Video: Ego-centric and Summarization

Presentation: Constance Clive
Computer Science Department
University of Pittsburgh
Nonchronological Video Synopsis and Indexing

Yael Pritch, Alex Rav-Acha, Shmuel Peleg
School of Computer Science and Engineering
The Hebrew University of Jerusalem
IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008
Motivation

- Effectively summarize activities from captured surveillance video
- Address queries on generated database objects
Fig. 4. In this space-time representation of video, moving objects create the “activity tubes”. The upper part represents the original video $I$, while the lower part represents the video synopsis $S$.

(a) The shorter video synopsis $S$ is generated from the input video $I$ by including most active pixels together with their spatio-temporal neighborhood. To assure smoothness, when pixel $A$ in $S$ corresponds to pixel $B$ in $I$, their “cross border” neighbors in space as well as in time should be similar.

(b) An approximate solution can be obtained by restricting consecutive synopsis pixels to come from consecutive input pixels.
Results

• Online phase requires less than one hour to process an hour of video (for typical surveillance video)
• Queries returned on the order of minutes depending on POI (Period of Interest)
Examples

• http://www.vision.huji.ac.il/video-synopsis
Detecting Activities of Daily Living in First-Person Camera Views

Hamed Pirsiavash, Deva Ramanan
Department of Computer Science, University of California, Irvine

Figure 1: Activities of daily living (ADL) captured by a wearable camera.
Motivation

- Tele-rehabilitation
- Life-logging for patients with memory loss
- Represent complex spatial-temporal relationships between objects
- Provide a large dataset of fully annotated ADLs
Challenges
long-scale temporal structure

“Classic” data: boxing

Wearable data: making tea

Difficult for HMMs to capture long-term temporal dependencies

*slide courtesy of Piriavash and Ramanan
Features

- Identify object:
  \[ f^t_i = \max_p \text{score}^t_i(p) \]  

- Aggregate features over time:
  \[ x^0_i = \frac{1}{|T|} \sum_{t \in T} f^t_i \]
Temporal Pyramid

- Generate temporal pyramid

\[ x_{i}^{j,k} = \frac{2^{j-1}}{|T|} \sum_{t \in T_{i,k}} f_{t}^{j}; \quad \forall k \in \{1...2^{j}\} \] (4)

- Learn SVM classifiers on features for activity recognition:

\[ x = \min \left( \left[ x_{1}^{0} \ldots x_{i}^{j,k} \ldots x_{K}^{L,2^{L}} \right]^{T}, 0.01 \right) \]
Temporal pyramid
Coarse to fine correspondence matching with a multi-layer pyramid
Inspired by “Spatial Pyramid” CVPR’06 and “Pyramid Match Kernels” ICCV’05

*slide courtesy of Piriavash and Ramanan
Active Object Models

- How to tell that an open fridge and a closed fridge are the same object?
- Train an additional object detector using the subset of “active” training images for a particular object
“Passive” vs “active” objects

Passive

Active

*slide courtesy of Piriyavash and Ramanan
Dataset

- 20 people
- 30 minutes of footage a day
- 10 hours of footage per person
- 18 different identified ADLs

<table>
<thead>
<tr>
<th>action name</th>
<th>mean of length (secs)</th>
<th>std. dev. of length</th>
</tr>
</thead>
<tbody>
<tr>
<td>combing hair</td>
<td>26.50</td>
<td>9.00</td>
</tr>
<tr>
<td>make up</td>
<td>108.00</td>
<td>85.44</td>
</tr>
<tr>
<td>brushing teeth</td>
<td>128.86</td>
<td>45.50</td>
</tr>
<tr>
<td>dental floss</td>
<td>92.00</td>
<td>23.58</td>
</tr>
<tr>
<td>washing hands/face</td>
<td>76.00</td>
<td>36.33</td>
</tr>
<tr>
<td>drying hands/face</td>
<td>26.67</td>
<td>13.06</td>
</tr>
<tr>
<td>laundry</td>
<td>215.50</td>
<td>142.81</td>
</tr>
<tr>
<td>washing dishes</td>
<td>159.60</td>
<td>154.39</td>
</tr>
<tr>
<td>moving dishes</td>
<td>143.00</td>
<td>159.81</td>
</tr>
<tr>
<td>making tea</td>
<td>143.00</td>
<td>71.81</td>
</tr>
<tr>
<td>making coffee</td>
<td>85.33</td>
<td>54.45</td>
</tr>
<tr>
<td>drinking water/bottle</td>
<td>70.50</td>
<td>30.74</td>
</tr>
<tr>
<td>drinking water/tap</td>
<td>8.00</td>
<td>5.66</td>
</tr>
<tr>
<td>making cold food/snack</td>
<td>117.20</td>
<td>96.63</td>
</tr>
<tr>
<td>vacuuming</td>
<td>77.00</td>
<td>60.81</td>
</tr>
<tr>
<td>watching tv</td>
<td>189.60</td>
<td>98.74</td>
</tr>
<tr>
<td>using computer</td>
<td>105.60</td>
<td>32.94</td>
</tr>
<tr>
<td>using cell</td>
<td>18.67</td>
<td>9.45</td>
</tr>
</tbody>
</table>
ADL vs. Image-Net

<table>
<thead>
<tr>
<th>Object</th>
<th>ADL</th>
<th>ImageNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>tap</td>
<td>40.4 ± 24.3</td>
<td>0.1</td>
</tr>
<tr>
<td>soap liquid</td>
<td>32.5 ± 28.8</td>
<td>2.5</td>
</tr>
<tr>
<td>fridge</td>
<td>19.9 ± 12.6</td>
<td>0.4</td>
</tr>
<tr>
<td>microwave</td>
<td>43.1 ± 14.1</td>
<td>20.2</td>
</tr>
<tr>
<td>oven/stove</td>
<td>38.7 ± 22.3</td>
<td>0.1</td>
</tr>
<tr>
<td>bottle</td>
<td>21.0 ± 27.0</td>
<td>9.8</td>
</tr>
<tr>
<td>kettle</td>
<td>21.6 ± 24.2</td>
<td>0.1</td>
</tr>
<tr>
<td>mug/cup</td>
<td>23.5 ± 14.8</td>
<td>14.8</td>
</tr>
<tr>
<td>washer/dryer</td>
<td>47.6 ± 15.7</td>
<td>1.8</td>
</tr>
<tr>
<td>tv</td>
<td>69.0 ± 21.7</td>
<td>26.9</td>
</tr>
</tbody>
</table>
Annotation

• 10 annotators, one annotation per 30 frames (1 second
  • Action Label
  • Object bounding box
  • Object identity
  • human-object interaction

• For co-occurring actions, the shorter interrupts the longer
Figure 6: We show different kitchen scenes in our dataset. Unlike many other manually constructed action datasets, we exhibit a large variety of scenes and objects.
Figure 7: Our manually-designed functional ADL taxonomy.
Experiment

• Leave-one-out cross-validation
• Average precision
• Class confusion matrices for classification error and taxonomy-derived loss
Training

• Off-the-shelf parts model for object detection
• 24 object categories
• 1200 training instances
• Inherent differences between training datasets:
Action Recognition results

Space-time interest points (STIP)
Bag-of-objects model (O)
Active-object model (AO)
Idealized perfect object detectors (IO)
Augmented Idealized object detectors (IA+IO)
Figure 10: Confusion matrix for temporal pyramid with vision based active object detectors on pre-segmented videos. Segment classification accuracy = 40.6%.
Discussion

• Limitations?
• Future Work?