Learning Realistic Human Actions from Movies

Ivan Laptev*, Marcin Marszałek**, Cordelia Schmid**, Benjamin Rozenfeld***

• INRIA Rennes, France
** INRIA Grenoble, France
*** Bar-Ilan University, Israel

Presented by: Nils Murrugarra
University of Pittsburgh
Motivation

- Internet has tons of video and still growing
- Human actions are very common in movies, TV news, personal video …

Action recognition useful for:
  - Content-based browsing
    e.g. fast-forward to the next goal scoring scene
  - Human scientists
    influence of smoking in movies on adolescent smoking

150,000 uploads every day
Motivation

• Actions in current datasets:

• Actions “In the Wild”:

Context

Web video search

- Useful for some action classes: kissing, hand shaking
- Noise results and not useful for most action

Web image search

- Useful for learning action context: static scenes and objects
- See also [Li-Jia & Fei-Fei ICCV07]

How to find real actions?
Context

**Movies** contains many classes and many examples of realistic actions

Problems:
- Only few class-samples per movie
- Manual annotation is very time consuming

How to annotate automatically?
Method – Annotation [1]

- Scripts available with no time synchronization
- Subtitles + time information

How to use the previous information?
- Identify an action and transfer time to scripts by text alignment

---

subtitles

01:20:17,240 --> 01:20:20,437
Why weren't you honest with me? Why'd you keep your marriage a secret?

01:20:20,640 --> 01:20:23,598
It wasn't my secret, Richard. Victor wanted it that way.

01:20:23,800 --> 01:20:26,189
Not even our closest friends knew about our marriage.

movie script

Rick sits down with Ilsa.

01:20:23

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

---

Method – Annotation

On the good side:

- Realistic variation of actions: subjects, views, etc…
- Many Classes and many examples per action
- No additional work for new classes
- Character names may be used to resolve “who is doing certain action?”

Problems:

- No spatial localization (no bounding box)
- Temporal localization may be poor
- Missing actions: e.g. scripts do not always follow the movie (not aligned)
- Annotation is incomplete, it can’t be a ground truth for test stage
- Large within-class variability per action *in text*
Method – Annotation - Evaluation

1. Annotate action samples *in text*
2. Perform automatic script-video alignment
3. Check the correspondence based on manual annotation

\[ a = \frac{\# \text{ matched words}}{\# \text{ all words}} \]

**Example of a “visual false positive”**

A black car pulls up, two army officers get out.

**Evaluation of retrieved actions on visual ground truth**

![Graph showing evaluation scores](image)

- \( a = 1.0 \)
- \( a \geq 0.5 \)

**How to improve?**

A black car pulls up, two army officers get out.

\[ a: \text{ quality of subtitle-script matching} \]
Method – Annotation – Text Approach

**Problem:** Text can express the same action in different ways:

<table>
<thead>
<tr>
<th>Action:</th>
<th>GetOutCar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example:</td>
<td>“… Will gets out of the Chevrolet. …” “… Erin exits her new truck…”</td>
</tr>
<tr>
<td>Potential false positives:</td>
<td>“…About to sit down, he freezes…”</td>
</tr>
</tbody>
</table>

