Activity Image-to-Video Retrieval by Disentangling Appearance and Motion

Liu Liu, Jiangtong Li, Li Niu, Ruicong Xu, Liqing Zhang

MoE Key Lab of Artificial Intelligence,
Department of Computer Science and Engineering,
Shanghai Jiao Tong University
Introduction

- Image-to-Video Retrieval:
  - retrieve relevant videos based on a query image

- Task Classification:
  - Instance-based Image-to-Video Retrieval (IIVR)
  - Activity-based Image-to-Video Retrieval (AIVR)
Introduction – IIVR & AIVR

- Existing IIVR research:
  - Early methods applied image retrieval methods for image-to-video retrieval,
  - Yu extracted object proposals from each frame and measured the similarity between the query object and the whole video through hamming distance,
  - Zhu and Wang introduced a large vocabulary quantization based Bag-of-Words to index videos.

Fig 1. Overview of Yu’s method

Fig 2. Overview of Zhu’s method
Introduction – IIVR & AIVR

- Limitations of IIVR:
  - focus more on object detection and representation while ignoring the dynamic motion tendency of different objects in videos and images

- Existing AIVR research:
  - APIVR projected the image features and activity proposal-based video features into a joint space and employed Graph Multi-Instance Learning module to filter out the noisy proposals.

Fig 1. Overview of APIVR
Introduction - Motivation

- Defects of APIVR: ignore the asymmetric relationship between images and videos.
- Image: appearance information; Video: appearance + motion information
  - Appearance information: the shape, pose, texture, and color of objects, …
  - Motion information: the trajectory of key points and variation of objects, …
Method - Overview

- Problem Definition:
  - Given an image and a video, return the similarity between them
- Two Modules: Feature Disentanglement and Video Feature Reconstruction

Fig 1. The flowchart of our method
Feature Extraction:

- Video Clip: 1) apply a R-C3D model pretrained on the ActivityNet dataset; 2) Choose top k proposals with largest confident scores; 3) Average;
- Image: apply the VGG-16 pretrained on ImageNet;
Method - Feature Disentanglement

- Asymmetric Disentanglement:
  - Video disentanglement:
    \[
    \mathbf{m}^v = E_v^{mo}(\mathbf{v}), \mathbf{a}^v = E_v^{ap}(\mathbf{v})
    \]
  - Image: \( \mathbf{u}^v = E_v^{ap}(\mathbf{u}) \)
  - Orthogonal Loss:
    \[
    \mathcal{L}_\text{orth} = \cos(\mathbf{m}^v, \mathbf{a}^v)
    \]
  - Classification loss:
    \[
    \mathcal{L}_\text{class} = -\log(p(\mathbf{a}^v)_y) - \log(p(\mathbf{a}^u)_y)
    \]
Since image-to-video translation is a multi-modal problem, inspired by VAE, we encode motion feature into motion uncertainty code $z$ ($p_\theta(z) = \mathcal{N}(0, 1)$).

- Apply two encoders: $\mu = E_\mu(m^v)$, $\sigma = E_\sigma(m^v)$

- Kullback–Leibler divergence Loss: $\mathcal{L}_{KL} = \text{KL}(q_\phi(z | m^v) || p_\theta(z))$
Method - Video Feature Reconstruction

- Generate motion uncertainty code: $\mathbf{z} = \mu + \epsilon\sigma$
- Video Feature Reconstruction: $\hat{\mathbf{v}} = D_v([\mathbf{a}^u, \mathbf{z}])$
- Reconstruction Loss: $\mathcal{L}_{re} = ||\bar{\mathbf{v}} - \hat{\mathbf{v}}||_2^2$
Method – Final Loss

- Final training loss: $\mathcal{L}_{total} = \lambda_0 \mathcal{L}_{orth} + \mathcal{L}_{class} + \mathcal{L}_{KL} + \mathcal{L}_{re}$
Method – Retrieval

- **Appearance Feature Space:**
  \[ S_A = 1 - \cos(a^u, a^v) \]

- **Video Feature Space:**
  - Sample motion uncertainty code from \( \mathcal{N}(0, 1) \) for \( h \) times;
  - Obtain \( h \) translated video features \( \{\hat{V}_i | i = 1 \ldots h\} \)
  - Calculate similarity: \( S_V = \min_{i=1}^{h} (1 - \cos(\bar{V}, \hat{V}_i)) \)

