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

Liu Liu, [1] **Jiangtong Li,** [1] **Li Niu***, [1] **Ruicong Xu,** [2] **Liqing Zhang** [1]
[1] MoE Key Lab of Artificial Intelligence, Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2] MEITUAN
{Shirlley, keep_moving-lee, ustcnewly}@sjtu.edu.cn, ranranxu95@gmail.com, zhang-lq@cs.sjtu.edu.cn

Abstract

With the rapid emergence of video data, image-to-video retrieval has attracted much attention. There are two types of image-to-video retrieval: instance-based and activity-based. The former task aims to retrieve videos containing the same main objects as the query image, while the latter focuses on finding the similar activity. Since dynamic information plays a significant role in the video, we pay attention to the latter task to explore the motion relation between images and videos. In this paper, we propose a Motion-assisted Activity Proposal-based Image-to-Video Retrieval (MAP-IVR) approach to disentangle the video features into motion features and appearance features and obtain appearance features from the images. Then, we perform image-to-video translation to improve the disentanglement quality. The retrieval is performed in both appearance and video feature spaces. Extensive experiments demonstrate that our MAP-IVR approach remarkably outperforms the state-of-the-art approaches on two benchmark activity-based video datasets.

Introduction

Since the rapid growth of multimedia data has brought great challenges to accurate retrieval across modalities, cross-modal retrieval has become a highlighted research topic in the retrieval area. Besides, along with the promotion of online video platforms like YouTube, video-related retrieval applications have drawn more and more attention from both academia and industry. One of the most basic applications is image-to-video retrieval, which aims to retrieve relevant videos based on a query image. In general, image-to-video retrieval can be categorized into Instance-based Image-to-Video Retrieval (IIVR) (Araujo and Girod 2018; Zhang et al. 2019) and Activity-based Image-to-Video Retrieval (AIVR) (Xu et al. 2020). IIVR aims to retrieve videos from the database based on the main instance (*i.e.*, main objects) in the query image, while AIVR aims to retrieve the videos containing a similar activity as the query image. Compared with IIVR, which mainly focuses on the static information of different instances in the video, AIVR pays attention to both static appearance and dynamic motion of these objects, which makes it much more challenging.

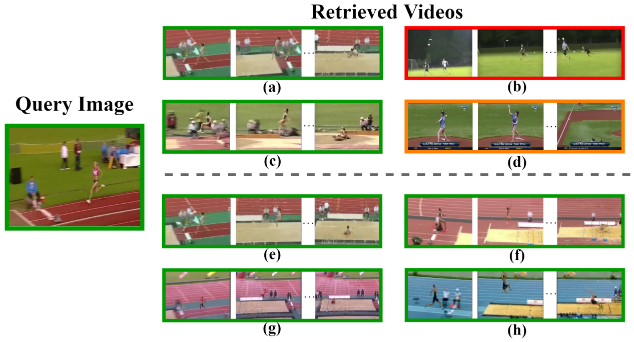

Figure 1: The example of the retrieved results of activity-based image-to-video retrieval with APIVR (Xu et al. 2020) (*resp.*, our expectation) (*i.e.*, a,b,c,d (*resp.*, e,f,g,h)), where different activities are outlined with different colors. In detail, (b) represents “Frisbee Catch”, (d) represents “Baseball Pitch”, while (a),(c),(e),(f),(g),(h), and the query image belong to “Long Jump”.

For instance-based image-to-video retrieval, a simple method (Sivic and Zisserman 2003) is to treat each frame in the video as an individual image and calculate the similarity between each frame and the query image. Nevertheless, the heavy computation makes it unsuitable for large-scale data, especially the long videos. To accelerate the retrieval process, Yu, Wang, and Yuan (2017) extracted object proposals from each frame and measured the similarity between the query object and the whole video through hamming distance. Additionally, to incorporate the object features from different views, Sivic, Schaffalitzky, and Zisserman (2004) learned object representations of query images and returned the objects of interest in video shots through object-level matching. However, all the instance-based image-to-video retrieval methods focus more on object detection and representation while ignoring the dynamic motion tendency of different objects in videos and images, which makes them difficult to discover the activity existing in the query image. Unlike IIVR, activity-based image-to-video retrieval not only searches for the same instances, but also explores the motion information in the videos. Considering this, in this paper, we focus on discovering the activity-related con-

*Corresponding author.

nection between images and videos.