**Solution:** Supervised text classification approach

- Given an scene description, predict if a target action is present or not
- Based on bag-of-words representation

```
  \[
  \begin{pmatrix}
    \text{cosmonaut} & d1 & 0 & 1 & 0 & 0 & 0 \\
    \text{astronaut} & 0 & 1 & 0 & 0 & 0 & 0 \\
    \text{moon} & 1 & 1 & 0 & 0 & 0 & 0 \\
    \text{car} & 1 & 0 & 0 & 1 & 1 & 0 \\
    \text{truck} & 0 & 0 & 0 & 1 & 0 & 1 \\
  \end{pmatrix}
  \]```
Method – Annotation – Text Approach

Features:
- Words
- Adjacent pair of words
- Non-adjacent pair of words within a small window
### Method – Annotation – Data

#### 12 movies
- **False**: 5, 6, 9, 7, 10, 21, 5, 33, 96
- **Correct**: 15, 6, 14, 8, 34, 30, 7, 29, 143
- **All**: 20, 12, 23, 15, 44, 51, 12, 62, 239

**automatically labeled training set**

- 22, 13, 20, 22, 49, 47, 11, 48, 232

**Goal**
- Compare performance of manual annotated data with automatic version

#### 20 different movies
- **test set**: 23, 13, 19, 22, 51, 30, 10, 49, 217

**60%**

**a>0.5**

video length <= 1000 frames
Method – Action Classifier [Overview]

Bag of space-time features + multi-channel SVM

[4], [5], [6]

Collection of space-time patches

Histogram of visual words

Multi-channel SVM Classifier

Method – Action Classifier - Features

- Space-time corner detector
  [7]

- Dense scale sampling (no explicit scale selection)
Method - Action Classifier - Descriptor

Multi-scale space-time patches from corner detector

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

Histogram of optical flow (HOF)
3x3x2x5bins HOF descriptor

Public code available at www.irisa.fr/vista/actions
Method - Action Classifier - Descriptor

**Visual Vocabulary Construction**

- Used a subset of 100’000 features sampled from training videos
- Identified 4000 clusters with k-means
  - Centroids = Visual Vocabulary Words

Bag-of-features
Method - Action Classifier - Descriptor

Vector BOF Generation

- Compute all features
- Assign each feature to the closest vocabulary word
- Compute vector of visual word occurrences.
Method - Action Classifier - Descriptor

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 and centre-focused ot2
Method - Action Classifier - Descriptor

Global spatio-temporal grids

Entire Action Sequence

Action Sequence Splitted on 2 over time

Action Sequence

Normalized
Method - Action Classifier - Descriptor

Global spatio-temporal grids

\[ y \quad t \]

\[ x \]

\[ h3 \times 1 \ t1 \]

\[ o2 \times 2 \ t1 \]
Method - Action Classifier - Learning

Non-Linear SVM:
• Map original space to a higher space, where the data is separable

\[ \Phi: x \rightarrow \varphi(x) \]
Method - Action Classifier - Learning

Multi-channel chi-square kernel

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

- Channel \( c \) is a combination of a descriptor (HOG or HOF) and a spatio-temporal grid
- \( D_c(H_i, H_j) \) is the chi-square distance between histograms
- \( A_c \) is the mean value of the distances between all training samples for the channel \( c \)
- The best set of channels \( C \) for a given training set is found based on a greedy approach
# Evaluation - Action Classifier

<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 l1)</td>
<td>41.5% (hof o2x2 t1, hog h3x1 t3)</td>
</tr>
<tr>
<td>Action HandShake</td>
<td>18.6%</td>
<td>12.1%</td>
<td>23.7% (hog h3x1 l1)</td>
<td>32.3% (hog h3x1 t1, hog o2x2 t1)</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 t3)</td>
</tr>
<tr>
<td>Action Kiss</td>
<td>52.0%</td>
<td>36.5%</td>
<td>52.0% (hog 1 l1)</td>
<td>53.3% (hog 1 t1, hog o2x2 f1)</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 h3x1 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 t3)</td>
</tr>
<tr>
<td>Action StandUp</td>
<td>45.4%</td>
<td>40.0%</td>
<td>45.4% (hog 1 l1)</td>
<td>50.5% (hog 1 t1, hog 1 t2)</td>
</tr>
</tbody>
</table>

**Findings**

- Different grids and channels combination are beneficial to increment performance
- HOG performs better for realistic actions (context, image content)
Evaluation - Action Classifier

Number of occurrences for each channel component within the optimized channel combinations for the KTH action dataset and our manually labelled movie dataset.
Evaluation - Action Classifier

Sample frames from the KTH actions sequences, all classes (columns) and scenarios (rows) are presented
Evaluation - Action Classifier

Average class accuracy on the KTH actions dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>Schuldt et al.</th>
<th>Niebles et al.</th>
<th>Wong et al.</th>
<th>Nowozin et al.</th>
<th>ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>71.7%</td>
<td>81.5%</td>
<td>86.7%</td>
<td>87.0%</td>
<td>91.8%</td>
</tr>
</tbody>
</table>

Confusion matrix for the KTH actions

<table>
<thead>
<tr>
<th></th>
<th>Walking</th>
<th>Jogging</th>
<th>Running</th>
<th>Boxing</th>
<th>Waving</th>
<th>Clapping</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walking</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jogging</td>
<td>.99</td>
<td>.01</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>Running</td>
<td>.04</td>
<td>.89</td>
<td>.07</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>Boxing</td>
<td>.01</td>
<td>.19</td>
<td>.80</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>Waving</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.97</td>
<td>.00</td>
<td>.03</td>
</tr>
<tr>
<td>Clapping</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.05</td>
<td>.00</td>
<td>.95</td>
</tr>
</tbody>
</table>
**Evaluation - Action Classifier**

**Noise Robustness**

- $p \leq 0.2$; performance decreases insignificantly
- $p = 0.4$; performance decreases by around 10%

**Why?** Automatic Annotation avoid cost of human annotation
Evaluation - Action Classifier

Evaluation in Real-World Videos

- TP: Class present, prediction says YES
- FN: Class present, prediction says NO
