Motion and Human Actions
Class overview

Motivation
- Historic review
- Modern applications

Appearance-based methods
- Motion history images
- Active shape models
- Tracking and motion priors

Motion-based methods
- Generic and parametric Optical Flow
- Motion templates

Space-time methods
- Local space-time features
- Action classification and detection
- Weakly-supervised action learning
Motivation

Goal: Interpreting complex dynamic scenes

Common methods:
• Segmentation
• Tracking

Common problems:
• Complex & changing BG
• Changing appearance

⇒ No global assumptions about the scene
Space-time

No global assumptions ⇒
Consider local spatio-temporal neighborhoods
Actions == Space-time objects?
Space-time local features
Space-Time Interest Points: Detection

What neighborhoods to consider?

Distinctive neighborhoods \implies High image variation in space and time \implies Look at the distribution of the gradient

Definitions:

\[ f : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R} \]  
Original image sequence

\[ g(x, y, t; \Sigma) \]  
Space-time Gaussian with covariance

\[ L_\xi(\cdot; \Sigma) = f(\cdot) \ast g_\xi(\cdot; \Sigma) \]  
Gaussian derivative of \( f \)

\[ \nabla L = (L_x, L_y, L_t)^T \]  
Space-time gradient

\[ \mu(\cdot; \Sigma) = \nabla L(\cdot; \Sigma)(\nabla L(\cdot; \Sigma))^T \ast g(\cdot; s\Sigma) = \begin{pmatrix} \mu_{xx} & \mu_{xy} & \mu_{xt} \\ \mu_{xy} & \mu_{yy} & \mu_{yt} \\ \mu_{xt} & \mu_{yt} & \mu_{tt} \end{pmatrix} \]  
Second-moment matrix
Space-Time Interest Points: Detection

Properties of \( \mu(\cdot; \Sigma) \)

\( \mu(\cdot; \Sigma) \) defines second order approximation for the local distribution of \( \nabla L \) within neighborhood \( \Sigma \)

- \( \text{rank}(\mu) = 1 \Rightarrow 1\text{D space-time variation of } f \text{ e.g. moving bar} \)
- \( \text{rank}(\mu) = 2 \Rightarrow 2\text{D space-time variation of } f \text{ e.g. moving ball} \)
- \( \text{rank}(\mu) = 3 \Rightarrow 3\text{D space-time variation of } f \text{ e.g. jumping ball} \)

Large eigenvalues of \( \mu \) can be detected by the local maxima of \( H \) over \((x,y,t)\):

\[
H(p; \Sigma) = \det(\mu(p; \Sigma)) + k \text{trace}^3(\mu(p; \Sigma)) \\
= \lambda_1 \lambda_2 \lambda_3 - k(\lambda_1 + \lambda_2 + \lambda_3)^3
\]

(similar to Harris operator [Harris and Stephens, 1988])
Space-Time Interest Points: Examples

Motion event detection
Local features for human actions
Local features for human actions

- boxing
- walking
- hand waving
Multi-scale space-time patches

Local space-time descriptor: HOG/HOF

Histogram of oriented spatial grad. (HOG)
3x3x2x4bins HOG descriptor

Histogram of optical flow (HOF)
3x3x2x5bins HOF descriptor
Visual Vocabulary: K-means clustering

- Group similar points in the space of image descriptors using K-means clustering
- Select significant clusters
Visual Vocabulary: K-means clustering

- Group similar points in the space of image descriptors using K-means clustering
- Select significant clusters
Local features: Matching

