Analyzing the similarities of actions based on video clustering

Vu Gia Truong  Do Hang Nga  Keiji Yanai

The University of Electro-Communications, Tokyo
Research area: computer vision

• Our lab are working on
  – Video recognition

• We are working on
  – Finding videos which have visually similar content.
Background

– Verbs

- Express actions.
- Represent by video shots
Background

– Analyzing visual similarities between verbs has not been conducted.

Walking  Eating  Running  Writing

Which pair is most similar? Which pair is most dissimilar?

– Difficulties

• How to represent verb by computable unit.
• How to collect data for several different actions.
Research purpose

• Analyzing visual similarities between actions.
  = Finding verbs pairs whose corresponding actions look similar

  e.g. 「play+piano」 vs 「type+keyboard」
Method summary

1. Extract bag-of-spatial-temporal(BOSFT) vectors
   1. Extract spatial-temporal features
   2. Encode features using Bag-of-Features approach

2. Calculate attribution vector for each verb
   1. Cluster all of BOSFT vectors
   2. Calculate distribution rate per cluster

3. Evaluate similarities
   1. Calculate distances between each verb
   2. Rank distances
      – Low distance = high similarity
1. Extract bag-of-spatial-temporal (BOSFT) vectors

1. Extract spatial-temporal features
   • Spatial-temporal feature tracks motion of special points in each frame during time.
   • For each video shot, several spatial-temporal feature vectors are extracted

2. Encode features using Bag-of-Features approach
   • Bag-of-Features counts occurrence frequency of each feature vector based on a dictionary
   • Vector represent occurrence frequency of spatial-temporal features is called Bag-of-spatial-temporal vector.
1.1 Extract spatial-temporal features

1.2 Encode features using Bag-of-Features approach
2. Calculate attribution vector

1. Cluster all of BOSFT vectors
   • Use Probabilistic latent semantic analysis (pLSA) to cluster
     – Set cluster count from 100 to 200
   • Calculate attribution vector of each video shot

2. Calculate attribution vector for each verb
   • Calculate mean value of all attribution vectors of all video shots belong to each verb
2.1 Cluster all of BOSFT vectors

- Use pLSA method to cluster all of BOSFT vectors

<table>
<thead>
<tr>
<th>Cluster i</th>
<th>Video</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.006</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cluster (i +1)</th>
<th>Video</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.004</td>
</tr>
</tbody>
</table>
2.1 Cluster all of BOSFT vectors

- Attribution vector of each video shot
  - Is distribution rate per cluster
  - Has sum equal to 1

Video shot $j$ of verb

<table>
<thead>
<tr>
<th>Cluster 1: $x_{j1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 2: $x_{j2}$</td>
</tr>
<tr>
<td>...</td>
</tr>
<tr>
<td>Cluster $i$: $x_{ji}$</td>
</tr>
<tr>
<td>...</td>
</tr>
<tr>
<td>Cluster $n$: $x_{jn}$</td>
</tr>
</tbody>
</table>

Attribution vector $x_{j1}, x_{j2}, \ldots, x_{jn}$
2.2 Calculate distribution vector for each verb

Video shots: Attribution vector of each video shot

\[(x_{11}, x_{12}, \ldots, x_{1n})\]

\[\vdots\]

\[(x_{m1}, x_{m2}, \ldots, x_{mn})\]

Attribution vector of verb:

\[x_i = \frac{1}{m} \sum_{j=1}^{m} x_{ij}\]
3. Evaluate similarities

1. Calculate distances between each verb
   • Use Jeisen-Shannon divergence to calculate

2. Evaluate similarities based on distances
   • Rank calculated distances
   • Find pairs of verbs which are visually similar
3.1 Calculate distances between each verb

- Jesen-Shannon divergence

\[ D_{JS}(q \parallel p) = \lambda D_{KL}(q \parallel \lambda q + (1 - \lambda) p) \]
\[ + (1 - \lambda) D_{KL}(p \parallel \lambda q + (1 - \lambda) p) \]

\[ D_{KL}(f(x) \parallel g(x)) = \int f(x) \log \frac{f(x)}{g(x)} \, dx \]
3.2 Evaluate similarities

1. Find pairs of verbs which are visually similar
   • Rank calculated distances ascending
   • Find pairs of verbs which are visually similar
     – Low rank score = high similarity
Experiment

• Dataset: 13320 videos for 101 verbs

1. Evaluate visual similarities between verb pairs

2. Find verb pairs which are very similar
Experiment data

- Dataset: UCF-101
  - Contain 13320 videos for 101 verbs
  - Divide into 5 main categories
    - Human–Object Interaction
    - Body–Motion
    - Human–Human Interaction
    - Playing Musical Instruments
    - Sports
## Experiment result: similar verb pairs

<table>
<thead>
<tr>
<th>Verb A</th>
<th>Verb B</th>
<th>Absolute distance</th>
<th>Relative distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mixing</td>
<td>CuttingInKitchen</td>
<td>0.0157</td>
<td>1.00</td>
</tr>
<tr>
<td>ThrowDiscs</td>
<td>HammerThrow</td>
<td>0.0181</td>
<td>1.16</td>
</tr>
<tr>
<td>ShavingBeard</td>
<td>Haircut</td>
<td>0.0197</td>
<td>1.26</td>
</tr>
<tr>
<td>SoccerJuggling</td>
<td>Nunchucks</td>
<td>0.0205</td>
<td>1.31</td>
</tr>
<tr>
<td>ShavingBeard</td>
<td>ApplyLipstick</td>
<td>0.0222</td>
<td>1.42</td>
</tr>
<tr>
<td>Haircut</td>
<td>BrushingTeeth</td>
<td>0.0224</td>
<td>1.43</td>
</tr>
<tr>
<td>FrontCrawl</td>
<td>BreastStroke</td>
<td>0.0234</td>
<td>1.49</td>
</tr>
<tr>
<td>PlayingTabla</td>
<td>PlayingGuitar</td>
<td>0.0234</td>
<td>1.50</td>
</tr>
<tr>
<td>PlayingTabla</td>
<td>PlayingCello</td>
<td>0.0239</td>
<td>1.53</td>
</tr>
<tr>
<td>MilitaryParade</td>
<td>BandMarching</td>
<td>0.0241</td>
<td>1.54</td>
</tr>
</tbody>
</table>
Examples of visually similar verb pair

- “ThrowDiscus” and “HammerThrow”
  - Body swings around
Examples of visually similar verb pair

• “PlayingGuitar” and “PlayingTablar”
  – Hands move in small area
Examples of visually un-similar verb pair

• “SoccerJuggling” and “NunChuck”
  – Body moves
Examples of visually un-similar verb pair

- "Haircut" and "BrushingTeeth"
  - Hand moves
Conclusion

• Propose a novel approach to represent verbs
• Evaluate visual similarities between verbs
• Future work
  – Improve represent method to find visual similar verb pairs
  – Expand dataset
    • Add cooking related verbs
    • Add eating related verbs
Thank you for listening
Future work details

1. Improve representing method
   – Apply mid-level feature approach
     • Extract feature descriptors from volumes of video

2. Expand dataset
   – Add other verbs
     • Add cooking, eating related verbs
   – Add more sample video shots for each verb
     • Expand to 200 vide shots per verb
Contributions of our work

• Analysis similarities between verbs

• Can be applied to
  – Text-based video retrieval
  – Content-based video retrieval