Activity Recognition

Yimeng Zhang

11/29/11

Adapted Slides from Efros’ and Laptev’s talks
Motivation

Activity
- Verb/predicate (object: noun) $\rightarrow$ human actions
- Usually detected from a video

Applications
- Content-based browsing
  - e.g. fast-forward to the next goal scoring scene
  - e.g. find “Bush shaking hands with Putin”
- Video Surveillance
  - Monitor the crime related activities
- Human scientists
  - influence of smoking in movies on adolescent smoking
KTH Activity Dataset
Hollywood Movie Dataset – v2
Early work: holistic model

- [Efros et al. ICCV 03]
- Tracking the person
Figure-centric Representation

- Stabilized spatio-temporal volume
- No translation information
- All motion caused by person’s limbs
Remembrance of Things Past

- “Explain” novel motion sequence by matching to previously seen video clips
- For each frame, match based on some temporal extent

**Challenge:** how to compare motions?
Spatial Motion Descriptor

Image frame

Optical flow

\( F_{x,y} \)

\( F_x, F_y \)

\( F_x^-, F_x^+, F_y^-, F_y^+ \)

blurred

\( F_x^-, F_x^+, F_y^-, F_y^+ \)
Football Actions: matching

Input Sequence

Matched Frames

input  matched
Classifying Tennis Actions

6 actions; 4600 frames; 7-frame motion descriptor
Woman player used as training, man as testing.
Holistic Model

- Advantage
  - Rich spatial modeling of different body parts
  - High discrimination

- Disadvantage
  - Require background subtraction
  - Require robust person tracking
  - Does not generalize well
Body Part based Model

- [Ali et al. ICCV 07]
Body Part Trajectories

Time series analysis of the trajectories
Problems of Holistic and Body Part Methods

Holistic or Body Part Methods:
- Camera stabilization
- Segmentation
- Tracking

Common problems:
- Complex & changing BG
- Appearance of new OBJ

[Laptev et al. CVPR08]
Activity Dataset “in the wild”

Laptev et al. CVPR 2008
What are human actions?

- Actions in current datasets:
- Actions “In the Wild”:

KTH action dataset
Actions in movies

- Realistic variation of human actions
- Many classes and many examples per class

Problems:
- Typically only a few class-samples per movie
- Manual annotation is very time consuming
Why weren't you honest with me? Why'd you keep your marriage a secret?

It wasn't my secret, Richard. Victor wanted it that way. Not even our closest friends knew about our marriage.
Script alignment: Evaluation

- Annotate action samples \textit{in text}
- Do automatic script-to-video alignment
- Check the correspondence of actions in scripts and movies

Example of a “visual false positive”

A black car pulls up, two army officers get out.
Text-based action retrieval

- Large variation of action expressions in text:
  - GetOutCar action: “... Will gets out of the Chevrolet. ...” “... Erin exits her new truck...”
  - Potential false positives: “...About to sit down, he freezes...”

- => Supervised text classification approach

Keywords action retrieval from scripts

Regularized Perceptron action retrieval from scripts
Movie actions dataset

<table>
<thead>
<tr>
<th>Action</th>
<th>Correct All</th>
<th>False</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;AnswerPhone&gt;</td>
<td>20</td>
<td>5</td>
<td>96</td>
</tr>
<tr>
<td>&lt;GetOutCar&gt;</td>
<td>12</td>
<td>6</td>
<td>143</td>
</tr>
<tr>
<td>&lt;HandShake&gt;</td>
<td>23</td>
<td>15</td>
<td>239</td>
</tr>
<tr>
<td>&lt;HugPerson&gt;</td>
<td>14</td>
<td>9</td>
<td>91</td>
</tr>
<tr>
<td>&lt;Kiss&gt;</td>
<td>20</td>
<td>7</td>
<td>51</td>
</tr>
<tr>
<td>&lt;SitDown&gt;</td>
<td>12</td>
<td>21</td>
<td>62</td>
</tr>
<tr>
<td>&lt;SitUp&gt;</td>
<td>15</td>
<td>5</td>
<td>29</td>
</tr>
<tr>
<td>&lt;StandUp&gt;</td>
<td>23</td>
<td>10</td>
<td>48</td>
</tr>
</tbody>
</table>

- **20 different movies**
- **12 movies**

Automatically labeled training set:
- 22, 13, 20, 22, 49, 47, 11, 48, 232

Manually labeled training set:
- 23, 13, 19, 22, 51, 30, 10, 49, 217

Test set:
- Hollywood-2 dataset: >1700 video clips of 12 categories
Training noise robustness