- Finds similar events in pairs of video sequences
Action Classification

Bag of space-time features + multi-channel SVM

[Laptev’03, Schuldt’04, Niebles’06, Zhang’07]
Action classification results

KTH dataset

Hollywood-2 dataset

<table>
<thead>
<tr>
<th>Channel</th>
<th>hoghof</th>
<th>Chance</th>
</tr>
</thead>
<tbody>
<tr>
<td>mAP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AnswerPhone</td>
<td>15.7</td>
<td>7.2</td>
</tr>
<tr>
<td>DriveCar</td>
<td>86.6</td>
<td>11.5</td>
</tr>
<tr>
<td>Eat</td>
<td>59.5</td>
<td>3.7</td>
</tr>
<tr>
<td>FightPerson</td>
<td>71.1</td>
<td>7.9</td>
</tr>
<tr>
<td>GetOutCar</td>
<td>29.3</td>
<td>6.4</td>
</tr>
<tr>
<td>HandShake</td>
<td>21.2</td>
<td>5.1</td>
</tr>
<tr>
<td>HugPerson</td>
<td>35.8</td>
<td>7.5</td>
</tr>
<tr>
<td>Kiss</td>
<td>51.5</td>
<td>11.7</td>
</tr>
<tr>
<td>Run</td>
<td>69.1</td>
<td>16.0</td>
</tr>
<tr>
<td>SitDown</td>
<td>58.2</td>
<td>12.2</td>
</tr>
<tr>
<td>SitUp</td>
<td>17.5</td>
<td>4.2</td>
</tr>
<tr>
<td>StandUp</td>
<td>51.7</td>
<td>16.5</td>
</tr>
</tbody>
</table>

[Laptev, Marszałek, Schmid, Rozenfeld 2008]
Action classification

Test episodes from movies “The Graduate”, “It’s a Wonderful Life”, “Indiana Jones and the Last Crusade”
Evaluation of local feature detectors and descriptors

Four types of detectors:
- Harris3D [Laptev 2003]
- Cuboids [Dollar et al. 2005]
- Hessian [Willems et al. 2008]
- Regular dense sampling

Four types of descriptors:
- HoG/HoF [Laptev et al. 2008]
- Cuboids [Dollar et al. 2005]
- HoG3D [Kläser et al. 2008]
- Extended SURF [Willems’et al. 2008]

Three human actions datasets:
- KTH actions [Schuldt et al. 2004]
- UCF Sports [Rodriguez et al. 2008]
Space-time feature detectors

Harris3D

Cuboids

Hessian

Dense
## Results on Hollywood-2

12 action classes collected from 69 movies

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

(Average precision scores)

- Best results for **dense** + HOG/HOF

[Wang, Ullah, Kläser, Laptev, Schmid, 2009]
Other recent local representations

- Y. and L. Wolf, "Local Trinary Patterns for Human Action Recognition", ICCV 2009

- P. Matikainen, R. Sukthankar and M. Hebert "Trajectons: Action Recognition Through the Motion Analysis of Tracked Features" ICCV VOEC Workshop 2009,

Dense trajectories [Wang et al. IJCV’13]

- Dense sampling
- Feature tracking based on optical flow
- Trajectory-aligned descriptors
Trajectory descriptors

Motion boundary descriptor
– spatial derivatives are calculated separately for optical flow in x and y, quantized into a histogram
– relative dynamics of different regions
– suppresses constant motions
Dense trajectories

- Advantages:
  - Captures the intrinsic dynamic structures in videos
  - MBH is robust to certain camera motion

- Disadvantages:
  - Generates irrelevant trajectories in background due to camera motion
  - Motion descriptors are modified by camera motion, e.g., HOF, MBH
**Improved dense trajectories** [Wang et al. ICCV’13]

- Improve dense trajectories by explicit camera motion estimation
- Detect humans to remove outlier matches for homography estimation
- Stabilize optical flow to eliminate camera motion
Camera motion estimation

- Find the correspondences between two consecutive frames:
  - Extract and match SURF features (robust to motion blur)
  - Use optical flow, remove uninformative points
- Combine SURF (green) and optical flow (red) results in a more balanced distribution
- Use RANSAC to estimate a homography from all feature matches

Inlier matches of the homography
Remove inconsistent matches due to humans

- Human motion is not constrained by camera motion, thus generates outlier matches
- Apply a human detector in each frame, and track the human bounding box forward and backward to join detections
- Remove feature matches inside the human bounding box during homography estimation

Inlier matches and warped flow, without or with HD
Remove background trajectories

- Remove trajectories by thresholding the maximal magnitude of stabilized motion vectors
- Our method works well under various camera motions, such as pan, zoom, tilt

**Successful examples**

**Failure cases**

- Failure due to severe motion blur; the homography is not correctly estimated due to unreliable feature matches

Removed trajectories (white) and foreground ones (green)
Experimental setting

- Motion stabilized trajectories and features (HOG, HOF, MBH)
- Square-root normalization for each descriptor, then PCA to reduce its dimension by a factor of two
- Use Fisher vector to encode each descriptor separately, set the number of Gaussians to K=256
- Use Power+L2 normalization for FV, and linear SVM with one-against-rest for multi-class classification

