**Motivation**

- Collecting action examples in video is even more tedious than labeling objects in images because temporal localization is required.
- We propose an unsupervised method to collect action examples from instructional videos.

**Intuition:** To help humans learn, the instructor will verbally say and visually demonstrate an action. The audio-visual correlation is the key to our method.

**Assumption:** If the instructor says an “action signer” + “action”, the action is going to be performed immediately in the video.

- Example action signifiers: “we will”, “we’ll”, “let us”, “let’s”, “going to”, “gonna”.

**Methodology**

1. Locate sentence with action signer.
   - **Instructional videos from video sharing websites.**
   - **Caption:** You are going to bend your left leg.
   - **Method:**
     - **Direct object (dobj) relation:** a verb and its accusative object.
     - Ex: bend leg, cut paper, cut tomatoes, cut onion, pour milk, throw ball, lift heel, bend elbows, drill hole.
     - Extracted with the Stanford Parser [3].

2. Find dobj relation in the sentence.
   - **Method:**
     - Pipeline is run over the whole data set to collect examples covering diverse actions.
   - Non-visual dobj relations such as “take care”, “ask me” are filtered out with pre-defined ontology of visible nouns.

3. Localize action segment in video with speech timing of dobj relation.

**Data Collection**

- Acquired 73,000 instructional videos with closed-captions from various sites.
- 25,835 examples belonging to 541 actions were extracted.

**Precision of collected action examples**

- Manually verified the precision of 38 actions.
- High audio-visual correlation in Yoga. Lower correlation in ball sports, but nouns are still 73% correct.

**Multimedia Event Detection (MED)**

- Compared generalization ability of different action datasets on the MED task.
- SVM detectors are trained on collected data using improved trajectory [1] fisher vectors [2].
- Action detectors are predicted onto each video to acquire an intermediate representation for each video.
- Randomly sampled action category detectors from the intermediate representation for MED and compute MAP on MED11 Development set (9746 videos).

**Advantage of unsupervised algorithm:**

1. As more instructional videos are uploaded, we acquire more data with no extra manual labor cost.
2. Action examples are associated with context. Verb and noun concepts are jointly learned.
3. Diverse actions are harvested as more instructional videos are collected.
4. The same action in diverse settings are also harvested: e.g., drilling hole in wood, walls and ice.

**Remarkable results**

- Table: Verb and Noun
<table>
<thead>
<tr>
<th>Category</th>
<th>Both Correct</th>
<th>Verb Correct</th>
<th>Noun Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ball Sports</td>
<td>0.296</td>
<td>0.280</td>
<td>0.727</td>
</tr>
<tr>
<td>Cooking</td>
<td>0.643</td>
<td>0.092</td>
<td>0.512</td>
</tr>
<tr>
<td>Exercise / Yoga</td>
<td>0.866</td>
<td>0.889</td>
<td>0.948</td>
</tr>
<tr>
<td>Sewing / Tools</td>
<td>0.727</td>
<td>0.735</td>
<td>0.460</td>
</tr>
<tr>
<td>Average</td>
<td>0.618</td>
<td>0.548</td>
<td>0.519</td>
</tr>
</tbody>
</table>

**References**