Figure: Performance of our video classification approach in the presence of wrong labels

- Up to $p=0.2$ the performance decreases insignificantly
- At $p=0.4$ the performance decreases by around 10%
### Hollywood Movie Result

<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>

**Table**: Average precision (AP) for each action class of our test set. We compare results for clean (annotated) and automatic training data. We also show results for a random classifier (chance)

[Laptev et al. CVPR08]
Space-Time Local Features

[Dollar et al. PETS Workshop 2005]
[Laptev et al. ICCV 03, CVPR08]
[Willems et al. ECCV 08]
[Wang et al. BMVC 09]
Space-time Local Features

Consider local spatio-temporal neighborhoods

[Dollar et al. PETS Workshop 2005]
[Laptev et al. ICCV 03, CVPR08]
[Wang et al. BMVC 09]
2D→3D Local Features

- Motivation:
  - *Sparse feature points* extended to the spatio-temporal case
Object Recognition

Advantages of Sparse Features

- Robustness
- Very good results

example from: http://www.robots.ox.ac.uk/~fergus/research/index.html
Bag-of-Words for Object Recognition

feature detection & representation

image representation

Length: dictionary size

Adopted figures from slides of Feifei, Li
Space-time Bag-of-Words

Bag of space-time features + multi-channel SVM

[Schuldt’04, Niebles’06, Zhang’07]

Feature detector

Collection of space-time patches

Descriptors

Histogram of visual words

Multi-channel SVM Classifier
Space-Time Feature: Detector (1)

- **Harris-3D**
  - [Laptev et al., ICCV 03]
  - Space-time corner detector

\[
H = \text{det}(\mu) + k \text{tr}^3(\mu)
\]

\[
\mu = \begin{pmatrix}
I_xI_x & I_xI_y & I_xI_t \\
I_xI_y & I_yI_y & I_yI_t \\
I_xI_t & I_yI_t & I_tI_t
\end{pmatrix} \ast g(\cdot; \sigma, \tau)
\]
Space-Time Feature: Detector (2)

- Cuboids
  - [Dollar et al. 2005]
- Response Function
  \[ R = (I * g * h_{ev})^2 + (I * g * h_{od})^2 \]
- Spatial Filter: Gaussian
- Temporal Filter: Gabor
  \[ h_{ev}(t; \tau, \omega) = -\cos(2\pi t \omega)e^{-t^2/\tau^2} \]
  \[ h_{od}(t; \tau, \omega) = -\sin(2\pi t \omega)e^{-t^2/\tau^2} \]
Space-Time Feature: Detector (3)

- Hessian – 3D
  - [Willems et al. ECCV 08]
  - Blob-like features

- Dense sampling [Wang et al. BMVC 09]

\[(\sigma^2, \tau^2) = S \times T, \ S = 2^{\{2,\ldots,6\}}, \ T = 2^{\{1,2\}}\]
Space-Time Feature: Detector (4)
Space-Time Features: Descriptor

Multi-scale space-time patches from corner detector

Histogram of oriented spatial grad. (HOG)

Histogram of optical flow (HOF)

3x3x2x4bins HOG descriptor

3x3x2x5bins HOF descriptor

Public code available at www.irisa.fr/vista/actions
## Comparison - KTH

2391 video clips of 6 categories

<table>
<thead>
<tr>
<th></th>
<th>walking</th>
<th>running</th>
<th>jogging</th>
<th>handwaving</th>
<th>handclapping</th>
<th>boxing</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Image](30x255 to 138x336)</td>
<td>[Image](144x255 to 252x336)</td>
<td>[Image](258x255 to 366x336)</td>
<td>[Image](372x257 to 474x339)</td>
<td>[Image](480x258 to 582x334)</td>
<td>[Image](588x258 to 690x334)</td>
<td>[Image](23x79 to 720x216)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>HOG3D</th>
<th>HOG/HOF</th>
<th>HOG</th>
<th>HOF</th>
<th>Cuboids</th>
<th>ESURF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harris3D</td>
<td>89.0%</td>
<td>91.8%</td>
<td>80.9%</td>
<td>92.1%</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Cuboids</td>
<td>90.0%</td>
<td>88.7%</td>
<td>82.3%</td>
<td>88.2%</td>
<td>89.1%</td>
<td>–</td>
</tr>
<tr>
<td>Hessian</td>
<td>84.6%</td>
<td>88.7%</td>
<td>77.7%</td>
<td>88.6%</td>
<td>–</td>
<td>81.4%</td>
</tr>
<tr>
<td>Dense</td>
<td>85.3%</td>
<td>86.1%</td>
<td>79.0%</td>
<td>88.0%</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