Datasets

- Hollywood2: 12 classes from 69 movies, report mAP
- HMDB51: 51 classes, report accuracy on three splits
- Olympic sports: 16 sport actions, report mAP
- UCF50: 50 classes, report accuracy over 25 groups
Evaluation of the intermediate steps

<table>
<thead>
<tr>
<th></th>
<th>HOG</th>
<th>HOF</th>
<th>MBH</th>
<th>HOF+MBH</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>DTF</td>
<td>38.4%</td>
<td>39.5%</td>
<td>49.1%</td>
<td>49.8%</td>
<td>52.2%</td>
</tr>
<tr>
<td>ITF</td>
<td>40.2%</td>
<td>48.9%</td>
<td>52.1%</td>
<td>54.7%</td>
<td>57.2%</td>
</tr>
</tbody>
</table>

Results on HMDB51 using Fisher vector

- Baseline: DTF = "dense trajectory feature"
- ITF = "improved trajectory feature"
- HOF improves significantly and MBH somewhat
- Almost no impact on HOG
- HOF and MBH are complementary, as they represent zero and first order motion information
Impact of feature encoding on improved trajectories

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Bag of features</th>
<th>Fisher vector</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DTF</td>
<td>ITF</td>
</tr>
<tr>
<td>Hollywood2</td>
<td>58.5%</td>
<td>62.2%</td>
</tr>
<tr>
<td>HMDB51</td>
<td>47.2%</td>
<td>52.1%</td>
</tr>
<tr>
<td>Olympic Sport</td>
<td>75.4%</td>
<td>83.3%</td>
</tr>
<tr>
<td>UCF50</td>
<td>84.8%</td>
<td>87.2%</td>
</tr>
</tbody>
</table>

Compare DTF and ITF using different feature encoding

- Standard bag of features: train a codebook of 4000 visual words with k-means for each descriptor type; RBF-$\chi^2$ kernel SVM for classification
- We observe a similar improvement of ITF over DTF when using BOF or FV for feature encoding
- The improvement of FV over BOF varies from 2% to 7% depending on the dataset
Impact of human detection and state of the art

- Human detection always helps. For Hollywood2 and HMDB51, the difference is more significant, as there are more humans present.

- Significantly outperforms the state of the art on all four datasets.
Activity recognition

- Complex events, i.e. making a sandwich, doing homework

Making sandwich

Doing homework

TrecVid Multi-media event detection dataset
Activity recognition

- Complex events, i.e. birthday party, parade

Birthday party

Parade

TrecVid Multi-media event detection dataset
TrecVid MED’13

- 100 positive video clips per event category, 5000 negative video clips
- Testing on 98000 videos clips, i.e., 4000 hours
- 20 known events, 10 adhoc events
- Videos come from publicly available, user-generated content on various Internet sites

- Descriptors: MBH, SIFT, audio, text & speech recognition
## Quantitative results on TrecVid MED’11

<table>
<thead>
<tr>
<th>Channel</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motion</td>
<td>44.65</td>
</tr>
<tr>
<td>Static</td>
<td>33.97</td>
</tr>
<tr>
<td>Audio</td>
<td>18.15</td>
</tr>
<tr>
<td>OCR</td>
<td>10.85</td>
</tr>
<tr>
<td>ASR</td>
<td>8.21</td>
</tr>
<tr>
<td>Visual=Motion+Static</td>
<td>47.22</td>
</tr>
<tr>
<td>Visual+Audio</td>
<td>50.41</td>
</tr>
<tr>
<td>Visual+OCR</td>
<td>48.97</td>
</tr>
<tr>
<td>Visual+ASR</td>
<td>48.28</td>
</tr>
<tr>
<td>Visual+Audio+OCR+ASR</td>
<td>52.28</td>
</tr>
</tbody>
</table>
## Quantitative results on TrecVid MED’11

