Spot On: Action Localization from Pointly-Supervised Proposals

Pascal Mettes
University of Amsterdam

Jan C. van Gemert
Delft University of Technology

Cees G. M. Snoek
University of Amsterdam
Goal

Kissing

Shaking hands
Related work: Action proposals at test time

**Supervoxels**
- Jain et al. CVPR’14
- Oneata et al. ECCV’14

**Trajectories**
- van Gemert et al. BMVC’15
- Puskas et al. ICCV’15

**Tracking/detection**
- Yu et al. CVPR’15
- Weinzaepfel et al. ICCV’15
Related work: Action proposals at test time

**Supervoxels**
- Jain et al. CVPR’14
- Oneata et al. ECCV’14

**Trajectories**
- van Gemert et al. BMVC’15
- Puskas et al. ICCV’15

**Tracking/detection**
- Yu et al. CVPR’15
- Weinzaepfel et al. ICCV’15

Action proposals
Related work: Training from box supervision

Annotate boxes for each frame of each train video.
Related work: Training from box supervision

Annotate boxes for each frame of each train video.

- **J-HMDB** Jhuang et al. ICCV’13
- **UCF Sports** Rodriguez et al. CVPR’08
- **UCF 101** Soomro et al. ’12

![Bar chart showing average number of frames](chart.png)
Our hypothesis

Training on bounding boxes not required. Training on proposals with fast point annotations is as effective.

Annotation time for video: 5 min. 11 sec.  
Annotation time for video: 25 sec.
Our contribution

- Human point supervision
- Compute proposal affinity
- Mine best proposal
Mining the best proposals

Train action classifiers using only best proposals.
Casted as a Multiple Instance Learning problem.
Mining the best proposals

Train action classifiers using only best proposals. Casted as a Multiple Instance Learning problem.

Use affinity with point annotations to guide the mining.
Proposal affinity

Novel overlap measure between point annotations and proposals.

No overlap
Small overlap
High overlap
Proposal affinity

Affinity = Proposal Match - Size Regularization
Proposal match

Affinity = Proposal Match - Size Regularization

Each point should match with the center of the proposal. Average the matches over all the points.

Match: 0.0

Match: 0.1

Match: 1.0
Size regularization

Affinity = Proposal Match - Size Regularization

Subtract the size of the proposal from the match.
To alleviate center bias of large proposals.
Mining recap

Point and Proposal Affinities

Select best proposals

Our MIL optimization

Create classifiers
Experiments

Unsupervised proposals from clustered trajectory features. Evaluated with Fisher Vectors and SVMs.

van Gemert et al. *BMVC’15*
Training without ground truth tubes

Best possible proposal performs as well as ground truth tubes.
Training without ground truth tubes

Best possible proposal performs as well as ground truth tubes.
Training without ground truth tubes

Mean AP maintained using our mined proposals.
Training without ground truth tubes

Similar performance from different tubes.

Ground truth box annotations
Our mined proposal
Lowering the annotation frame-rate

Points as effective as boxes, while faster to annotate.
Lowering the annotation frame-rate

Up to 50 times speed-up at similar performance.
Lowering the annotation frame-rate

Up to 50 times speed-up at similar performance.
Lowering the annotation frame-rate

Points are fast. Competitive even at 10% annotation effort.
Hollywood2Tubes

Dataset to demonstrate how easy action annotation becomes. Contains actions and instances new to action localization.

Download: tinyurl.com/hollywood2tubes

Multi-label videos.
Contextual actions.
Group interactions.

Videos from Hollywood2 by Marszalek et al. CVPR’09
Hollywood2Tubes - good example

Localization result

![Graph showing localization result with threshold and average precision.](image-url)
Hollywood2Tubes - bad example

Localization result

- Getting out of a car

Average Precision

Threshold
Hollywood2Tubes - Overall performance

Action recall comparable to current datasets.
Action localization leaves much room for improvement.
Conclusions

Carefully annotated boxes not required for action localization.

We train on unsupervised proposals, guided by point annotations.

