Unsupervised Co-Segmentation for Athlete Movements and Live Commentaries Using Crossmodal Temporal Proximity

Yasunori Ohishi, Yuki Tanaka, and Kunio Kashino

NTT Corporation, Japan
Related work

Embedding model (DAVEnet) that can directly associate visual objects with spoken words [Harwath+2016]

Image network (Pre-trained VGG16)

Speech network (CNN-based)

Image

Audio caption

Higher similarity

• Triplet loss function
• Margin softmax loss function
• Noise contrastive estimation

• 400K English captions [Harwath+2019]
• 100K Hindi captions [Harwath+2018]
• 100K Japanese captions [Ohishi+2020]
Our challenge

Co-segmentation of sports actions and live commentary

Video frames

Mel-spectrogram

Temporal proximity

“はっけよいのこった” (Ready go!)

“正面からあたって” (Frontal attack)

“相手の上半身を強く押し、土俵の外へ出しました”  
(Push hard against the opponents upper body to force him out of the ring)

Copyright 2021 NTT CORPORATION
Guided attention scheme to efficiently detect and utilize temporal co-occurrences of audio and video information

Model

Video network (ECO) → Visual feature → Guided attention scheme

32 video frames (10-second video)

Speech network (DAVEnet) → Audio feature

Mel spectrogram (10-second audio)

Element-wise product → Mean pooling → Similarity

Time axis

Similarity matrix

Time axis

$G_{i,j} = \exp \left\{ \frac{-\left( \frac{i}{T} - \frac{j}{T} \right)^2}{2\sigma_x^2} \right\}$
Guided attention scheme to efficiently detect and utilize temporal co-occurrences of audio and video information

Model

Video network (ECO) → Visual feature → Audio network (DAVEnet)

32 video frames (10-second video)

Mel spectrogram (10-second audio)

Temporal pooling

Spatial and temporal pooling

Temporal information is averaged or discarded.

Dot product

Similarity

Existing approaches (Baseline)
Dataset

- 170 hours of NHK broadcast of grand sumo tournaments
- 1,218 matches of nine frequent winning techniques
- 10-second video clips and their raw audio waveforms centered around labeled times as audio-visual pairs

<table>
<thead>
<tr>
<th>Winning techniques</th>
<th>Training</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frontal push out</td>
<td>365</td>
<td>10</td>
</tr>
<tr>
<td>Frontal force out</td>
<td>362</td>
<td>10</td>
</tr>
<tr>
<td>Slap down</td>
<td>141</td>
<td>10</td>
</tr>
<tr>
<td>Thrust down</td>
<td>77</td>
<td>10</td>
</tr>
<tr>
<td>Over arm throw</td>
<td>45</td>
<td>10</td>
</tr>
<tr>
<td>Frontal thrust out</td>
<td>42</td>
<td>10</td>
</tr>
<tr>
<td>Frontal crush out</td>
<td>34</td>
<td>10</td>
</tr>
<tr>
<td>Rear push out</td>
<td>34</td>
<td>10</td>
</tr>
<tr>
<td>Frontal push down</td>
<td>28</td>
<td>10</td>
</tr>
</tbody>
</table>

1,128 90

10-second video 10-second audio
## Crossmodal search results

Audio-visual retrieval recall scores when the correct result was defined as the clips with the same winning techniques as the query.

<table>
<thead>
<tr>
<th>$\sigma_g$</th>
<th>Audio to Video</th>
<th></th>
<th>Video to Audio</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R@1</td>
<td>R@3</td>
<td>R@5</td>
<td>R@1</td>
<td>R@3</td>
</tr>
<tr>
<td>0.001</td>
<td>.289</td>
<td>.600</td>
<td>.739</td>
<td>.294</td>
<td>.611</td>
</tr>
<tr>
<td>0.01</td>
<td>.348</td>
<td>.656</td>
<td>.770</td>
<td>.304</td>
<td>.604</td>
</tr>
<tr>
<td>0.1</td>
<td>.304</td>
<td>.648</td>
<td>.763</td>
<td>.307</td>
<td>.581</td>
</tr>
<tr>
<td>1</td>
<td>.289</td>
<td>.600</td>
<td>.711</td>
<td>.211</td>
<td>.511</td>
</tr>
<tr>
<td>10</td>
<td>.211</td>
<td>.461</td>
<td>.611</td>
<td>.144</td>
<td>.389</td>
</tr>
<tr>
<td>100</td>
<td>.122</td>
<td>.389</td>
<td>.511</td>
<td>.056</td>
<td>.211</td>
</tr>
<tr>
<td>Baseline</td>
<td>.256</td>
<td>.422</td>
<td>.589</td>
<td>.233</td>
<td>.511</td>
</tr>
</tbody>
</table>
Our method better captures the correspondence between audio and visual information and the edges of the segments.