<table>
<thead>
<tr>
<th>Channel</th>
<th>mAP</th>
<th>Birthday party</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motion</td>
<td>44.65</td>
<td>30.7</td>
</tr>
<tr>
<td>Static</td>
<td>33.97</td>
<td>25.9</td>
</tr>
<tr>
<td>Audio</td>
<td>18.15</td>
<td>33.3</td>
</tr>
<tr>
<td>OCR</td>
<td>10.85</td>
<td>10.1</td>
</tr>
<tr>
<td>ASR</td>
<td>8.21</td>
<td>3.6</td>
</tr>
<tr>
<td>Visual=Motion+Static</td>
<td>47.22</td>
<td>34.8</td>
</tr>
<tr>
<td>Visual+Audio</td>
<td>50.41</td>
<td>47.7</td>
</tr>
<tr>
<td>Visual+OCR</td>
<td>48.97</td>
<td>35.8</td>
</tr>
<tr>
<td>Visual+ASR</td>
<td>48.28</td>
<td>35.0</td>
</tr>
</tbody>
</table>
| Visual+Audio+OCR+ASR     | 52.28| 48.4          
## Quantitative results on TrecVid MED’11

<table>
<thead>
<tr>
<th>Channel</th>
<th>mAP</th>
<th>Birthday party</th>
<th>Repair appliance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motion</td>
<td>44.65</td>
<td>30.7</td>
<td>42.6</td>
</tr>
<tr>
<td>Static</td>
<td>33.97</td>
<td>25.9</td>
<td>43.6</td>
</tr>
<tr>
<td>Audio</td>
<td>18.15</td>
<td>33.3</td>
<td>43.3</td>
</tr>
<tr>
<td>OCR</td>
<td>10.85</td>
<td>10.1</td>
<td>32.1</td>
</tr>
<tr>
<td>ASR</td>
<td>8.21</td>
<td>3.6</td>
<td>39.2</td>
</tr>
<tr>
<td>Visual=Motion+Static</td>
<td>47.22</td>
<td>34.8</td>
<td>47.5</td>
</tr>
<tr>
<td>Visual+Audio</td>
<td>50.41</td>
<td>47.7</td>
<td>54.5</td>
</tr>
<tr>
<td>Visual+OCR</td>
<td>48.97</td>
<td>35.8</td>
<td>50.8</td>
</tr>
<tr>
<td>Visual+ASR</td>
<td>48.28</td>
<td>35.0</td>
<td>54.5</td>
</tr>
<tr>
<td>Visual+Audio+OCR+ASR</td>
<td>52.28</td>
<td>48.4</td>
<td>57.2</td>
</tr>
</tbody>
</table>
## Quantitative results on TrecVid MED’11

<table>
<thead>
<tr>
<th>Channel</th>
<th>mAP</th>
<th>Birthday party</th>
<th>Repair appliance</th>
<th>Make sandwich</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motion</td>
<td>44.65</td>
<td>30.7</td>
<td>42.6</td>
<td>22.5</td>
</tr>
<tr>
<td>Static</td>
<td>33.97</td>
<td>25.9</td>
<td>43.6</td>
<td>21.5</td>
</tr>
<tr>
<td>Audio</td>
<td>18.15</td>
<td>33.3</td>
<td>43.3</td>
<td>11.2</td>
</tr>
<tr>
<td>OCR</td>
<td>10.85</td>
<td>10.1</td>
<td>32.1</td>
<td>19.4</td>
</tr>
<tr>
<td>ASR</td>
<td>8.21</td>
<td>3.6</td>
<td>39.2</td>
<td>6.7</td>
</tr>
<tr>
<td>Visual=Motion+Static</td>
<td>47.22</td>
<td>34.8</td>
<td>47.5</td>
<td>27.8</td>
</tr>
<tr>
<td>Visual+Audio</td>
<td>50.41</td>
<td>47.7</td>
<td>54.5</td>
<td>27.3</td>
</tr>
<tr>
<td>Visual+OCR</td>
<td>48.97</td>
<td>35.8</td>
<td>50.8</td>
<td>35.7</td>
</tr>
<tr>
<td>Visual+ASR</td>
<td>48.28</td>
<td>35.0</td>
<td>54.5</td>
<td>28.8</td>
</tr>
<tr>
<td>Visual+Audio+OCR+ASR</td>
<td>52.28</td>
<td>48.4</td>
<td>57.2</td>
<td>35.4</td>
</tr>
</tbody>
</table>
TrecVid MED 2013 – example results

Horse riding competition

rank 1

rank 2

rank 3
TrecVid MED 2013 – example results

Tuning a musical instrument

rank 1

rank 2

rank 3

Tuning a lever harp to the key of E Flat Major
Recent CNN methods

Two-Stream Convolutional Networks for Action Recognition in Videos [Simonyan and Zisserman NIPS14]

