Learning realistic human actions from movies

Ivan Laptev, Marcin Marszalek, Cordelia Schmid, Benjamin Rozenfeld
CVPR 2008

Presented by Piotr Mirowski
Courant Institute, NYU
Advanced Vision class, November 3, 2009
Challenge

• Existing human action recognition datasets show:
  • few action classes
  • in simplified settings
• In **real movies**, individual **variations**:
  • expression, posture, motion and clothing
  • perspective and camera motion
  • illumination
  • occlusion and scene surroundings
Outline

• Automatic annotation of human actions
• Video classification for action recognition
• Results
  • Evaluation of spatio-temporal grids and noise
  • Classification performance on KTH dataset
  • Real-world videos
• Space-time interest points
Annotation of movies

- Even common action classes appear few times per movie
- Need to annotate hundreds of hours
- Solution: use movie scripts
  - Advantage: publicly available
  - Drawback: no time alignment with movie
  - Drawback: variation of action expression
  - Drawback: script and action can mismatch
Time-alignment of scripts and subtitles

- **Scripts** in plain text
- use indentation to separate actions from dialogue
- word matching of script with subtitles (dynamic programming)
- misalignment score: \( a = \frac{\#\text{matched words}}{\#\text{all words}} \)
- **subtitles** have precise time information


[Figure 2. Example of matching speech sections (green) in subtitles and use of time information (yellow).]
Manual evaluation of script-based annotation

- **if perfect** text alignment between script and subtitles \( (a=1) \)
  - 70% actions match
  - 10% outside field
  - 10% misaligned

- **if imperfect** alignment \( (a<1) \)
  - <60% actions match

---

Evaluation of retrieved actions on visual ground truth


A black car pulls up. Two army officers get out.

\[1:13:41 - 1:13:45\]
Text retrieval of human actions from scripts

8 action classes
~400 action samples
>17k non-action samples

- **Text features:**
  - words
  - adjacent pairs of words (bigrams)
  - Bag of Features (BoF) sparse vector
  - Multi-class classification with **regularized perceptron**

Regularized perceptron better than keywords

Construction of video datasets

- classify 8 classes of actions on:
  - **training set:** 12 movies
  - **testing set:** 20 movies
- **manual training dataset:** select visually correct samples from manual action annotation
- **automatic training dataset:** no visual check


![Table 1](http://www.irisa.fr/vista/actions)
Outline

• Automatic annotation of human actions
• Video classification for action recognition
• Results
  • Evaluation of spatio-temporal grids and noise
  • Classification performance on KTH dataset
  • Real-world videos
• Space-time interest points
Multi-scale space-time interest points

Figure 5. Space-time interest points detected for two video frames with human actions hand shake (left) and get out car (right).

Spatial scale: $\sigma = 2$ to 16
Time scale: $\tau = \sqrt{2}$ or 2

Feature descriptors using local histograms

SIFT-like histograms of:
- image gradient (HoG)
- optical flow (HoF)

In a volume: centered around interest point, of size $18\sigma \times 18\sigma \times 18\tau$, subdivided into $3 \times 3 \times 2$ cuboids

Spatio-temporal Bag of Features (BoF)

- **“visual vocabulary”**
- >100,000 interest points and HoG or HoF feature descriptors on training samples
- **k-means clustering** to assign interest points using Euclidian distance in HoG or HoF space
- Empirical choice k=4000

Space-time volumes

Spatio-temporal Bag of Features (BoF)

- "visual vocabulary"
- >100,000 interest points and HoG or HoF feature descriptors on training samples
- k-means clustering to assign interest points using Euclidian distance in HoG or HoF space
- Empirical choice $k=4000$
- each interest point is assigned to the closest among $k$ clusters
- each assignment is a "word"
- histogram of words within a HoG or HoF "channel"
- 24 channels i.e. video subdivisions into grid cells
Spatio-temporal Bag of Features (BoF)

- each interest point is assigned to the closest among k clusters

- histogram of words within a HoG or HoF “channel”

- 24 channels i.e. video subdivisions into grid cells

---

[Figure 6. Examples of a few spatio-temporal grids.]