To fulfil the Activity-based Image-to-Video Retrieval (AIVR) task, Xu et al. (2020) proposed Activity Proposal-based Image-to-Video Retrieval, which 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. The proposal-based method is capable of matching images and videos in a shared embedding space and avoiding the heavy computation of frame-based retrieval. However, simply matching images and videos in a shared embedding space is ill-suited for this task, which ignores the asymmetric relationship between images and videos. Specifically, the image features only contain appearance information (*e.g.*, the shape, pose, texture, and color of objects) while the video proposal features contain both appearance information and motion information (*e.g.*, the trajectory of key points and variation of objects). As shown in the top of Figure 1, we list top four retrieval results of Xu et al. (2020). Without considering the motion information in the videos, the retrieved results (*i.e.*, (b) and (d)) may belong to different activities, such as “Frisbee Catch” and “Baseball Pitch”, since these results share similar appearance features (*e.g.*, green background, people with running posture) with the query image.

Being aware of the asymmetric relationship between images and videos, we deem that the challenge of activity-based image-to-video retrieval is how to use the motion features from videos to assist the matching between images and videos. In this paper, we leverage the motion features disentangled from video features to guide the translation from image features to video features, which brings in additional motion information to facilitate the AIVR task. For example, in the bottom of Figure 1, we aim to explore the motion information inferred from image appearance (*e.g.*, the emergence of the jumping pit and the parabola of long jump).Therefore, the expected retrieval results (*i.e.*, (e), (f), (g), and (h)) could be decided from both static appearance and dynamic motion.

In the paper, we propose our Motion-assisted Activity Proposal-based Image-to-Video Retrieval (MAP-IVR) approach for the AIVR task, as illustrated in Figure 2. For each image, we first use a pre-trained model (*i.e.* VGG-16) to extract image features. For each video, we apply R-C3D (Xu, Das, and Saenko 2017), an extension of 3D CNN, to generate video proposal features, which are averaged to yield the video features. After that, the video features are disentangled into appearance features and motion features, while the image features are projected to the same appearance feature space. Then, we translate image appearance features to video features aided by motion uncertainty code. At length, considering the asymmetric relationship between videos and images, image-to-video translation from image appearance features have multiple possibilities (*e.g.*, shape variation and action details of different human body parts), which are characterized by the motion uncertainty code. Therefore, we integrate image appearance feature with the motion uncertainty code derived from certain video feature to reconstruct this video feature, in which the image and the video belong to the same activity category. Note that the motion uncertainty code not only compensates the motion uncertainty during image-to-video translation, but also facilitates motion-assisted image-to-video retrieval. Finally, we conduct retrieval in both appearance feature space and video feature space to combine the best of both worlds. Comprehensive experimental results on two benchmark activity-based video datasets verify the effectiveness of our method. Our contributions can be summarized as follows:

- Considering the asymmetric relation between image modality and video modality, we propose to facilitate the AIVR task with the assistance of video motion feature.
- We design a novel Motion-assisted Activity Proposal-based Image-to-Video Retrieval (MAP-IVR) approach to capture the activity-related correlation between images and videos.
- Experiment results on two benchmark activity-based video datasets demonstrate the superiority of our approach compared to state-of-the-art methods.

Related Work

Image-to-Video Retrieval

The goal of image-to-video retrieval is using a query image to retrieve relevant videos, which can be categorized into Instance-based Image-to-Video Retrieval (IIVR) (Araujo and Girod 2018; Zhang et al. 2019) and Activity-based Image-to-Video Retrieval (AIVR) (Xu et al. 2020). The IIVR task pays more attention to the appearance relationship between videos and images, so the retrieved videos contain the same main objects as the query image. A common solution is treating each video as a sequence of images, which inspires many works (Sivic and Zisserman 2003; Yu, Wang, and Yuan 2017; Xu et al. 2017) to apply image retrieval methods for image-to-video retrieval. Through the temporal consistency, Sivic, Schaffalitzky, and Zisserman (2004) could find different views of the same object in videos and return the objects of interest in video shots. To formulate the relevant video segments searching, Zhu and Satoh (2012); Wang et al. (2015) introduced a large vocabulary quantization based Bag-of-Words (Harris 1954) to index videos. Besides, Araujo and Girod (2018) developed video representation with Fisher Vectors and Bloom filters, aiming to find the visual information in videos. Additionally, Xu et al. (2017) measured the similarity through the distance between the query image and its orthogonal projection in the subspace spanned by video key frames.