Learning Spatiotemporal Features with 3D Convolutional Networks [Tran et al. ICCV15]

Action recognition with trajectory pooled convolutional descriptors [Wang et al. CVPR15]
Recent CNN methods

Two-Stream Convolutional Networks for Action Recognition in Videos
[Simonyan and Zisserman NIPS14]
Recent CNN methods

Learning Spatiotemporal Features with 3D Convolutional Networks
[Tran et al. ICCV15]

Figure 1. 2D and 3D convolution operations. a) Applying 2D convolution on an image results in an image. b) Applying 2D convolution on a video volume (multiple frames as multiple channels) also results in an image. c) Applying 3D convolution on a video volume results in another volume, preserving temporal information of the input signal.
Recent CNN methods

Action recognition with trajectory pooled convolutional descriptors
[Wang et al. CVPR15]
Action recognition - tasks

- Action classification: assigning an action label to a video clip

Making sandwich: present
Feeding animal: not present...

[Image of a child making a sandwich]
Action recognition - tasks

- Action classification: assigning an action label to a video clip

- Action localization (temporal): search temporal locations of an action in a video

Making sandwich: present
Feeding animal: not present
...
Action recognition - tasks

- Action localization (spatio-temporal) + interaction with an object, human, etc.

[Prest et al., PAMI 13]
Why automatic action localization?

• Query for specific videos in professional Archives and YouTube
• Analyze and describe content of videos
• Produce audio descriptions for visual impaired

Education: How do I make a pizza?

Sociology research: Influence of character smoking in movies
Why automatic action localization?

- Car safety & self-driving and video surveillance
- Detection of humans (pedestrians) and their motion, detection of unusual behavior
Temporal action localization

• Temporal sliding window
  – Robust video repres. for action recognition, Oneata et al., IJCV’15
  – Automatic annotation of actions in video, Duchenne et al., ICCV’09
  – Temporal localization of actions with actoms, Gaidon et al., PAMI’13

• Shot detection
  – ADSC Submission at Thumos Challenge 2015
Spatio-temporal action localization

[Retrieving actions in movies, I. Laptev and P. Pérez, ICCV’07]
Action representation

Hist. of Gradient
Hist. of Optic Flow

\( \begin{pmatrix} X \\ Y \\ T \end{pmatrix} \)

features: \( f_1, f_2, f_3, \ldots \)

\( \Delta T \)

\( \Delta Y \)

\( \Delta X \)

block-histogram features:

\( f = H \)

\( f = (H_1, H_2) \)

\( f = (H_1, H_2, H_3, H_4) \)

Plain
Temp-2
Spat-4
Action learning

- Efficient discriminative classifier [Freund & Schapire’97]
- Good performance for face detection [Viola & Jones’01]

AdaBoost:

- Pre-aligned samples
- Haar features
- Histogram features

Boosting:

\[ H(z) = \text{sgn}\left( \sum_{t=1}^{T} \alpha_t h_t(f_t) \right) \]

Selected features

Weak classifier

Optimal threshold

Fisher discriminant

[Laptev, Perez 2007]
Manual annotation of drinking actions in movies: “Coffee and Cigarettes”; “Sea of Love”

“Drinking”: 159 annotated samples
“Smoking”: 149 annotated samples

Temporal annotation

Spatial annotation

First frame  →  Keyframe  →  Last frame

head rectangle

torso rectangle
Action Detection

Test episodes from the movie “Coffee and cigarettes”

[Laptev, Perez 2007]
20 most confident detections
Spatio-temporal action localization

Spatio-temporal action localization

- Modeling temporal human-object interaction

[Explicit modeling of human-object interactions in realistic videos, Prest et al., PAMI 13]
Tracking humans and objects

• Fully automatic human tracks: state of the art detector + Brox tracks
• Object tracks: detector learnt from annotated training images + Brox tracks
• Extraction of a large number of human-object track pairs
Action descriptors

• Interaction descriptor: relative location, area and motion between human and object tracks

• Human track descriptor: 3DHOG-track [Klaeser et al.'10]
Experimental results on C&C Drinking
Experimental results on C&C

Smoking
Experimental results on C&C

Coffee and Cigarettes (drinking)