- **Combination:**
  \[ S_{all} = (1 - \lambda_v)S_A + \lambda_v S_V \]
  - \( \lambda_v \) is a hyper-parameter to balance two feature spaces
Follow the APIVR method to construct the dataset:
- Divide long videos and select video clips
- Randomly sample a frame as its paired image for each video clip

**ActivityNet:**
- 4727 validation videos from 200 activity categories
- Final obtain: 4739 image-video pairs = 3790 training pairs + 949 test pairs

**THUMOS’14:**
- 200 validation videos and 213 test videos from 20 different sports activities
- merge similar activity categories
- Final obtain: 7028 image-video pairs = 5614 training pairs + 1414 test pairs
## Experiment - Comparison with Other Methods

<table>
<thead>
<tr>
<th>Method</th>
<th>ActivityNet</th>
<th></th>
<th></th>
<th>THUMOS’14</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CMDN (Peng, Huang, and Qi 2016)</td>
<td>0.289</td>
<td>0.280</td>
<td>0.269</td>
<td>0.257</td>
<td>0.518</td>
<td>0.513</td>
</tr>
<tr>
<td>DSPE (Wang, Li, and Lazebnik 2016)</td>
<td>0.281</td>
<td>0.273</td>
<td>0.261</td>
<td>0.249</td>
<td>0.507</td>
<td>0.505</td>
</tr>
<tr>
<td>JFSSL (Wang et al. 2016)</td>
<td>0.277</td>
<td>0.268</td>
<td>0.256</td>
<td>0.244</td>
<td>0.476</td>
<td>0.473</td>
</tr>
<tr>
<td>ACMR (Wang et al. 2017)</td>
<td>0.294</td>
<td>0.288</td>
<td>0.273</td>
<td>0.259</td>
<td>0.526</td>
<td>0.522</td>
</tr>
<tr>
<td>CCL (Peng et al. 2018)</td>
<td>0.287</td>
<td>0.279</td>
<td>0.267</td>
<td>0.256</td>
<td>0.512</td>
<td>0.509</td>
</tr>
<tr>
<td>DSCMR (Zhen et al. 2019)</td>
<td>0.297</td>
<td>0.292</td>
<td>0.281</td>
<td>0.269</td>
<td>0.625</td>
<td>0.623</td>
</tr>
<tr>
<td>SDML (Hu et al. 2019)</td>
<td>0.304</td>
<td>0.301</td>
<td>0.289</td>
<td>0.279</td>
<td>0.648</td>
<td>0.647</td>
</tr>
<tr>
<td>BPBC (Xu et al. 2017)</td>
<td>0.295</td>
<td>0.287</td>
<td>0.275</td>
<td>0.258</td>
<td>0.514</td>
<td>0.511</td>
</tr>
<tr>
<td>APIVR (Xu et al. 2020)</td>
<td>0.308</td>
<td>0.298</td>
<td>0.283</td>
<td>0.269</td>
<td>0.655</td>
<td>0.653</td>
</tr>
<tr>
<td>MAP-IVR (Appearance)</td>
<td>0.304</td>
<td>0.297</td>
<td>0.284</td>
<td>0.273</td>
<td>0.643</td>
<td>0.641</td>
</tr>
<tr>
<td>MAP-IVR (Video)</td>
<td>0.323</td>
<td>0.313</td>
<td>0.296</td>
<td>0.282</td>
<td>0.691</td>
<td>0.689</td>
</tr>
<tr>
<td>MAP-IVR (Comb)</td>
<td>0.357</td>
<td>0.346</td>
<td>0.329</td>
<td>0.314</td>
<td>0.721</td>
<td>0.719</td>
</tr>
</tbody>
</table>

Table 1. Comparison with existing methods on ActivityNet and THUMOS’14.
Table 2. The ablation study of different loss terms. “Comb” represents the retrieval in the combination of two spaces; “Ap” and “Vi” represent the retrieval in appearance feature and video feature space, respectively. √ (resp., ×) means adding (resp., removing) this loss during training.
Table 2. The ablation study of different loss terms. “Comb” represents the retrieval in the combination of two spaces; “Ap” and “Vi” represent the retrieval in appearance feature and video feature space, respectively. √ (resp., ×) means adding (resp., removing) this loss during training.
Experiment - Ablation Study

Table 2. The ablation study of different loss terms. “Comb” represents the retrieval in the combination of two spaces; “Ap” and “Vi” represent the retrieval in appearance feature and video feature space, respectively. √ (resp., ×) means adding (resp., removing) this loss during training.