Recently, considering the real-world applications, Xu et al. (2020) raised the Activity-based Image-to-Video Retrieval (AIVR) task and proposed an activity proposal-based approach. They adopted R-C3D (Xu, Das, and Saenko 2017) to generate temporal proposal features to preserve the activity information. Then, following ACMR (Wang et al. 2017), they projected the image features and activity proposal-based video features into a shared embedding space to measure similarities. Actually, the image features are not motion-aware, so it is hard to guarantee that the shared embedding space contains motion information. Therefore, in

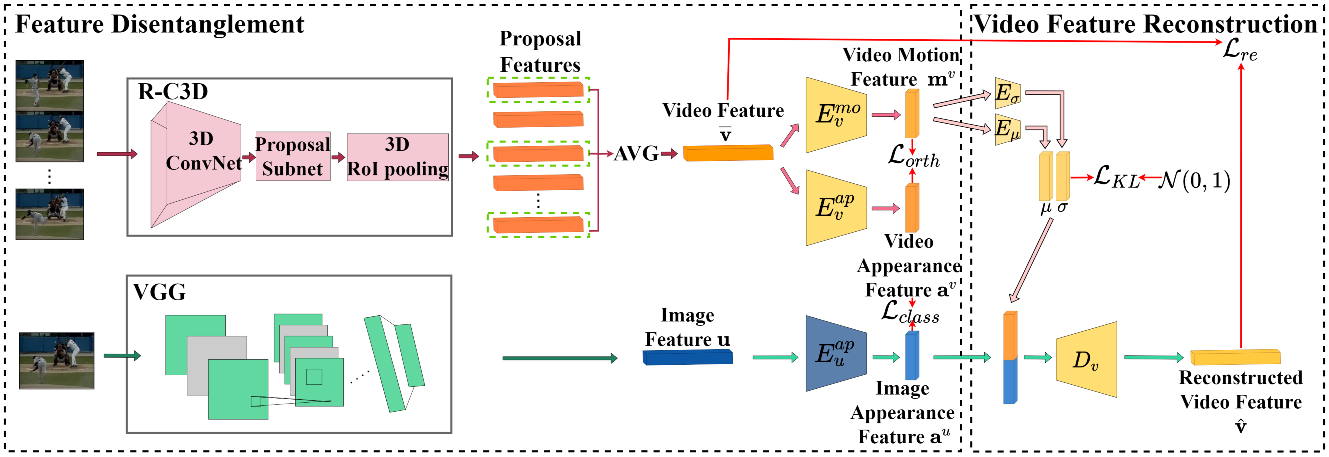

Figure 2: The flowchart of our MAP-IVR approach. We employ the R-C3D model (Xu, Das, and Saenko 2017) pretrained on the ActivityNet dataset and VGG-16 (Simonyan and Zisserman 2015) pretrained on ImageNet (Deng et al. 2009) to extract video proposal features and image features, respectively. Video proposal features are averaged to produce the video features. Then, we disentangle video features into motion features and appearance features, and project image features into the same appearance feature space. Finally, we perform image-to-video translation to reconstruct video features.

this paper, we employ video feature disentanglement and reconstruction to avoid losing motion information.

Video Representation Disentanglement

Since videos are essentially moving contents, it is natural to factor them into static and dynamic components. Lin et al. (2017) separated video into motion, foreground, and background under an unsupervised framework. Denton and Birodkar (2017) designed a predictive auto-encoder with adversarial loss to learn disentangled representation.

With the rapid growth of video data, learning disentangled video representation has benefited various areas such as video prediction (Villegas et al. 2017; Hsieh et al. 2018) and video generation (Tulyakov et al. 2018; Wang et al. 2020). Video prediction aims to predict and forecast what will happen in video sequences, and has been studied in several contexts such as activity prediction (Soran, Farhadi, and Shapiro 2015) and future frame prediction (Fan, Zhu, and Yang 2019). Under the help of disentanglement, Villegas et al. (2017) proposed to decompose the video into motion and content, which are encoded independently to predict the next frame. Video generation targets at generating realistic temporal dynamics, which can be divided into generation from additional input (Ohnishi et al. 2018) and noise (Saito, Matsumoto, and Saito 2017). Through video disentanglement, Tulyakov et al. (2018) adopted decomposed motion and content representation for video generation.

