Keypoint-Based Action Recognition

Presenter: Jianchao Yang
Course Instructor: Prof. Derek Hoiem
Papers to discuss

- Behavior recognition via sparse spatio-temporal features.

- Learning realistic human actions from movies.
Behavior Recognition via Sparse Spatio-Temporal Features

• Motivated by the success application of key points in object recognition

• Designed a spatio-temporal feature for behavior recognition
Approach

• Similar to what seen in object recognition
Keypoints detection

• Extension from 2D

• Localization proceeds along the spatial dimensions $x$ and $y$, as well as the temporal dimension $t$.

• 3D corners too rare
Keypoints detection (cont’)

• Response function:

\[ R = (I \ast g \ast h_{ev})^2 + (I \ast g \ast h_{od})^2 \]

– Spatial kernel \( g(x, y; \sigma) \) is 2D Gaussian

– Temporal kernel

\[ h_{ev}(t; \tau, \omega) = -\cos(2\pi t \omega) e^{-t^2/\tau^2} \]
\[ h_{od}(t; \tau, \omega) = -\sin(2\pi t \omega) e^{-t^2/\tau^2} \]
Keypoints detection (cont’)

• Keypoints
  – By pooling maxima of the filter responses
  – Emphasize \textit{temporal} information other than spatial information
  – Strong response to periodic motions
  – Does not respond to pure translation motion
  – Totally unsupervised
Cuboid descriptor

Key points ➔ Cuboids ➔ Spatio-temporal descriptor

Descriptor:
Normalized pixel values;
Gradients;
Windowed optical flow, etc.

Transform:
Vectorize directly;
Histogram (global or local).
Cuboid descriptor (cont’)

Gradient is best!
Vectorize directly is best! ??
Behavior descriptor

Key points → Cuboids → Spatio-temporal descriptor → Behavior descriptor

Transform

Prototypes
Experiment results

- Datasets: facial expression, mouse, human actions
Experiments results (cont’)

Human Facial Expression Database.

Mouse Database.
Learning realistic human actions from movies

- Automatic annotation of human actions in video.
- Video classification by space-time features.
Bag-of-feature approach

• Extension of recent advances in bag-of-feature approaches
  – Spatial pyramid $\rightarrow$ more general spatial grids
  – Fixed weights for each pyramid level $\rightarrow$ optimized
  – Spatial grid $\rightarrow$ space-time grids
Space-time features

• Interest point detection: Harris operator
Spatio-temporal bag-of-features

• Hierarchical structure
Spatio-temporal grids
Experiment results

• Evaluation of spatio-temporal grids
Experiments results (cont’)

<table>
<thead>
<tr>
<th>Task</th>
<th>HoG BoF</th>
<th>HoF BoF</th>
<th>Best channel</th>
<th>Best combination</th>
</tr>
</thead>
<tbody>
<tr>
<td>KTH multi-class</td>
<td>81.6%</td>
<td>89.7%</td>
<td>91.1% (hof h3x1 t3)</td>
<td>91.8% (hof 1 t2, hog 1 t3)</td>
</tr>
<tr>
<td>Action AnswerPhone</td>
<td>13.4%</td>
<td>24.6%</td>
<td>26.7% (hof h3x1 t3)</td>
<td>32.1% (hof o2x2 t1, hog h3x1 t3)</td>
</tr>
<tr>
<td>Action GetOutCar</td>
<td>21.9%</td>
<td>14.9%</td>
<td>22.5% (hof o2x2 1)</td>
<td>41.5% (hof o2x2 t1, hog h3x1 t1)</td>
</tr>
<tr>
<td>Action HandShake</td>
<td>18.6%</td>
<td>12.1%</td>
<td>23.7% (hof h3x1 1)</td>
<td>32.3% (hog h3x1 t1, hog o2x2 t3)</td>
</tr>
<tr>
<td>Action HugPerson</td>
<td>29.1%</td>
<td>17.4%</td>
<td>34.9% (hog h3x1 t2)</td>
<td>40.6% (hog 1 t2, hog o2x2 t2, hog h3x1 t2)</td>
</tr>
<tr>
<td>Action Kiss</td>
<td>52.0%</td>
<td>36.5%</td>
<td>52.0% (hog 1 1)</td>
<td>53.3% (hog 1 t1, hog 1 t1, hog o2x2 t1)</td>
</tr>
<tr>
<td>Action SitDown</td>
<td>29.1%</td>
<td>20.7%</td>
<td>37.8% (hog 1 t2)</td>
<td>38.6% (hog 1 t2, hog 1 t3)</td>
</tr>
<tr>
<td>Action SitUp</td>
<td>6.5%</td>
<td>5.7%</td>
<td>15.2% (hog h3x1 t2)</td>
<td>18.2% (hog o2x2 t1, hog o2x2 t2, hog h3x1 t2)</td>
</tr>
<tr>
<td>Action StandUp</td>
<td>45.4%</td>
<td>40.0%</td>
<td>45.4% (hog 1 1)</td>
<td>50.5% (hog 1 t1, hog 1 t2)</td>
</tr>
</tbody>
</table>
Experiment results (cont’)

- KTH action database

Walking  Jogging  Running  Boxing  Waving  Clapping
Experiment results (cont’)

<table>
<thead>
<tr>
<th>Action</th>
<th>Clean</th>
<th>Automatic</th>
<th>Chance</th>
</tr>
</thead>
<tbody>
<tr>
<td>AnswerPhone</td>
<td>32.1%</td>
<td>16.4%</td>
<td>10.6%</td>
</tr>
<tr>
<td>GetOutCar</td>
<td>41.5%</td>
<td>16.4%</td>
<td>6.0%</td>
</tr>
<tr>
<td>HandShake</td>
<td>32.3%</td>
<td>9.9%</td>
<td>8.8%</td>
</tr>
<tr>
<td>HugPerson</td>
<td>40.6%</td>
<td>26.8%</td>
<td>10.1%</td>
</tr>
<tr>
<td>Kiss</td>
<td>53.3%</td>
<td>45.1%</td>
<td>23.5%</td>
</tr>
<tr>
<td>SitDown</td>
<td>38.6%</td>
<td>24.8%</td>
<td>13.8%</td>
</tr>
<tr>
<td>SitUp</td>
<td>18.2%</td>
<td>10.4%</td>
<td>4.6%</td>
</tr>
<tr>
<td>StandUp</td>
<td>50.5%</td>
<td>33.6%</td>
<td>22.6%</td>
</tr>
</tbody>
</table>
Comments

• The two methods are extensions of key-points based image classification. Will dense descriptors be better?

• Key-points based methods work surprisingly well for image and sequence classification, why?

• Issues needed to address:
  – Discriminative key-points learning or design for the given task
  – Discriminative key-points selection for the given task
  – More efficient way to use location information