[Wang et al. BMVC 09]
Comparison – UCF Sport

150 video clips of 10 categories

<table>
<thead>
<tr>
<th></th>
<th>HOG3D</th>
<th>HOG/HOF</th>
<th>HOG</th>
<th>HOF</th>
<th>Cuboids</th>
<th>ESURF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harris3D</td>
<td>79.7%</td>
<td>78.1%</td>
<td>71.4%</td>
<td>75.4%</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Cuboids</td>
<td>82.9%</td>
<td>77.7%</td>
<td>72.7%</td>
<td>76.7%</td>
<td>76.6%</td>
<td>–</td>
</tr>
<tr>
<td>Hessian</td>
<td>79.0%</td>
<td>79.3%</td>
<td>66.0%</td>
<td>75.3%</td>
<td>–</td>
<td>77.3%</td>
</tr>
<tr>
<td>Dense</td>
<td>85.6%</td>
<td>81.6%</td>
<td>77.4%</td>
<td>82.6%</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

[Wang et al. BMVC 09]
Comparison – Hollywood-2 Dataset

1707 video clips of 12 actions

<table>
<thead>
<tr>
<th></th>
<th>HOG3D</th>
<th>HOG/HOF</th>
<th>HOG</th>
<th>HOF</th>
<th>Cuboids</th>
<th>ESURF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harris3D</td>
<td>43.7%</td>
<td>45.2%</td>
<td>32.8%</td>
<td>43.3%</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Cuboids</td>
<td>45.7%</td>
<td>46.2%</td>
<td>39.4%</td>
<td>42.9%</td>
<td>45.0%</td>
<td>–</td>
</tr>
<tr>
<td>Hessian</td>
<td>41.3%</td>
<td>46.0%</td>
<td>36.2%</td>
<td>43.0%</td>
<td>–</td>
<td>38.2%</td>
</tr>
<tr>
<td>Dense</td>
<td>45.3%</td>
<td>47.4%</td>
<td>39.4%</td>
<td>45.5%</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

[Wang et al. BMVC 09]
Long-range Spatio-Temporal Information
Spatio-temporal Pyramid Matching

We use global spatio-temporal grids

- In the spatial domain:
  - 1x1 (standard BoF)
  - 2x2, o2x2 (50% overlap)
  - h3x1 (horizontal), v1x3 (vertical)
  - 3x3

- In the temporal domain:
  - t1 (standard BoF), t2, t3

Figure: Examples of a few spatio-temporal grids

[Laptev et al. CVPR08]
Multi-channel chi-square kernel

We use SVMs with a multi-channel chi-square kernel for classification

\[ K(H_i, H_j) = \exp \left( - \sum_{c \in C} \frac{1}{A_c} D_c(H_i, H_j) \right) \]

- Channel \( c \) is a combination of a detector, descriptor and a grid
- \( D_c(H_i, H_j) \) is the chi-square distance between histograms
- The best set of channels \( C \) for a given training set is found based on a greedy approach
Combining channels

<table>
<thead>
<tr>
<th>Task</th>
<th>HoG BoF</th>
<th>HoF BoF</th>
<th>Best chan.</th>
<th>Best comb.</th>
</tr>
</thead>
<tbody>
<tr>
<td>KTH multi-class</td>
<td>81.6%</td>
<td>89.7%</td>
<td>91.1%</td>
<td>91.8%</td>
</tr>
<tr>
<td>Action AnswerPhone</td>
<td>13.4%</td>
<td>24.6%</td>
<td>26.7%</td>
<td>32.1%</td>
</tr>
<tr>
<td>Action GetOutCar</td>
<td>21.9%</td>
<td>14.9%</td>
<td>22.5%</td>
<td>41.5%</td>
</tr>
<tr>
<td>Action HandShake</td>
<td>18.6%</td>
<td>12.1%</td>
<td>23.7%</td>
<td>32.3%</td>
</tr>
<tr>
<td>Action HugPerson</td>
<td>29.1%</td>
<td>17.4%</td>
<td>34.9%</td>
<td>40.6%</td>
</tr>
<tr>
<td>Action Kiss</td>
<td>52.0%</td>
<td>36.5%</td>
<td>52.0%</td>
<td>53.3%</td>
</tr>
<tr>
<td>Action SitDown</td>
<td>29.1%</td>
<td>20.7%</td>
<td>37.8%</td>
<td>38.6%</td>
</tr>
<tr>
<td>Action SitUp</td>
<td>6.5%</td>
<td>5.7%</td>
<td>15.2%</td>
<td>18.2%</td>
</tr>
<tr>
<td>Action StandUp</td>
<td>45.4%</td>
<td>40.0%</td>
<td>45.4%</td>
<td>50.5%</td>
</tr>
</tbody>
</table>