<table>
<thead>
<tr>
<th></th>
<th>$\mathcal{L}_{\text{class}}$</th>
<th>$\mathcal{L}_{\text{KL}}$</th>
<th>$\mathcal{L}_{\text{re}}$</th>
<th>$\mathcal{L}_{\text{orth}}$</th>
<th>Comb</th>
<th>Ap</th>
<th>Vi</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>0.357</td>
<td>0.304</td>
<td>0.323</td>
</tr>
<tr>
<td>2</td>
<td>×</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>0.296</td>
<td>—</td>
<td>0.296</td>
</tr>
<tr>
<td>3</td>
<td>√</td>
<td>×</td>
<td>√</td>
<td>√</td>
<td>0.221</td>
<td>0.297</td>
<td>0.047</td>
</tr>
<tr>
<td>4</td>
<td>√</td>
<td>√</td>
<td>×</td>
<td>√</td>
<td>0.299</td>
<td>0.299</td>
<td>—</td>
</tr>
<tr>
<td>5</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>×</td>
<td>0.334</td>
<td>0.303</td>
<td>0.307</td>
</tr>
<tr>
<td>6</td>
<td>√</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>0.285</td>
<td>0.285</td>
<td>—</td>
</tr>
</tbody>
</table>
### Table 2. Ablation Study

<table>
<thead>
<tr>
<th></th>
<th>( L_{\text{class}} )</th>
<th>( L_{KL} )</th>
<th>( L_{re} )</th>
<th>( L_{orth} )</th>
<th>Comb</th>
<th>Ap</th>
<th>Vi</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>0.357</td>
<td>0.304</td>
<td>0.323</td>
</tr>
<tr>
<td>2</td>
<td>×</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>0.296</td>
<td>—</td>
<td>0.296</td>
</tr>
<tr>
<td>3</td>
<td>√</td>
<td>×</td>
<td>√</td>
<td>√</td>
<td>0.221</td>
<td>0.297</td>
<td>0.047</td>
</tr>
<tr>
<td>4</td>
<td>√</td>
<td>√</td>
<td>×</td>
<td>√</td>
<td>0.299</td>
<td>0.299</td>
<td>—</td>
</tr>
<tr>
<td>5</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>×</td>
<td>0.334</td>
<td>0.303</td>
<td>0.307</td>
</tr>
<tr>
<td>6</td>
<td>√</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>0.285</td>
<td>0.285</td>
<td>—</td>
</tr>
</tbody>
</table>

Table 2. The ablation study of different loss terms. “Comb” represents the retrieval in the combination of two spaces; “Ap” and “Vi” represent the retrieval in appearance feature and video feature space, respectively. √ (resp., ×) means adding (resp., removing) this loss during training.
Experiment - Hyper-parameter Analysis

Analysis of different hyper-parameters

(a) The retrieval results with various combination ratios of two spaces.
(b) The retrieval results in video feature space and values of orthogonal loss with different $\lambda_o$.
(c) The retrieval results based on the combination of two spaces with different $h$ motion uncertain codes in the testing stage.
Experiment - Effect of Appearance and Motion Feature Dimension

Analysis of appearance and motion feature dimension.

(a) The retrieval results based on the combination of two spaces, where $\lambda_v = 0.5$.

(b) The retrieval results in appearance feature space and video feature space, respectively.
# Experiment - Visualization of Retrieved Videos

<table>
<thead>
<tr>
<th>Query Image</th>
<th>Appearance Feature Space</th>
<th>Retrieved Videos</th>
<th>Video Feature Space</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Query Image" /></td>
<td><img src="image2.png" alt="Appearance Feature Space" /></td>
<td><img src="image3.png" alt="Retrieved Videos" /></td>
<td><img src="image4.png" alt="Video Feature Space" /></td>
</tr>
</tbody>
</table>
Thanks!