- FP: Class not present, prediction says YES
- TN: Correct Prediction
Evaluation - Action Classifier

Evaluation in Real-World Videos

<table>
<thead>
<tr>
<th>AnswerPhone</th>
<th>GetOutCar</th>
<th>HandShake</th>
<th>HugPerson</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TN</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FN</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Action Classification example results based on **automatic annotated data**
Evaluation - Action Classifier

<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>

Evaluation based on Average precision (AP) over actions.

**Clean** = Annotated  
**Chance** = Random Classifier
Demo - Action Classifier

Test episodes from movies “The Graduate”, “It’s a wonderful life”, “Indiana Jones and the Last Crusade”
Conclusion

Summary

- Automatic generation of realistic action samples
  - New action dataset available [www.irisa.fr/vista/actions](http://www.irisa.fr/vista/actions)
  - Bag-of-features expanded to video domain
  - Best performance on KTH benchmark
  - Promising results for actions in the “wild”

Disadvantages

- Still improvement in automatic annotation is required. Only a 60% was achieved.
- Parameters for the grid of cuboids are not well-justified, how were determined. Similarly, the # of visual words for k-means algorithm.
  - K-means is susceptible to outliers.
- A greedy approach for determine the best set of channels can achieve sub-optimal results.

Future directions

- Automatic action class discovery
- Internet-scale video search
Questions
References


Action Recognition Using a Distributed Representation of Pose and Appearance

Subhransu Maji\textsuperscript{1}, Lubomir Bourdev \textsuperscript{1,2}, and Jitendra Malik\textsuperscript{1}

\textsuperscript{1}University of California at Berkeley
\textsuperscript{2}Adobe Systems, Inc.

Presented by: Nils Murrugarra
University of Pittsburgh
Goal

Static Image Action Recognition

Motivation:

• Humans can easily recognize pose and actions from Limited Views of a single image.

• Action and pose is identified by body parts (occlusions) at different locations and scales.
Poselets

Poselet:

- Body part detectors of joint locations of people in images.
- They are used to find patches related to a given configuration of joints.

L. Bourdev, S. Maji, T. Brox and J. Malik, Detection People using Mutually Consistent Poselet Activations, ECCV 2010
Robust Representation of Pose and Appearance

Poselet Activation Vector

- Poselet annotation are reused from a previous article.
- Represent each example by the poselets that are active.

Estimate 3D Orientation of Head and Torso
Data Collection

Manual Verification
- Discard images with high disagreement
  - Low resolution and high occlusion
- Only used rotation in Y

Human Error
- Small error in canonical views (front, back, left and right)
- Measured as average of standard deviation

3D pose of head and torso Annotations

Amazon Mechanical Turk
3D Estimation - Goal

Goal

- Given a bounding box of a person, estimate its 3D orientation of head and torso.
3D Estimation - Descriptor

Procedure

• Discretize 3D orientation $[-180, 180]$ in 8 bins [Classification].
• Angled estimation based on interpolation
  • Highest predicted bin
  • Two adjacent neighbors

<table>
<thead>
<tr>
<th>pt1</th>
<th>pt2</th>
<th>pt3</th>
<th>.</th>
<th>.</th>
<th>.</th>
<th>.</th>
<th>ptn</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7</td>
<td>0.8</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>0.2</td>
<td>0.9</td>
</tr>
</tbody>
</table>

Each entry correspond to a poselet type
3D Estimation – Example Results
3D Estimation – Evaluation

Head Orientation: 62.1 %

Torso Orientation: 61.71 %
Goal

• Given a bounding box, estimate an action category
Pose alone can’t learn to identify actions
2D Action Classifier - Method

Solution
- Learn appearance considering poselets per action category
- Based on HOG and SVM

Appearance information would help
2D Action Classifier - Method

**Approach**
- Find Poselet k-Nearest Neighbors
- Select the more discriminative
- Learn appearance model based on HOG and SVM
Object Interaction can help?

• It was considered an interaction with horse, motorbike, bicycle and TV.
• A people-object model spatial location was learnt [object activation vector]
2D Action Classifier - Method

Context can still help us?

Add action classifier for other people in image
### 2D Action Classifier - Evaluation

<table>
<thead>
<tr>
<th>Category</th>
<th>Validation</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PAV</td>
<td>w/ OAV</td>
</tr>
<tr>
<td>phoning</td>
<td>63.3</td>
<td>62.0</td>
</tr>
<tr>
<td>playing instrument</td>
<td>44.2</td>
<td>44.4</td>
</tr>
<tr>
<td>reading</td>
<td>37.4</td>
<td>44.4</td>
</tr>
<tr>
<td>ridingbike</td>
<td>62.0</td>
<td><strong>84.7</strong></td>
</tr>
<tr>
<td>ridinghorse</td>
<td>91.1</td>
<td><strong>97.7</strong></td>
</tr>
<tr>
<td>running</td>
<td>82.4</td>
<td>84.1</td>
</tr>
<tr>
<td>taking photo</td>
<td>21.1</td>
<td>22.9</td>
</tr>
<tr>
<td>using computer</td>
<td>54.2</td>
<td><strong>64.9</strong></td>
</tr>
<tr>
<td>walking</td>
<td>82.0</td>
<td>83.6</td>
</tr>
<tr>
<td>average</td>
<td>59.8</td>
<td>65.3</td>
</tr>
</tbody>
</table>

*Image context: best in “ridingbike”, second best in “ridinghorse”*
Conclusion

Summary

- A method for Action Recognition in static image was presented.
- It is based mainly in:
  - Poselet features
  - An Appearance model
  - Object Interaction
  - Context information

Disadvantages

- The use of bounding-boxes is not realistic. A better scenario is that given an image, an algorithm should detect all people actions automatically.
- Related to the Poselet Activation Vector, a intersection threshold of 0.15 is defined. How this threshold was determined? A similar situation happens with the Spatial Model of Object Interaction.
Questions
References

