Hybrid Active Learning via Deep Clustering for Video Action Detection
THU-AM-228

Aayush J Rana, Yogesh S Rawat
Center for Research in Computer Vision (CRCV)
University of Central Florida (UCF)
Challenges

• Training requires dense annotation
  • Dense data $\propto$ large annotation cost
• Unnecessary cost
  • Repetitive nearby frames
  • Unrelated frames annotated
• Comparison across videos
  • Varying length
  • Varying actors
  • Class-wise difficulty
  • No difficulty metric
Previous work

• Annotation selection at frame level
• Assumes all videos annotated
  • Partial annotations
  • No metric to compare between videos
Motivation

• Reduce annotation cost
  • Video level selection
  • Frame level selection
  • Remove redundant videos
• Enable video comparison using
  • Informativeness
  • Diversity
• Improve sparse training
  • Improve pseudo-label usage
Contributions

• Hybrid selection (CLAUS)
  • AL based strategy
  • Video + frame selection
  • Uncertainty based video ranking
  • Clustering based video selection

• Improved pseudo-label loss (STeW)
  • Pixel-level weight
  • BG/FG consistency
Proposed approach

Model Training

Labeled Videos

Updated Annotations

Inter-sample Selection

CLAUS Hybrid AL

Intra-Sample

\[ \sum A_t \rightarrow V_{score} \]
Model Training Objectives

• Classification loss
• Localization loss
  • Spatio-Temporally Weighted loss (STeW)
  • Uses pixel-level consistency as weight
• Cluster loss
  • K arbitrary clusters
  • Adjust centers using video features

\[
\min_{\theta} \mathcal{L} = \mathcal{L}^{Cluster} + \mathcal{L}^{STeW} + \mathcal{L}^{Cls}
\]
Intra-sample selection

• Frame-level selection
• Uncertainty based score
• Distance based redundancy reduction
• Top $t$ frames used for video score
Inter-sample selection

- Video-level selection
- Video score from intra-sample
- Cluster assignment per video
- Top $V$ videos per cluster selected
  - Frames from intra-sample
Datasets

• UCF-101
  • 3207 videos
  • 24 action classes
  • Spatio-temporal bounding box annotation

• J-HMDB
  • 928 videos
  • 21 action classes
  • Spatio-temporal pixel-wise annotation
Comparing with baselines

<table>
<thead>
<tr>
<th>Method</th>
<th>A%</th>
<th>UCF-101-24</th>
<th>J-HMDB-21</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>v-mAP</td>
<td>f-mAP</td>
</tr>
<tr>
<td>Random</td>
<td>1%</td>
<td>52.6</td>
<td>54.1</td>
</tr>
<tr>
<td>Equi.</td>
<td>1%</td>
<td>53.3</td>
<td>55</td>
</tr>
<tr>
<td>Entropy [1] †</td>
<td>1%</td>
<td>52.2</td>
<td>53.5</td>
</tr>
<tr>
<td>Uncertainty [14] †</td>
<td>1%</td>
<td>44.0</td>
<td>46.7</td>
</tr>
<tr>
<td>Our</td>
<td>1%</td>
<td>61.8</td>
<td>61.6</td>
</tr>
<tr>
<td>Random</td>
<td>5%</td>
<td>67.5</td>
<td>67.3</td>
</tr>
<tr>
<td>Equi.</td>
<td>5%</td>
<td>67.2</td>
<td>67.0</td>
</tr>
<tr>
<td>Entropy [1] †</td>
<td>5%</td>
<td>71.3</td>
<td>70.2</td>
</tr>
<tr>
<td>Uncertainty [14] †</td>
<td>5%</td>
<td>69.7</td>
<td>68.2</td>
</tr>
<tr>
<td>Our</td>
<td>5%</td>
<td>72.2</td>
<td>72.1</td>
</tr>
</tbody>
</table>
Comparing with prior weakly-supervised

<table>
<thead>
<tr>
<th>Method</th>
<th>A%</th>
<th>f-mAP@</th>
<th>v-mAP@</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.5</td>
<td>0.1</td>
</tr>
<tr>
<td>Mettes et al. [40]</td>
<td>V</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Escorcia et al. [12]</td>
<td>V</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Zhang et al. [67]</td>
<td>V</td>
<td>30.4</td>
<td>62.1</td>
</tr>
<tr>
<td>Arnab et al. [3]</td>
<td>V</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Mettes et al. [39]</td>
<td>P</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Cheron et al. [9]</td>
<td>P</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Weinz. et al. [64]</td>
<td>1.1%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Weinz. et al. [64]</td>
<td>2.8%</td>
<td>63.8</td>
<td>-</td>
</tr>
<tr>
<td>MixMatch [5]</td>
<td>S-20%</td>
<td>20.2</td>
<td>-</td>
</tr>
<tr>
<td>Pseudo-label [32]</td>
<td>S-20%</td>
<td>64.9</td>
<td>-</td>
</tr>
<tr>
<td>Co-SSD(CC) [24]</td>
<td>S-20%</td>
<td>65.3</td>
<td>-</td>
</tr>
<tr>
<td>Kumar et al. [31]</td>
<td>S-20%</td>
<td>69.9</td>
<td>-</td>
</tr>
<tr>
<td>Ours</td>
<td>1%</td>
<td>61.6</td>
<td>98.1</td>
</tr>
<tr>
<td>Ours</td>
<td>5%</td>
<td>72.1</td>
<td>98.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>A%</th>
<th>f-mAP@</th>
<th>v-mAP@</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.5</td>
<td>0.1</td>
</tr>
<tr>
<td>Zhang et al. [67]</td>
<td>V</td>
<td>65.9</td>
<td>81.5</td>
</tr>
<tr>
<td>Weinz. et al. [64]</td>
<td>6%</td>
<td>50.7</td>
<td>-</td>
</tr>
<tr>
<td>Weinz. et al. [64]</td>
<td>15%</td>
<td>56.5</td>
<td>-</td>
</tr>
<tr>
<td>MixMatch [5]</td>
<td>S-30%</td>
<td>7.5</td>
<td>-</td>
</tr>
<tr>
<td>Pseudo-label [32]</td>
<td>S-30%</td>
<td>57.4</td>
<td>-</td>
</tr>
<tr>
<td>Co-SSD(CC) [24]</td>
<td>S-30%</td>
<td>60.7</td>
<td>-</td>
</tr>
<tr>
<td>Kumar et al. [31]</td>
<td>S-30%</td>
<td>64.4</td>
<td>-</td>
</tr>
<tr>
<td>Ours</td>
<td>1%</td>
<td>61.9</td>
<td>99.0</td>
</tr>
<tr>
<td>Ours</td>
<td>5%</td>
<td>72.7</td>
<td>99.1</td>
</tr>
</tbody>
</table>

J-HMDB-21
Action Detection Results

Soccer Juggling

Salsa Dancing

Floor Gymnastics

Horse Riding

Long Jumping

Red: GT
Blue: Our detection
Cluster representation

a: CLAUS

b: Entropy

c: Uncertainty

d: Random
Cluster effectiveness

![Graphs showing the relationship between cluster effectiveness and % Frames annotated.](image)
Loss effectiveness
Selection method analysis
Summary

• Hybrid selection improves performance
  • Clustering-aware selection strategy
  • Reduces similar video
  • Enables inter-sample comparison

• $STeW$ loss improves sparse label training
Thank You

Project Link: https://tinyurl.com/hybridclaus