Audio-based Near-Duplicate Video Retrieval with Audio Similarity Learning

Pavlos Avgoustinakis
Giorgos Kordopatis-Zilos
Symeon Papadopoulos
Andreas L. Symeonidis
Ioannis Kompatsiaris
Problem statement

Duplicate Audio Video Retrieval (DAVR)

- Given a video query, search a video database and retrieve videos that share the same audio content

\[ f(\text{query audio}, \text{database audio}) = \text{similarity} \]
Motivation

State-of-the-art limitations
• Use of handcrafted approaches for fingerprint extraction
• Rigid aggregation schemes for similarity calculation
• Lack of evaluation benchmarks with user-generated content annotated based on audio duplicity

Our objective
• Leverage of deep learning and transfer learning
• Composition of evaluation benchmarks
Contribution

- Robust audio-based video similarity calculation
  - Transfer learning from a pre-trained CNN
  - Audio similarity learning
- Two annotated datasets that serve as benchmarks for DAVR
Feature Extraction

- Generate audio Mel-spectrograms of audio signals
  - Divide into overlapping time segments
- Employ a pre-trained CNN (Kumar et al. 2018)
  - Max Activation of Convolutions (MAC) on intermediate CNN layers
  - Apply PCA whitening and attention-based weighting

Similarity calculation

- Generation of the similarity matrix
  \[ S_{qp} = Q \cdot P^T \]

- Audio Similarity Learning network
  - Four-layer CNN
  - Captures the temporal structures

- Chamfer Similarity
  \[
  CS(q, p) = \frac{1}{X'} \sum_{i=1}^{X'} \max_{j \in [1, Y']} \text{Htanh}(S_{qp}^{ij}(i, j))
  \]
Evaluation datasets

**FIVR-200K**

- FIVR-200K (Kordopatis-Zilos et al., 2019) annotated for Fine-grained Incident Video Retrieval
- 76 video queries
- 3,392 audio duplicate pairs

**SVD**

- SVD (Jiang et al., 2019) annotated for Near-Duplicate Video Retrieval
- 167 video queries
- 1,492 audio duplicate pairs

Experimental results

- Duplicate Audio Video Retrieval (DAVR)

<table>
<thead>
<tr>
<th>Method</th>
<th>FIVR-200K_\alpha</th>
<th>SVD_\alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dejavu [17]</td>
<td>0.726</td>
<td>0.874</td>
</tr>
<tr>
<td>Spectro Slides [14]</td>
<td>0.588</td>
<td>0.716</td>
</tr>
<tr>
<td>Spectro Tiles [16]</td>
<td>0.510</td>
<td>0.605</td>
</tr>
<tr>
<td>AuSiL (ours)</td>
<td>0.701</td>
<td>0.940</td>
</tr>
</tbody>
</table>

- Audio speed transformations

<table>
<thead>
<tr>
<th>Method</th>
<th>FIVR-200K^T_\alpha</th>
<th>SVD^T_\alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dejavu [17]</td>
<td>0.443</td>
<td>0.741</td>
</tr>
<tr>
<td>AuSiL (ours)</td>
<td>0.865</td>
<td>0.923</td>
</tr>
</tbody>
</table>
Experimental results

• Visual-based video retrieval

<table>
<thead>
<tr>
<th>Method</th>
<th>FIVR-200K</th>
<th></th>
<th>SVD</th>
<th>EVVE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DSVR</td>
<td>CSVR</td>
<td>ISVR</td>
<td></td>
</tr>
<tr>
<td>Dejavu [17]</td>
<td>0.352</td>
<td>0.324</td>
<td>0.230</td>
<td>0.477</td>
</tr>
<tr>
<td>Spectro Slides [14]</td>
<td>0.288</td>
<td>0.269</td>
<td>0.189</td>
<td>0.406</td>
</tr>
<tr>
<td>Spectro Tiles [16]</td>
<td>0.249</td>
<td>0.228</td>
<td>0.159</td>
<td>0.323</td>
</tr>
<tr>
<td>AuSiL (ours)</td>
<td>0.327</td>
<td>0.310</td>
<td>0.232</td>
<td>0.516</td>
</tr>
<tr>
<td>Best visual</td>
<td>0.892</td>
<td>0.841</td>
<td>0.702</td>
<td>0.785</td>
</tr>
</tbody>
</table>
Thank you!

Get in touch:
Giorgos Kordopatis-Zilos: georgekordopatis@iti.gr / @g_kordo

MeVer team:
https://mever.iti.gr/web/ / @meverteam

Code available in:
https://github.com/mever-team/ausil

With the support of: