Efficient feature extraction, encoding and classification for action recognition
Vadim Kantorov, Ivan Laptev
INRIA – WILLOW / École Normale Supérieure, Paris, France

**Goal**
- Fast action recognition.
- State-of-the-art performance.

**Motivation**
- Huge amounts of video: Decades of TV channels
- High speed video transfer each year: >100x speed
- Current state-of-the-art methods for action recognition typically process ~1 frame per second

**Contributions**
- >100x speed-up of video feature extraction.
- 4x real-time action recognition (CPU).
- Minor decrease in recognition accuracy.
- Publicly available implementation

**Related work**

**MPEG flow**
- Estimated motion vectors are part of the most compressed video representations: MPEG, H.264, VP9.
- MPEG motion vectors are sparse, typically defined on a 16x16 pixel grid.
- The quality of MPEG flow is comparable to motion estimation by standard Optical Flow algorithms.

Motion in the synthetic MPI Sintel Flow dataset:
- Quantized ground truth flow
- Quantized MPEG flow, err<0.263

Motion in movie frames:
- Quant. LK flow, err<0.314
- Quant. Farneback flow, err<0.296

**Local motion descriptor**
- Use sparse MPEG flow vectors to compute HOF: Histograms of flow MBH: Motion boundary histograms
- Grid cells of two scales: 16x16 pixels, 5 frames 24x24 pixels, 5 frames
- Dense descriptor sampling with 16 pixels spatial stride 5 frames temporal stride

**Descriptor aggregation**
- Feature encoding and classification schemes: Histogram encoding + \( \phi \)-dimensional signature encoding + linear SVM Fisher Vector [1] + linear SVM
- Descriptor assignment using approximate Nearest Neighbor search (FLANN) [4].
- Approximate FV aggregation with updates of five nearest centroids only.

**Results**

**Hollywood2**
- Classification (mAP): MF [out] DT [1], MBH [out] DT [1]
- Speed (fps): MF [1], MBH [1]

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>MF (out)</th>
<th>DT [1]</th>
<th>MBH (out)</th>
<th>DT [1]</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOF</td>
<td>47.2%</td>
<td>52.0%</td>
<td>346.8</td>
<td>330.3</td>
</tr>
<tr>
<td>MBHx</td>
<td>49.0%</td>
<td>52.0%</td>
<td>330.3</td>
<td></td>
</tr>
<tr>
<td>MBHy</td>
<td>50.4%</td>
<td>56.1%</td>
<td>330.3</td>
<td></td>
</tr>
<tr>
<td>MBH+MBHx+MBHy</td>
<td>53.9%</td>
<td>58.0%</td>
<td>218.7</td>
<td></td>
</tr>
<tr>
<td>MBH+MBHx+MBHy+HOG</td>
<td>56.2%</td>
<td>60.0%</td>
<td>166.4</td>
<td>1.2</td>
</tr>
</tbody>
</table>

**Parameter sensitivity**
- Sampling stride mAP fps
  - 16: 58.5% 15.3
  - 8: 58.6% 24.1
  - 4: 59.2% 13.7

- OF stride marginally affects accuracy
- Stable recognition across codecs and bit-rates

**Comparison to the state of the art**

**HMDDB 51**
- Acc.Feat. Quant. Total
- MF FLANN4-32 55.8% 16.4 32.2 60.0
- MF VLASTD4 55.7% 16.5 34.9 70.4
- MF VLASTD2 55.7% 16.5 32.2 68.9
- DT [1] 59.9% 1.2 5.1 6.4

**UCF 50**
- Acc.Feat. Quant. Total
- MF ALL FV32 46.7% 455.6 129.7 515.3
- MF ALL FV32 46.7% 455.6 129.7 515.3
- MF ALL VLASTD32 46.7% 455.6 129.7 515.3
- MBH [2] HOG+FV [1] 33.3% 40.6 240.8 284.4
- MBH [2] FV [1] 48.5% 40.6 240.8 284.4

**Code available**
http://www.di.ens.fr/willow/research/fastvideofeat