[C1]
[C2]
[C3]
[C4]
[C4000]

[Image 717x343 to 806x407]

[Image 224x126 to 349x317]

[Image 224x54 to 349x113]
SVM classification of action words

- **Word** histograms for samples $i, j$: $H_i = \{h_{in}\}$ $H_j = \{h_{jn}\}$
- Vocabulary of $V$ words
- **Histogram distance**: $D_c(H_i, H_j) = \frac{1}{2} \sum_{n=1}^{V} \frac{(h_{in} - h_{jn})^2}{h_{in} + h_{jn}}$
- one distance for each channel $c$
- multi-channel **kernel** for Support Vector Machine (SVM):
  $K(H_i, H_j) = \exp (- \sum_{c \in C} \frac{1}{A_c} D_c(H_i, H_j))$
- grow set of channels $C$ in a greedy way
Summary of video classification method

- Find **space-time interest points**

- Describe, using **local histograms**, the **gradient and optical flow** in the neighborhood of each interest point

- Classify the **feature descriptor** to the **closest word**

- Describe, using **histograms, word occurrences** within large-scale **space-time channels**

- Classify a video snippet, using a set of **channels** and **word histograms**, and an **SVM multi-class classifier**
Outline

• Automatic annotation of human actions
• Video classification for action recognition

• Results
  • Evaluation of spatio-temporal grids and noise
  • Classification performance on KTH dataset
  • Real-world videos

• Space-time interest points
Evaluation of spatio-temporal grids

- Select best spatio-temporal channels for the KTH dataset and real movies dataset
- HoG
- HoF
- Grids
- Significant gain over Bag of Features (BoF)

Figure 7. Number of occurrences for each channel component within the optimized channel combinations for the KTH action dataset and our manually labeled movie dataset.

<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>91.6%</td>
<td>89.7%</td>
<td>91.1% (hof h3x1 t3)</td>
<td>91.8% (hof 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 t1)</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% (hog h3x1 t1)</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 t1)</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 t1)</td>
<td>50.5% (hog 1 t1, hog t2)</td>
</tr>
</tbody>
</table>

Robustness to noise in training data labels

- Add noise (wrong labels) into training dataset (KTH)
- Generate wrong labels with probability $p$
- **Insensitivity of classification accuracy to wrong labels**
  - $p<0.2$: no effect
  - $p<0.4$: decrease by 10%

[Fig. 9] Performance of our video classification approach in the recognition scenario, where the observed amount of wrong labels is significantly, i.e., by less than two percent. At errors. Up to errors. Note that we have observed a comparable level of resistance to noise in the training data or the problem is decomposed into simpler examples the occluded handshake or the hardly visible person many display features not unusual for the classified action, for example the rapid getting up is typical for "GetOutCar". Most of the false negatives are very difficult to recognize, see for example the stretched hands are typical for "HugPerson". Most of the false negatives are very difficult to recognize, see for example the stretched hands are typical for "HugPerson".

The classification results are good for the manual training data. We also show results for a random classifier (chance). Table 5. Average precision (AP) for each action class of our test set. We compare results for clean (annotated) and automatic training. Table 5. Average precision (AP) for each action class of our test set. We compare results for clean (annotated) and automatic training.

**Figure 9**

- Average class accuracy
- Probability of a wrong label

**Table 5**

<table>
<thead>
<tr>
<th>Action</th>
<th>Clean</th>
<th>Automatic</th>
</tr>
</thead>
<tbody>
<tr>
<td>StandUp</td>
<td>58%</td>
<td>56%</td>
</tr>
<tr>
<td>SitUp</td>
<td>62%</td>
<td>60%</td>
</tr>
<tr>
<td>SitDown</td>
<td>65%</td>
<td>63%</td>
</tr>
<tr>
<td>Kiss</td>
<td>70%</td>
<td>68%</td>
</tr>
<tr>
<td>HugPerson</td>
<td>75%</td>
<td>73%</td>
</tr>
<tr>
<td>HandShake</td>
<td>80%</td>
<td>78%</td>
</tr>
<tr>
<td>AnswerPhone</td>
<td>85%</td>
<td>83%</td>
</tr>
</tbody>
</table>