We introduce Hollywood2Tubes.
<table>
<thead>
<tr>
<th>Method</th>
<th>Supervision</th>
<th>UCF Sports AUC</th>
<th>UCF 101 mAP</th>
<th>Hollywood2Tubes mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lan et al. [14]</td>
<td>box</td>
<td>0.380</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Tian et al. [1]</td>
<td>box</td>
<td>0.420</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Wang et al. [18]</td>
<td>box</td>
<td>0.470</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Jain et al. [2]</td>
<td>box</td>
<td>0.489</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Chen et al. [20]</td>
<td>box</td>
<td>0.528</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>van Gemert et al. [4]</td>
<td>box</td>
<td>0.546</td>
<td>0.345</td>
<td>-</td>
</tr>
<tr>
<td>Soomro et al. [5]</td>
<td>box</td>
<td>0.550</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Gkioxari et al. [15]</td>
<td>box</td>
<td>0.559</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Weinzaepfel et al. [16]</td>
<td>box</td>
<td>0.559</td>
<td>0.468</td>
<td>-</td>
</tr>
<tr>
<td>Jain et al. [47]</td>
<td>zero-shot</td>
<td>0.232</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Cinbis et al. [8]*</td>
<td>video label</td>
<td>0.278</td>
<td>0.136</td>
<td>0.009</td>
</tr>
<tr>
<td>This work</td>
<td>points</td>
<td>0.545</td>
<td>0.348</td>
<td>0.143</td>
</tr>
<tr>
<td>Method</td>
<td>Supervision</td>
<td>UCF Sports AUC</td>
<td>UCF 101 mAP</td>
<td>Hollywood2Tubes mAP</td>
</tr>
<tr>
<td>----------------------</td>
<td>---------------</td>
<td>----------------</td>
<td>-------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>Lan et al. [14]</td>
<td>box</td>
<td>0.380</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Tian et al. [1]</td>
<td>box</td>
<td>0.420</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Wang et al. [18]</td>
<td>box</td>
<td>0.470</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Jain et al. [2]</td>
<td>box</td>
<td>0.489</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Chen et al. [20]</td>
<td>box</td>
<td>0.528</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>van Gemert et al. [4]</td>
<td>box</td>
<td>0.546</td>
<td>0.345</td>
<td>-</td>
</tr>
<tr>
<td>Soomro et al. [5]</td>
<td>box</td>
<td>0.550</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Gkioxari et al. [15]</td>
<td>box</td>
<td>0.559</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Weinzaepfel et al. [16]</td>
<td>box</td>
<td>0.559</td>
<td>0.468</td>
<td>-</td>
</tr>
<tr>
<td>Jain et al. [47]</td>
<td>zero-shot</td>
<td>0.232</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Cinbis et al. [8]*</td>
<td>video label</td>
<td>0.278</td>
<td>0.136</td>
<td>0.009</td>
</tr>
<tr>
<td>This work</td>
<td>points</td>
<td>0.545</td>
<td>0.348</td>
<td>0.143</td>
</tr>
<tr>
<td>Method</td>
<td>Supervision</td>
<td>UCF Sports AUC</td>
<td>UCF 101 mAP</td>
<td>Hollywood2Tubes mAP</td>
</tr>
<tr>
<td>---------------------</td>
<td>-------------</td>
<td>----------------</td>
<td>-------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>Lan et al. [14]</td>
<td>box</td>
<td>0.380</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Tian et al. [1]</td>
<td>box</td>
<td>0.420</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Wang et al. [18]</td>
<td>box</td>
<td>0.470</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Jain et al. [2]</td>
<td>box</td>
<td>0.489</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Chen et al. [20]</td>
<td>box</td>
<td>0.528</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>van Gemert et al. [4]</td>
<td>box</td>
<td>0.546</td>
<td>0.345</td>
<td>-</td>
</tr>
<tr>
<td>Soomro et al. [5]</td>
<td>box</td>
<td>0.550</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Gkioxari et al. [15]</td>
<td>box</td>
<td>0.559</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Weinzaepfel et al. [16]</td>
<td>box</td>
<td>0.559</td>
<td>0.468</td>
<td>-</td>
</tr>
<tr>
<td>Jain et al. [47]</td>
<td>zero-shot</td>
<td>0.232</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Cinbis et al. [8]*</td>
<td>video label</td>
<td>0.278</td>
<td>0.136</td>
<td>0.009</td>
</tr>
<tr>
<td>This work</td>
<td>points</td>
<td>0.545</td>
<td>0.348</td>
<td>0.143</td>
</tr>
</tbody>
</table>
4- Comparison to state-of-the-art

Competitive to boxes, better than other weak supervision.
Related work: Training from box supervision

Annotate boxes for *each* frame of *each* video.

<table>
<thead>
<tr>
<th>Database</th>
<th>Number of videos</th>
<th>Average number of frames</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCF 101</td>
<td>3,500</td>
<td>750</td>
</tr>
<tr>
<td>MSR-II Cao et al. CVPR'10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UCF Sports Rodriguez et al. CVPR'08</td>
<td></td>
<td></td>
</tr>
<tr>
<td>J-HMDB Jhuang et al. ICCV'13</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>