>0.016</td>
<td>0.000</td>
<td>0.006</td>
</tr>
<tr>
<td>VGG_top500</td>
<td>0.029</td>
<td><strong>0.032</strong></td>
<td>0.000</td>
<td><strong>0.010</strong></td>
</tr>
<tr>
<td>C3D+VGG_top1000*</td>
<td>0.040</td>
<td>0.017</td>
<td>0.000</td>
<td>0.002</td>
</tr>
<tr>
<td>Best System (F)**</td>
<td>0.054</td>
<td><strong>0.435</strong></td>
<td><strong>0.176</strong></td>
<td><strong>0.229</strong></td>
</tr>
</tbody>
</table>

* The system runs that we submitted
** Excluding our system runs
A person sitting down with a laptop
Outline

• System Overview

• Selected Topics
  – Webly-Labeled Learning
  – Experimental Results
    • FCVID and YFCC (-)
    • AVS Extra

• Conclusions & Future Work
Conclusion & Future Work

• We present a Webly-Labeled Learning framework for video detector learning
• It utilizes prior knowledge from the Internet to allow fully automatic video query with no annotation
• In the future, we will incorporate SQG and object detection for certain type of queries
INF@TREC 2016: Surveillance Event Detection

Jia Chen¹, Jiande Sun², Yang Chen³, Alexandar Hauptmann¹
¹Carnegie Mellon University
²Shandong University
³Zhejiang University
System overview

- **Mixed strategy approach**
  - ‘Static’ actions primarily defined by key poses
    - Embrace, Pointing, Cell2Ear
  - ‘Dynamic’ action primarily defined by motions
    - Running, People meeting, ...

```
Event
---
static
(Embrace, Pointing)
dynamic
(PersonRunning, PeopleMeet,...)

Solution
---
pose appearance as object detection (Faster-RCNN)
motion feature (dense trajectory) + multi class SVM
```
Static action

- Object detection for pose overall appearance
- One model for all cameras (camera irrelevant)
- Train data
  - manually label the bounding box for the corresponding people involved in the event
  - Embrace (1,853 bounding boxes)
  - Pointing (2,518 bounding boxes)
  - Cell2Ear (1,391 bounding boxes)
Pose modeling

- Overall appearance vs key point skeleton
Unsupervised data generation for hard negative class

- Other poses are used as hard negatives
- Automatically generate labels for this negative class using a pre-trained person detector
Prediction in test stage

• predict pose on images per 10 frames (0.4s)
• threshold the score at 0.1
• average pooling score in sliding windows
  – width: 50 frames
  – stride: 50 frames
Dynamic actions (from 2015)

- **Raw feature extraction**
  - dense trajectory and improved dense trajectory
- **Feature Encoding**
  - fish vector and spatial fish vector

<table>
<thead>
<tr>
<th>tra</th>
<th>hog</th>
<th>hof</th>
<th>MBHx</th>
<th>MBHy</th>
</tr>
</thead>
<tbody>
<tr>
<td>sfv</td>
<td>fv</td>
<td>sfv</td>
<td>fv</td>
<td>sfv</td>
</tr>
<tr>
<td>sfv</td>
<td>fv</td>
<td>sfv</td>
<td>fv</td>
<td>sfv</td>
</tr>
</tbody>
</table>

- SVM as multi-class classifier (one model for one camera)
- Score fusion
Performance

• Object detection metric
  – AP is much lower than that on object detection dataset ($\geq 0.8$), e.g. MSCOCO
  – Embrace/Pointing/Cell2Ear pose is more fine-grained and much harder than person detection
  – Ratio of pos/neg in SED test data much smaller than 1:6 (1:921)

<table>
<thead>
<tr>
<th></th>
<th>mAP (1:6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Embrace</td>
<td>0.425</td>
</tr>
<tr>
<td>Pointing</td>
<td>0.263</td>
</tr>
<tr>
<td>Cell2Ear</td>
<td>0.024</td>
</tr>
</tbody>
</table>
Performance

- Event detection metric*
  - promising performance on PMiss for Embrace
  - promising performance on RFA for Cell2Ear
  - mediocre performance of Pointing on actualRFA and actual PMiss leads to worst performance on actual DCR

<table>
<thead>
<tr>
<th></th>
<th>actualDCR</th>
<th>minDCR</th>
<th>actualRFA</th>
<th>actualPMiss</th>
<th>#CorDet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell2Ear</td>
<td>0.9901</td>
<td>0.9308</td>
<td>5.57</td>
<td>0.962</td>
<td>12</td>
</tr>
<tr>
<td>Embrace</td>
<td>0.7335</td>
<td>0.7006</td>
<td>40.93</td>
<td>0.529</td>
<td>139</td>
</tr>
<tr>
<td>Pointing</td>
<td>0.9648</td>
<td>0.9550</td>
<td>22.33</td>
<td>0.853</td>
<td>254</td>
</tr>
</tbody>
</table>

*Evaluated on Eev08
Embrace case study (true positive)

predict score: 1.00

predict score: 0.71
Embrace case study (false positive)

predict score: 1.00

fusion with motion feature will help solve such cases

predict score: 0.95

3d information will help solve such cases
Pointing case study (true positive)

predict score: 1.00

predict score: 0.87
Pointing case study (false positive)

need key point information to guide the model to attend to certain regions (e.g. palm, elbow and shoulder)

need additional motion information to solve such cases
Cell2Ear case study (true positive)

predict score: 0.49

predict score: 0.25
Cell2Ear case study (false positive)

- Predict score: 0.88
- Predict score: 0.88

Need additional motion information to solve such cases.

Need key point information to guide the model to attend to certain regions (e.g. palm, elbow and shoulder).
Preliminary experiment to verify the need of skeleton key-points

- sample 900 images
  - Embrace: 150 (100 for train and 50 for test)
  - Pointing: 150 (100 for train and 50 for test)
  - Cell2Ear: 150 (100 for train and 50 for test)
  - Other: 450 (100 for train and 150 for test)
Preliminary experiment to verify the need of skeleton key-points

- Manually label the key point pose
  - Head, neck, L shoulder, R shoulder, L elbow, Relbow, L palm, R palm
Preliminary experiment to verify the need of skeleton key-points

- keypoint information alone performs 10% over appearance information alone
- keypoint position + global appearance fail to improve over key point position alone (need attention-based approach)

<table>
<thead>
<tr>
<th>feature</th>
<th>accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>keypoint position</td>
<td>66.0</td>
</tr>
<tr>
<td>appearance</td>
<td>56.3</td>
</tr>
<tr>
<td>keypoint position + appearance</td>
<td>59.3</td>
</tr>
</tbody>
</table>
Conclusion and future work

• Pose based approach for static action type is promising
• Need key-point poses for better performance

• Combining pose detection with motion
  – Using pose detection with motion features can solve some of the hard cases in single frame key pose detection alone

• 3-D reconstruction is necessary for interaction events such as Embrace