Note that results obtained by Jhuang et al. [63] and Wong et al. [19] are not comparable to ours, as they are based on non-standard experimental setups: they either use more classes except "HandShake" the automatic training obtains results significantly above chance level. This shows that an automatically trained system can successfully recognize human actions in real-world videos. For kissing, the performance is evaluated with the average precision (AP) of the probability that a label being wrong. Training for the problem of wrong labels in the training set. In this section we evaluate the robustness of our action classification method on the natural-scene images of the PASCAL training set and lower for the automatic one. However, for all cases except "HandShake" the automatic training obtains results significantly above chance level. This shows that an automatically trained system can successfully recognize human actions in real-world videos. For kissing, the performance is evaluated with the average precision (AP) of the probability that a label being wrong. Training for the problem of wrong labels in the training set.
Outline

• Automatic annotation of human actions
• Video classification for action recognition
• Results
  • Evaluation of spatio-temporal grids and noise
  • Classification performance on KTH dataset
  • Real-world videos
• Space-time interest points
KTH dataset

- 6 types of human actions
- 4 scenarios:
  - outdoors
  - outdoors + scale variation
  - outdoors + different clothes
  - indoors
- 2391 sequences

Outperforms state-of-the-art on KTH dataset

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>71.7%</td>
<td>81.5%</td>
<td>86.7%</td>
<td>91.8%</td>
</tr>
</tbody>
</table>

Table 3. Average class accuracy on the KTH actions dataset.

<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>.99</td>
<td>.01</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>Jogging</td>
<td>.04</td>
<td>.89</td>
<td>.07</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>Running</td>
<td>.01</td>
<td>.19</td>
<td>.80</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>Boxing</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.97</td>
<td>.00</td>
<td>.03</td>
</tr>
<tr>
<td>Waving</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.91</td>
<td>.09</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>

Table 4. Confusion matrix for the KTH actions.

[Schuldt et al. (2008) “Learning realistic human actions from movies”]
Outline

• Automatic annotation of human actions
• Video classification for action recognition

• Results
  • Evaluation of spatio-temporal grids and noise
  • Classification performance on KTH dataset
  • Real-world videos

• Space-time interest points
Recognition results in real-world videos

- 217 test videos
- one classifier per action (present or absent)
- results reported as Average Precision (AP)
- manual training achieves better AP than automatic training
- results are significative

<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 5. 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).

\[
\text{Precision} = \frac{tp}{tp + fp}
\]

Examples of positives and negatives

AnswerPhone  GetOutCar  HandShake  HugPerson  Kiss  SitDown  SitUp  StandUp

TP  TN  FN  FP
Outline

- Automatic annotation of human actions
- Video classification for action recognition
- Results
  - Evaluation of spatio-temporal grids and noise
  - Classification performance on KTH dataset
  - Real-world videos
- Space-time interest points
Introduction

- interest points in spatio-temporal domain
- = “interesting events”
- = sparse coding of video
- = extend Harris interest point detector to time
- = corners in space-time

Harris interest point detector

- **Scale-space representation** of image $f^{sp}$
  
  \[ L^{sp}(x, y; \sigma_l^2) = g^{sp}(x, y; \sigma_l^2) \ast f^{sp}(x, y), \]

  (smoothed image)

- Gaussian blur:
  
  \[ g^{sp}(x, y; \sigma_l^2) = \frac{1}{2\pi\sigma_l^2} \exp\left(-\frac{(x^2 + y^2)}{2\sigma_l^2}\right) \]

- Gaussian derivatives:
  
  \[ L_{x}^{sp} = \partial_x (g^{sp}(\cdot; \sigma_l^2) \ast f^{sp}(\cdot)) \]