Coffee and Cigarettes (smoking)
## Comparison to the state of the art

<table>
<thead>
<tr>
<th>Method</th>
<th>Drinking</th>
<th>Smoking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interaction classifier</td>
<td>31.60</td>
<td>16.20</td>
</tr>
<tr>
<td>Object classifier</td>
<td>4.30</td>
<td>5.50</td>
</tr>
<tr>
<td>3DHOG-track classifier</td>
<td>52.20</td>
<td>21.50</td>
</tr>
<tr>
<td>Combination</td>
<td>62.10</td>
<td>32.80</td>
</tr>
<tr>
<td>Laptev et al. [22]</td>
<td>43.40</td>
<td>-</td>
</tr>
<tr>
<td>Willems et al. [35]</td>
<td>45.20</td>
<td>-</td>
</tr>
<tr>
<td>Klaeser et al. [20]</td>
<td>54.10</td>
<td>24.50</td>
</tr>
</tbody>
</table>
Experimental results on Rochester dataset

- Rochester daily activities dataset
  - 150 videos of 5 persons
  - leave-one-person-out test scenario
Experimental results on Rochester dataset
Learning to track for spatio-temporal action localization

frame-level object proposals and CNN action classifier
[Gkioxari and Malik, CVPR 2015]

tracking best candidates
Instant & class level tracking

scoring with CNN + IDT

temporal detection sliding window

Frame-level candidates

- For each frame
  - Compute object proposals (EdgeBoxes [Zitnick et al. 2014])
  - Extract CNN features (training similar to R-CNN [Girshick et al. 2014])
  - Score each object proposal

[Gkioxari and Malik’15, Simonyan and Zisserman’14]
Tracking best candidates

• Select the top scoring proposals

• For each selected candidate
  ▶ Learn an instance-level detector
  ▶ For each frame
    • Perform a sliding-window and select the best box according to the class-level detector and the instance-level detector
    • Update instance-level detector

class-level → robustness to drastic change in poses (Diving, Swinging)
instance-level → sufficiently specific
Rescoring and temporal sliding window

- To capture the dynamics
  ▶ Dense trajectories

- Temporal sliding window
Datasets (spatial localization)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of videos</td>
<td>150</td>
<td>928</td>
</tr>
<tr>
<td>Number of classes</td>
<td>10</td>
<td>21</td>
</tr>
<tr>
<td>Average length</td>
<td>63 frames</td>
<td>34 frames</td>
</tr>
</tbody>
</table>

![Images of datasets examples]
Datasets

- **UCF-101** [Soomro et al. 2012]
  - Spatio-temporal localization for a subset of the dataset
  - 3207 videos, 24 classes
  - Average length: 176 frames
## Results

### Impact of the tracker

<table>
<thead>
<tr>
<th>Detectors in the tracker</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>UCF-Sports</td>
</tr>
<tr>
<td>instance-level + class-level</td>
<td>90.50%</td>
</tr>
<tr>
<td>instance-level</td>
<td>74.27%</td>
</tr>
<tr>
<td>class-level</td>
<td>85.67%</td>
</tr>
</tbody>
</table>

### Comparison to SOA on UCF-Sports

<table>
<thead>
<tr>
<th>mAP</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Gkioxari and Malik 2015</td>
<td>75.8</td>
</tr>
<tr>
<td>Ours</td>
<td>90.5</td>
</tr>
</tbody>
</table>

### Comparison to SOA on J-HMDB

<table>
<thead>
<tr>
<th>mAP</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Gkioxari and Malik 2015</td>
<td>53.3</td>
</tr>
<tr>
<td>Ours</td>
<td>59.7</td>
</tr>
</tbody>
</table>
Quantitative evaluation (UCF-101)

<table>
<thead>
<tr>
<th></th>
<th>mAP</th>
<th>0.05</th>
<th>0.2</th>
<th>0.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yu and Yuan’15</td>
<td></td>
<td>42.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td></td>
<td>54.28</td>
<td>46.7</td>
<td>37.8</td>
</tr>
</tbody>
</table>
Spatio-temporal action localization

UCF-101
Spatio-temporal video tubes

• Brox and Malik, Object segmentation by long term analysis of point trajectories, ECCV’10

• Oneata et al., Spatio-temporal object detection proposals, ECCV’14

• Gemert et al., Action localization proposals from dense trajectories, BMVC’15

• Yu and Yuan, Fast action proposals for human action detection and search, CVPR’15
Human pose estimation + action recognition