**Table:** Classification performance of different channels and their combinations

- It is worth trying different grids
- It is beneficial to combine channels
Aligned Space-Time Pyramid Matching

[Duan et al. CVPR 10]

Find best matching of a binary graph
### Aligned Pyramid Matching Result

Table 1. Means and standard deviations (%) of MAPs at different levels using SVM with the default kernel parameter for SIFT features.

<table>
<thead>
<tr>
<th></th>
<th>Gaussian</th>
<th>Laplacian</th>
<th>ISD</th>
<th>ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level-0</td>
<td>41.4 ± 3.7</td>
<td>44.2 ± 3.8</td>
<td>45.0 ± 3.5</td>
<td>46.2 ± 4.0</td>
</tr>
<tr>
<td>Level-1 (Unaligned)</td>
<td>43.0 ± 2.7</td>
<td>47.7 ± 1.7</td>
<td>49.0 ± 1.6</td>
<td>48.2 ± 1.5</td>
</tr>
<tr>
<td>Level-1 (Aligned)</td>
<td>50.4 ± 3.7</td>
<td>53.8 ± 1.8</td>
<td>52.9 ± 3.6</td>
<td>51.0 ± 2.5</td>
</tr>
</tbody>
</table>

Kodak Event Dataset: 1358 video clips of 6 Events
Trajectory of Local Features

- [Messing et al. ICCV 09]
- Track local space-time features
Feature Flow
Figure 1. Graphical model for our tracked keypoint velocity history model (Dark circles denote observed variables).

\[
P(A, O) = \sum_M P(A, M, O) = \\
= P(A) \prod_{f} \sum_{i} P(M_f^i | A) P(O_{0,f} | M_f^i) \\
= \prod_{t=1}^{T_f} P(O_{t,f} | O_{t-1,f}, M_f^i)
\]
Trajectory words
Trajectory of Local Features Result

- Daily Living Dataset
## Trajectory of Local Features Result

<table>
<thead>
<tr>
<th>Method</th>
<th>Percent Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temporal Templates [6]</td>
<td>33</td>
</tr>
<tr>
<td>Spatio-Temporal Cuboids [7]</td>
<td>36</td>
</tr>
<tr>
<td>Space-Time Interest Points [12]</td>
<td>59</td>
</tr>
<tr>
<td>Velocity Histories (Sec. 3)</td>
<td>63</td>
</tr>
<tr>
<td>Latent Velocity Histories (Sec. 7)</td>
<td>67</td>
</tr>
<tr>
<td>Augmented Velocity Histories (Sec. 6)</td>
<td>89</td>
</tr>
</tbody>
</table>
String of Feature Graphs

- [Gaur et al. ICCV 11]
- Interactive activities among different people
Match Feature Graphs

Matching individual feature collections

Query feature collection

Dataset feature collection

Local matching scores

Matching video-strings using DTW

query video-string

dataset video-string
Interactive Activities

Interactive Activities:
shake hands, hug, kick, point, push, punch
Example Results

<table>
<thead>
<tr>
<th>Query</th>
<th>Retrieved Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hand shake</td>
<td></td>
</tr>
<tr>
<td>Punching</td>
<td></td>
</tr>
<tr>
<td>Hugging</td>
<td></td>
</tr>
<tr>
<td>Pointing</td>
<td></td>
</tr>
<tr>
<td>Kicking</td>
<td></td>
</tr>
</tbody>
</table>
Group-level activities
Conclusion

- Activity Recognition
  - Single Person
  - Multiple Persons
- Environments
  - Controlled environment, Stable cameras
  - Complex scenes – youtube, movie
- Approach
  - Holistic / Body part
  - Space-time local features
  - Incorporate long-range dependencies
Discussion

• Need for a large dataset of more activities
  • Current dataset: around 10 activity categories
  • ActivityNet?
  • A hierarchical dataset: high jump, long jump, ski jump

• Current algorithm is far from perfect
  • More suitable features?

• Speed is important
  • Avoid processing every frame
Events in Crowd

Images from [Wu et al. CVPR 2010]
Thank you