- Harris interest point detector localizes significant changes in both directions
  
  \[ \mu^{sp}(\cdot; \sigma_l^2, \sigma_i^2) = g^{sp}(\cdot; \sigma_l^2) \ast ((\nabla L(\cdot; \sigma_l^2))(\nabla L(\cdot; \sigma_l^2))^T) \]

  \[ = g^{sp}(\cdot; \sigma_l^2) \ast \begin{pmatrix} (L_{x}^{sp})^2 & L_{x}^{sp}L_{y}^{sp} \\ L_{x}^{sp}L_{y}^{sp} & (L_{y}^{sp})^2 \end{pmatrix} \]

- Second-moment matrix:
Harris corner function $H$

- Second-moment descriptor equivalent to the covariance matrix of 2D distribution of image orientation vectors, in a local neighborhood
  \[
  \mu_{sp}(\cdot; \sigma_i^2, \sigma_l^2) = g_{sp}(\cdot; \sigma_i^2) * ((\nabla L(\cdot; \sigma_i^2))(\nabla L(\cdot; \sigma_i^2))^T)
  \]

- Eigenvalues $\lambda_1, \lambda_2, (\lambda_1 \leq \lambda_2)$ describe the variations along the 2 image directions
  \[
  \lambda_1, \lambda_2 = g_{sp}(\cdot; \sigma_i^2) * \begin{pmatrix}
  (L_{sp}^x)^2 & L_{sp}^x L_{sp}^y \\
  L_{sp}^x L_{sp}^y & (L_{sp}^y)^2
  \end{pmatrix}
  \]

- Two large eigenvalues indicate interest points
  \[
  H_{sp} = \det(\mu_{sp}) - k \text{trace}^2(\mu_{sp}) = \lambda_1 \lambda_2 - k(\lambda_1 + \lambda_2)^2
  \]

- Maximize corner function:
Interest points in spatio-temporal domain

\[ L(\cdot; \sigma_i^2, \tau_i^2) = g(\cdot; \sigma_i^2, \tau_i^2) * f(\cdot) \]

\[ g(x, y, t; \sigma_i^2, \tau_i^2) = \frac{1}{\sqrt{(2\pi)^3 \sigma_i^4 \tau_i^2}} \exp\left(-\frac{(x^2 + y^2)}{2\sigma_i^2} - \frac{t^2}{2\tau_i^2}\right) \]

\[ \mu = g(\cdot; \sigma_i^2, \tau_i^2) * \begin{pmatrix} L_x^2 & L_xL_y & L_xL_t \\ L_xL_y & L_y^2 & L_yL_t \\ L_xL_t & L_yL_t & L_t^2 \end{pmatrix} \]

\[ H = \det(\mu) - k \text{trace}^3(\mu) = \lambda_1 \lambda_2 \lambda_3 - k(\lambda_1 + \lambda_2 + \lambda_3)^3 \]
Synthetic image results (mono-scale)

Figure 2: Results of detecting spatio-temporal interest points on synthetic image sequences. (a): A moving corner; (b) A merge of a ball and a wall; (c): Collision of two balls with interest points detected at scales $\sigma^2 l = 8$ and $\tau^2 l = 8$; (d): the same signal as in (c) but with the interest points detected at coarser scales $\sigma^2 l = 16$ and $\tau^2 l = 16$.

The space-time location of the interest point and the length of the semi-axes proportional to the local scale parameters $\sigma^2 l$ and $\tau^2 l$ used in the computation of $H$.

Figure 2a shows a sequence with a moving corner. The interest point is detected at the moment in time when the motion of the corner changes direction. This type of event occurs frequently in natural sequences, such as sequences of articulated motion. Note that according to the definition of spatio-temporal interest points, image structures with constant motion do not give rise to responses of the detector. Other typical types of events that can be detected by the proposed method are splits and unifications of image structures. In figure 2b, the interest point is detected at the moment and the position corresponding to the collision of a ball and a wall. Similarly, interest points are detected at the moment of collision and bouncing of two balls as shown in figure 2c-d. Note, that different types of events are detected depending on the scale of observation.