From the perspective of technical approach, similar to Villegas et al. (2017), we decompose the video into two complementary parts (*i.e.*, motion and appearance) with independent encoders. However, instead of processing frames of each video, we perform the disentanglement based on video-level features, which is simple yet effective. From the perspective of application, our work is the first one to introduce asymmetric representation disentanglement into the image-to-video retrieval task.

Methodology

Overall

In this section, we will elaborate the details of our proposed Motion-assisted Activity Proposal-based Image-to-Video Retrieval (MAP-IVR) approach for the AIVR task, as illustrated in Figure 2. Our proposed MAP-IVR can be divided into feature disentanglement module and video feature reconstruction module. In the first module, we disentangle the video feature into appearance feature and motion feature, and project the image feature into the same appearance feature space, which will be described in Section . In the second module, video feature reconstruction is performed based on the image appearance feature and the motion uncertainty code derived from motion feature, which will be described in Section . With two different feature spaces, our retrieval strategy will be discussed in Section .

Problem Formulation and Notation For concise mathematical expression, we denote a matrix (*e.g.*, $\mathbf{X}$) and vector (*e.g.*, $\mathbf{x}$) using an uppercase and lowercase letter in boldface, respectively, and denote a scalar (*e.g.*, x) using a lowercase letter. Besides, we adopt $[\cdot,\cdot]$ to represent the concatenation of two vectors, and $\cos(\cdot,\cdot)$ to denote the cosine similarity between two features.

In the training stage, we assume that there are n image-video pairs, denoted as $(\mathbf{u}_i, \bar{\mathbf{v}}_i)|_{i=1}^{n}$, where $\mathbf{u}_i \in \mathbb{R}^{d_u}$ is an image feature vector and $\bar{\mathbf{v}}_i \in \mathbb{R}^{d_v}$ is a video feature vector with d_u (*resp.*, d_v) being the feature dimension of $\mathbf{u}_i$ (*resp.*, $\bar{\mathbf{v}}_i$). In the meanwhile, each pair $(\mathbf{u}_i, \bar{\mathbf{v}}_i)$ has the same activity category label y_i. In the process of disentanglement, image appearance feature $\mathbf{a}_i^u \in \mathbb{R}^d$, video appearance feature $\mathbf{a}_i^v \in \mathbb{R}^d$, and video motion feature $\mathbf{m}_i^v \in \mathbb{R}^d$ can be learned, where d is the dimension of these features. Then, based on $\mathbf{a}_i^u$ and $\mathbf{m}_i^v$, we perform image-to-video translation to reconstruct video feature $\hat{\mathbf{v}}_i \in \mathbb{R}^{d_v}$. Our objective is to calculate the similarity between images and videos in appearance feature space (*i.e.*, $\mathbf{a}_i^u$, $\mathbf{a}_i^v$) and video feature space (*i.e.*, $\hat{\mathbf{v}}_i$, $\bar{\mathbf{v}}_i$). In each training iteration, we feed a pair $(\mathbf{u}_i, \bar{\mathbf{v}}_i)$ with the same activity category label into our method, and the subscript i is omitted in the remainder of this paper for simplicity.

Feature Disentanglement

Feature Extraction Given an image, we use VGG-16 (Simonyan and Zisserman 2015) pretrained on ImageNet (Deng et al. 2009) to extract its image feature $\mathbf{u}$.

Given a video clip, we first apply a R-C3D model (Xu, Das, and Saenko 2017) pretrained on the ActivityNet dataset (Heilbron et al. 2015) to extract a bag of temporal proposals which are very likely to contain the activity. Proposal generation is capable of generating candidate action proposals and filtering out background noise. Besides, for each proposal, the R-C3D model can predict its confidence scores corresponding to all activity categories. We use the largest confidence score among all activity categories as the confidence score for each proposal, and choose top k proposals with largest confident scores. For computation efficiency, we simply average the top k proposal features as the video feature $\bar{\mathbf{v}}$.