- Estimation of body joints in video

Poses in the wild dataset [Cherian’14]  Pose results [Pfister’15]
Potential impact of human pose on action classification

- Systematically replace steps of “dense trajectories” with ground truth
- Ground-truth annotations for a subset of HMDB (Joint-HMDB)
- Pose features (joint position and spatio-temporal relations) results in a significant improvement

[H. Jhuang et al.'13]
Robust pose features – Pose-CNN

- Track human pose in a video → body part track
- Extract CNN features (appearance and motion) per part-track
- Train SVM classifier

1) input video
2) video pose estimation [Cherian'14]
3) crop human body parts
4) extract CNN features (appearance and motion) per part and per frame
5) video descriptors: aggregation of frame features (max/min)
6) P-CNN: concatenation of part features from appearance and flow
Datasets used for evaluation

• JHMB as described previously

• MPI cooking
  – 64 fine grained actions
  – a total of 5609 clips, 7 training/test splits
  – similar action, i.e. cut dice, cut slices, and cut stripes

• Sub-MPI
  – selection of two similar classes
  – wash hands and wash objects with GT pose
## Performance of the individual features

- Different body parts are complementary
- Appearance and flow are complementary

<table>
<thead>
<tr>
<th>Parts</th>
<th>JHMDB-GT</th>
<th>MPII Cooking-Pose [8]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>App</td>
<td>OF</td>
</tr>
<tr>
<td>Hands</td>
<td>46.3</td>
<td>54.9</td>
</tr>
<tr>
<td>Upper body</td>
<td>52.8</td>
<td>60.9</td>
</tr>
<tr>
<td>Full body</td>
<td>52.2</td>
<td>61.6</td>
</tr>
<tr>
<td>Full image</td>
<td>43.3</td>
<td>55.7</td>
</tr>
<tr>
<td>All</td>
<td>60.4</td>
<td>69.1</td>
</tr>
</tbody>
</table>
Robustness of P-CNN

<table>
<thead>
<tr>
<th>JHMDB</th>
<th>GT</th>
<th>Pose [7]</th>
<th>Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>P-CNN</td>
<td>74.6</td>
<td>61.1</td>
<td>13.5</td>
</tr>
<tr>
<td>HLPF</td>
<td>77.8</td>
<td>25.3</td>
<td>52.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>sub-MPII Cooking</th>
<th>GT</th>
<th>Pose [7]</th>
<th>Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>P-CNN</td>
<td>83.6</td>
<td>67.5</td>
<td>16.1</td>
</tr>
<tr>
<td>HLPF</td>
<td>76.2</td>
<td>57.4</td>
<td>18.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MPII Cooking</th>
<th>Pose [7]</th>
</tr>
</thead>
<tbody>
<tr>
<td>P-CNN</td>
<td>62.3</td>
</tr>
<tr>
<td>HLPF</td>
<td>32.6</td>
</tr>
</tbody>
</table>

- P-CNN on par with HLPF for GT
- P-CNN significantly more robust for real noisy poses
Comparison to state of the art

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>P-CNN</td>
<td>74.6</td>
<td>61.1</td>
<td>62.3</td>
</tr>
<tr>
<td>DT-FV</td>
<td>65.9</td>
<td>65.9</td>
<td>67.6</td>
</tr>
<tr>
<td>P-CNN + DT-FV</td>
<td>79.5</td>
<td>72.2</td>
<td>71.4</td>
</tr>
</tbody>
</table>

- P-CNN better than IDT on ground-truth
- P-CNN and IDT are complementary
Conclusion

• Excellent results with P-CNN features
  – more robust than existing pose features (HLPF)
  – outperform IDT significantly for ground-truth annotations, on par for real pose

• Pose does help to describe human actions
  – Further improvement with very recent pose algorithms
  – Modeling interactions with objects
Where to get training data?

Weakly-supervised learning
Actions in movies

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

- 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?

---

Rick sits down with Ilsa.

Oh, it wasn't my secret, Richard. Victor wanted it that way. Not even our closest friends knew about our marriage.

