SSS-31: Automatic Action Video Dataset Construction from Web using Density-based Cluster Analysis and Outlier Detection
Do Hang Nga and Yanai Keiji (The University of Electro-Communications, Tokyo)

Introduction
- Previous work: require additional data (e.g., tags[3]), ignore concept diversity problem
- This work: exploits only visual features of Web videos, copes with concept diversity

Proposed Approach
- Action Concept → Web videos
- Automatic Approach → Relevant Shots
- Action Dataset

i. Word preparation
   - “verb” (dive), “verb+non-verb” (throw hammer), “non-verb” (vault)
ii. Video search
   - “verb” & “verb-ing” (dive & diving)
iii. Video filtering
   - No videos of “Entertainment”
iv. Video downloading
   - Web API (e.g., Youtube API)
v. Shot segmentation
   - Color histogram

Shot Collection
- Video search, download & segmentation

Shot Clustering with OPTICS[1]
- A low reachability distance indicates an object within a cluster.
- A high reach-dist indicates a noise or a jump from one cluster to another.

LOF (Local Outlier Factor) [5]
- \( \text{LOF}_{\text{MinPts}}(p) = \frac{\sum_{o \in \text{MinPts}} \text{dist}(p)}{\text{MinPts} - \text{dist}(o)} \)
- Small \( \text{MinPts} - \text{dist} \) corresponds to a region with high density.
- Shots with low LOF are considered as relevant shots and ranked to the top.
- Shots are selected from all clusters to guarantee diversity of selection results.

Experiments and Results

Experiment 1: Dataset Construction
- Data: Web videos (YouTube)
- Actions: 11 actions in UCF11[2]
- Precision rate = percentage of relevant shots among top 100 shots [3]
- Baseline[3]: VisualRank based method

<table>
<thead>
<tr>
<th>Action</th>
<th>Proposed</th>
<th>Baseline</th>
<th>Action</th>
<th>Proposed</th>
<th>Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>basketball</td>
<td>59</td>
<td>67</td>
<td>swing</td>
<td>36</td>
<td>22</td>
</tr>
<tr>
<td>biking</td>
<td>30</td>
<td>35</td>
<td>tennis_swing</td>
<td>38</td>
<td>37</td>
</tr>
<tr>
<td>diving</td>
<td>25</td>
<td>19</td>
<td>trampoline_jumping</td>
<td>42</td>
<td>44</td>
</tr>
<tr>
<td>golf_swing</td>
<td>59</td>
<td>52</td>
<td>volleyball_spiking</td>
<td>36</td>
<td>45</td>
</tr>
<tr>
<td>horse_riding</td>
<td>49</td>
<td>48</td>
<td>walking</td>
<td>25</td>
<td>11</td>
</tr>
<tr>
<td>soccer_juggling</td>
<td>76</td>
<td>72</td>
<td>Average</td>
<td>43.2</td>
<td>41.1</td>
</tr>
</tbody>
</table>

Experiment 2: Action Classification
- Dataset: UCF11[2]
- Precision = average of 25-fold validation
- Training data: standard data[2] & shots automatically obtained in Experiment 1
- With standard training data [2]: 81.5%