Active Learning of an Action Detector from Untrimmed Videos
Sunil Bandla and Kristen Grauman
Department of Computer Science, University of Texas at Austin

**Motivation**
- Realistic unlabeled videos are "untrimmed" to temporal regions of interest, and each video contains multiple actions.
- This yields unlabeled feature distribution where useful and redundant candidates are hard to distinguish for active learning.

**Main Idea**
- We introduce a detection-based active learning approach to select videos for annotation, while accounting for their untrimmed nature.
- Voting-based detector is robust to partial evidence and supports fast incremental updates during active learning.
- Learn accurate action recognition models with fewer annotations.

**Hough-based Action Detector**
**Building the Detector**
- Extract HoG/HoF features at STIPs detected in training videos and build Hough tables, and sort words by discriminative power.

**Applying the Detector to a Novel Video**
- Use the Hough table entries to vote on the probable action centers.
- Reduces number of candidate intervals per video for active selection.

**Active Selection of Untrimmed Videos**
We seek the unlabeled video that, if used to augment the action detector, will more confidently localize actions in all unlabeled videos.

$$v^* = \arg\max_{v \in \mathcal{X}} \max_{T \in \mathbb{Z}} S(T \cup v^*),$$

where $T$ is the training set, $\mathbb{Z} = \{1, -1\}$ is the set of possible labels, and $v^*$ denotes that the unlabeled video has been given label $I$.

- Treating the unlabeled video as positive, we score the value of probable action intervals in the video to the current detector $I$.

$$S(T \cup v^*) = \max_{k=1,...,K} \text{VALUE}(D(T \cup v^*_k))$$

- Treating $v$ as negative,

$$S(T \cup v^*) = \text{VALUE}(D(T \cup v^-_k))$$

where VALUE is our novel entropy-based detector confidence defined below.

**Estimating Detector Confidence with Space-Time Entropy**
- Quantize unlabeled video's 3D vote space and compute normalized entropy
- A vote space with good cluster(s) indicates consensus on the location(s) of the action

**Annotations**
- Our interface that annotators use to label action intervals in the actively requested videos.
- Available on the project webpage.

**Results**

**UT Interaction (6 classes)**
- Passive vs. Active: Annotation effort saved by intelligent label requests.
- Active Classifier < Ours: Accounting for untrimmed nature of video is critical.
- Active Entropy < Ours: Simply estimating individual video uncertainty is insufficient.
- Active GT-Ints > Active Pred-Ints: Room for improvement in interval estimates.

**MSR Actions 1 (3 classes)**

<table>
<thead>
<tr>
<th>Train Set</th>
<th>HandShake</th>
<th>Hug</th>
<th>Kick</th>
<th>Point</th>
<th>Punch</th>
<th>Push</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial L ex only</td>
<td>0.1981</td>
<td>0.3029</td>
<td>0.1466</td>
<td>0.0107</td>
<td>0.1094</td>
<td>0.2022</td>
</tr>
<tr>
<td>After 15 rounds active</td>
<td>0.2574</td>
<td>0.3904</td>
<td>0.2175</td>
<td>0.0164</td>
<td>0.1758</td>
<td>0.2689</td>
</tr>
<tr>
<td>Full train set (42 ex)</td>
<td>0.3218</td>
<td>0.3218</td>
<td>0.0478</td>
<td>0.1987</td>
<td>0.2022</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test Set</th>
<th>Clapping</th>
<th>Waving</th>
<th>Boxing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial L ex only</td>
<td>0.2288</td>
<td>0.2318</td>
<td>0.1135</td>
</tr>
<tr>
<td>After 10 rounds active</td>
<td>0.3739</td>
<td>0.3134</td>
<td>0.1043</td>
</tr>
<tr>
<td>Full train set (27 ex)</td>
<td>0.3132</td>
<td>0.2582</td>
<td>0.0819</td>
</tr>
</tbody>
</table>

Our active method achieves good accuracy using much less annotations.