To further emphasize the importance of the spatial and the temporal scales of observation, let us consider an oscillating signal with different spatial and temporal frequencies defined by $f(x, y, t) = \text{sgn}(y - \sin(x/4))\sin(t/4))$, where $\text{sgn}(u) = 1$ for $u > 0$ and $\text{sgn}(u) = -1$ for $u < 0$ (see figure 3). A scan can be seen from the illustration, the result of detecting the strongest interest points highly depends on the choice of the scale parameters $\sigma^2 l$ and $\tau^2 l$.

We can observe that space-time structures with long temporal extents are detected for large values of $\tau^2 l$ while short events are preferred by the detector with small values of $\sigma^2 l$.

Synthetic image results (multi-scale)

\[ \sigma_l^2 = 8, \tau_l^2 = 8 \]

\[ \sigma_l^2 = 2, \tau_l^2 = 8 \]

\[ \sigma_l^2 = 8, \tau_l^2 = 2 \]

\[ \sigma_l^2 = 2, \tau_l^2 = 2 \]

Figure 3: Results of detecting interest point at different spatial and temporal scales for a synthetic sequence with impulses having varying extents in space and time. The extents of the detected events roughly correspond to the scale parameters \( \sigma_l^2 \) and \( \tau_l^2 \) used for computing \( H(8) \).

From the presented examples, we can conclude that a correct selection of temporal and spatial scales is crucial when capturing events with different spatial and temporal extents. Moreover, estimating the spatio-temporal extents of events is important for their further interpretation. In the next section, we will present a mechanism for simultaneous estimation of spatio-temporal scales. This mechanism will then be combined with the interest point detector resulting in scale-adapted interest points in section 3.2.

3 Spatio-temporal scale adaptation

3.1 Scale selection in space-time

During recent years, the problem of automatic scale selection has been addressed in several different ways, based on the maximization of normalized derivative expressions over scale, or the behavior of entropy measures or error measures over scales (see Lindeberg and Bretzner (2003) for a review). To estimate the spatio-temporal extent of an event in [Laptev & Lindenberg (2004) "On Space-Time Interest Points"]
Real video, multi-scale, with scale selection

Figure 7: Results of detecting spatio-temporal interest points from the motion of the legs of a walking person. (a): 3-D plot with a thresholded level surface of a leg pattern (here shown upside down to simplify interpretation) and the detected interest points illustrated by ellipsoids; (b): spatio-temporal interest points overlayed on single frames in the original sequence.

The top rows of figure 8 show interest points detected in an outdoor sequence with a walking person and a zooming camera. The changing values of the selected spatial scales (illustrated by the size of the circles) illustrate the invariance of the method with respect to spatial scale changes of the image structures. Note that besides events in the leg pattern, the detector finds spurious points due to the non-constant motion of the coat and the arms. Image structures with constant motion in the background, however, do not result in the response of the detector. The pure spatial interest operator on the contrary gives strong responses in the static background as shown at the bottom row of figure 8.

The third example explicitly illustrates how the proposed method is able to estimate the temporal extent of detected events. Figure 9 shows a person making hand-waving gestures with a high frequency on the left and a low frequency on the right. The distinct interest points are detected at the moments and at the spatial positions where the palm of hand changes direction from motion. Whereas the selected spatial scale remains constant, the selected temporal scale depends on the frequency of the wave pattern. The high frequency pattern results in short events and gives rise to interest points with small temporal extent (see figure 9a). Correspondingly, hand motions with low frequency result in interest points with long temporal extent as shown in figure 9b.

3 Here, we used the scale-adapted Harris interest point detector (Mikolajczyk and Schmid, 2001) that detects maxima of $H_{sp}$ in space and extrema of normalized Laplacian operator over scales (Lindeberg, 1998).