---

Script-based video annotation

- Scripts available for >500 movies (no time synchronization)
- Subtitles (with time info.) are available for the most of movies
- Can transfer time to scripts by text alignment

[Laptev, Marszałek, Schmid, Rozenfeld 2008]
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
Hollywood-2 actions dataset

<table>
<thead>
<tr>
<th>Actions</th>
<th>Training subset (clean)</th>
<th>Training subset (automatic)</th>
<th>Test subset (clean)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AnswerPhone</td>
<td>66</td>
<td>59</td>
<td>64</td>
</tr>
<tr>
<td>DriveCar</td>
<td>85</td>
<td>90</td>
<td>102</td>
</tr>
<tr>
<td>Eat</td>
<td>40</td>
<td>44</td>
<td>33</td>
</tr>
<tr>
<td>FightPerson</td>
<td>54</td>
<td>33</td>
<td>70</td>
</tr>
<tr>
<td>GetOutCar</td>
<td>51</td>
<td>40</td>
<td>57</td>
</tr>
<tr>
<td>HandShake</td>
<td>32</td>
<td>38</td>
<td>45</td>
</tr>
<tr>
<td>HugPerson</td>
<td>64</td>
<td>27</td>
<td>66</td>
</tr>
<tr>
<td>Kiss</td>
<td>114</td>
<td>125</td>
<td>103</td>
</tr>
<tr>
<td>Run</td>
<td>135</td>
<td>187</td>
<td>141</td>
</tr>
<tr>
<td>SitDown</td>
<td>104</td>
<td>87</td>
<td>108</td>
</tr>
<tr>
<td>SitUp</td>
<td>24</td>
<td>26</td>
<td>37</td>
</tr>
<tr>
<td>StandUp</td>
<td>132</td>
<td>133</td>
<td>146</td>
</tr>
<tr>
<td>All Samples</td>
<td>823</td>
<td>810</td>
<td>884</td>
</tr>
</tbody>
</table>

Training and test samples are obtained from 33 and 36 distinct movies respectively.

Hollywood-2 dataset is on-line: http://www.irisa.fr/vista/actions/hollywood2

[Laptev, Marszałek, Schmid, Rozenfeld 2008]
# Action classification results

<table>
<thead>
<tr>
<th>Channel</th>
<th>Clean hogho/hof</th>
<th>Automatic hogho/hof</th>
<th>Chance</th>
</tr>
</thead>
<tbody>
<tr>
<td>mAP</td>
<td></td>
<td></td>
<td>9.2</td>
</tr>
<tr>
<td>AnswerPhone</td>
<td>15.7</td>
<td>18.2</td>
<td>7.2</td>
</tr>
<tr>
<td>DriveCar</td>
<td>86.6</td>
<td>78.2</td>
<td>11.5</td>
</tr>
<tr>
<td>Eat</td>
<td>59.5</td>
<td>13.0</td>
<td>3.7</td>
</tr>
<tr>
<td>FightPerson</td>
<td>71.1</td>
<td>52.9</td>
<td>7.9</td>
</tr>
<tr>
<td>GetOutCar</td>
<td>29.3</td>
<td>13.8</td>
<td>6.4</td>
</tr>
<tr>
<td>HandShake</td>
<td>21.2</td>
<td>12.8</td>
<td>5.1</td>
</tr>
<tr>
<td>HugPerson</td>
<td>35.8</td>
<td>15.2</td>
<td>7.5</td>
</tr>
<tr>
<td>Kiss</td>
<td>51.5</td>
<td>43.2</td>
<td>11.7</td>
</tr>
<tr>
<td>Run</td>
<td>69.1</td>
<td>54.2</td>
<td>16.0</td>
</tr>
<tr>
<td>SitDown</td>
<td>58.2</td>
<td>28.6</td>
<td>12.2</td>
</tr>
<tr>
<td>SitUp</td>
<td>17.5</td>
<td>11.8</td>
<td>4.2</td>
</tr>
<tr>
<td>StandUp</td>
<td>51.7</td>
<td>40.5</td>
<td>16.5</td>
</tr>
</tbody>
</table>

Average precision (AP) for Hollywood-2 dataset
Scripts as weak supervision

Challenges:

- Imprecise temporal localization
- No explicit spatial localization
- NLP problems, scripts ≠ training labels

“… Will gets out of the Chevrolet. …” vs. Get-out-car
“… Erin exits her new truck…”

Subtitles

00:24:22 → 00:24:25
- Yes, Monsieur Laszlo. Right this way.

00:24:51 → 00:24:53
Two Cointreaux, please.

Script

Speech
Monsieur Laszlo. Right this way.

Scene description
As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilisa...

Speech
Two cointreaux, please.