Action Recognition with Improved Trajectories
Heng WANG and Cordelia SCHMID
LEAR team, INRIA, France

**Contribution**
- Improve dense trajectories by explicit camera motion estimation
- Use human detection to remove outlier matches for robust homography estimation
- Remove trajectories caused by camera motion, warp optical flow to eliminate camera motion

**Dense trajectories revisited**
- State of the art results on a large variety of datasets for action classification
- MBH is based on derivatives of optical flow, thus can suppress constant camera motion
- Many irrelevant trajectory features in the background due to camera motion
- Motion descriptors based on optical flow can be corrupted due to camera motion, e.g., HOF, MBH

**Camera motion estimation**
- Match feature points between frames using SURF descriptors (green) and dense optical flow (red)
- The combination of SURF and optical flow results in a more balanced distribution
- Use RANSAC to estimate a homography from all feature matches

**Remove inconsistent matches due to humans**
- Human motion is not constrained by camera motion
- Apply a human detector in each frame, and remove feature matches inside the human bounding box

**Warp optical flow**
- Warp the second frame of two consecutive frames with the homography and recompute the optical flow
- For the HOF descriptor, the warped flow removes irrelevant camera motion, thus only encodes foreground motion
- For the MBH descriptor, it also improves, as the motion boundaries are enhanced

**Remove background trajectories**
- Remove trajectories due to camera motion, which has similar effects as sampling features with visual saliency
- Our method works well under various camera motions, such as pan, zoom, tilt

**Experimental setting**
- "RootSIFT" 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>Trajectory</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>25.4%</td>
<td>38.4%</td>
<td>39.5%</td>
<td>49.1%</td>
<td>49.8%</td>
<td>52.2%</td>
</tr>
<tr>
<td>WarpFlow</td>
<td>31.0%</td>
<td>38.7%</td>
<td>48.5%</td>
<td>50.9%</td>
<td>53.5%</td>
<td>55.6%</td>
</tr>
<tr>
<td>RmTrack</td>
<td>26.9%</td>
<td>39.6%</td>
<td>41.6%</td>
<td>50.8%</td>
<td>51.0%</td>
<td>53.9%</td>
</tr>
<tr>
<td>ITF</td>
<td>32.4%</td>
<td>40.2%</td>
<td>48.9%</td>
<td>52.1%</td>
<td>54.7%</td>
<td>57.2%</td>
</tr>
</tbody>
</table>

**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>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>

**Impact of human detection and state of the art**

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Hollywood2</th>
<th>HMDB51</th>
<th>Olympic Sports</th>
<th>UCF50</th>
</tr>
</thead>
<tbody>
<tr>
<td>With HD</td>
<td>84.3%</td>
<td>85.7%</td>
<td>91.1%</td>
<td>91.2%</td>
</tr>
<tr>
<td>Without HD</td>
<td>83.0%</td>
<td>85.5%</td>
<td>90.2%</td>
<td>90.5%</td>
</tr>
</tbody>
</table>

Source code: http://lear.inrialpes.fr/~wang/improved_trajectories