Asymmetric Disentanglement To perform the motion-assisted image-to-video retrieval, the first step is to disentangle motion and appearance features from the video features. Besides, the appearance features from both images and videos are aligned in a shared appearance feature space. Therefore, we use video motion encoder E_v^{mo} and video appearance encoder E_v^{ap} to disentangle the video feature $\bar{\mathbf{v}}$ into motion feature $\mathbf{m}^v$ and appearance feature $\mathbf{a}^v$ respectively. The image feature $\mathbf{u}$ is projected to image appearance feature $\mathbf{a}^u$ through another image appearance encoder E_u^{ap}. The above procedure can be formulated as

$$\mathbf{m}^v = E_v^{mo}(\bar{\mathbf{v}}),\ \mathbf{a}^v = E_v^{ap}(\bar{\mathbf{v}}),\ \mathbf{a}^u = E_u^{ap}(\mathbf{u}). \quad (1)$$

The goal of our feature disentangle module is separating motion and appearance information apart from the video feature. Therefore, to maximize the divergence between motion features and appearance features, we employ an orthogonal constraint (Shukla et al. 2019) to reinforce the disentanglement. Specifically, we adopt cosine similarity to measure the coherence between motion feature $\mathbf{m}^v$ and video appearance feature $\mathbf{a}^v$:

$$\mathcal{L}_{orth} = \cos(\mathbf{m}^v, \mathbf{a}^v), \quad (2)$$

where the results are non-negative because both $\mathbf{m}^v$ and $\mathbf{a}^v$ are the output of ReLU activation. With the cosine similarity decreasing, we expect the motion and appearance information to be disentangled from the video feature $\bar{\mathbf{v}}$. The motion feature concerns about dynamic changes, while the appearance feature focuses more on the static objects, which behaves similarly to the image feature.

When an image and a video belong to the same activity, they are supposed to contain visually similar objects. Therefore, we align images and videos in the shared appearance feature space. Specifically, in the appearance feature space, the images and videos belonging to the same activity category should be pulled close. Simultaneously, the appearance features of different activities should be pushed apart. Therefore, we employ an appearance classifier p to distinguish the features from different activity categories regardless of their modalities (image or video), with the cross-entropy classification loss:

$$\mathcal{L}_{class} = -\log(p(\mathbf{a}^v)_y) - \log(p(\mathbf{a}^u)_y), \quad (3)$$

in which y is the activity category label shared by $\mathbf{a}^u$ and $\mathbf{a}^v$. $p(\cdot)_j$ means the classification score corresponding to the j-th category.

Video Feature Reconstruction

To ensure that the video feature is successfully disentangled into motion feature and appearance feature, we perform video feature reconstruction on the basis of image appearance feature $\mathbf{a}^u$ and video motion feature $\mathbf{m}^v$.

Because images lack motion information, image-to-video translation is a multi-modal problem because the translated video could have multiple possibilities instead of a single deterministic result. For example, given an image with a man playing basketball, even if we can tell this man is going to shoot the ball, the shape variation and action details of different human body parts in the shooting process are still uncertain, leading to multiple possible motion information compatible with this image.

Inspired by Kingma and Welling (2014), we encode motion feature into motion uncertainty code $\mathbf{z}$, which compensates for the motion uncertainty when translating images to their corresponding videos of the same activity. To support stochastic sampling in the testing stage, we enforce the motion uncertainty code to follow unit Gaussian distribution, where $p_\theta(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{1})$. Specifically, we first apply motion uncertainty encoders E_μ and E_σ on motion feature $\mathbf{m}^v$ to obtain $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ respectively, which form the conditional probability $q_\phi(\mathbf{z}|\mathbf{m}^v) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$. Then, we employ the Kullback–Leibler divergence (Kullback and Leibler 1951) loss to encourage $q_\phi(\mathbf{z}|\mathbf{m}^v)$ to be close to the prior $p_\theta(\mathbf{z})$:

$$\mathcal{L}_{KL} = KL(q_\phi(\mathbf{z}|\mathbf{m}^v)||p_\theta(\mathbf{z})), \quad (4)$$

where ϕ indicates the model parameters of motion uncertainty encoders E_μ and E_σ.

In the training stage, with $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$, we adopt the common reparameterization trick (Kingma and Welling 2014) to generate $\mathbf{z}$ as $\mathbf{z} = \boldsymbol{\mu} + \epsilon\boldsymbol{\sigma}$, where ϵ is a random value sampled from $\mathcal{N}(0, 1)$. The concatenation of image appearance feature $\mathbf{a}^u$ and motion uncertainty code $\mathbf{z}$ is sent to the video feature decoder D_v, yielding the reconstructed video feature $\hat{\mathbf{v}}$:

$$\hat{\mathbf{v}} = D_v([\mathbf{a}^u, \mathbf{z}]). \quad (5)$$

Recall that $\bar{\mathbf{v}}$ is decomposed into $\mathbf{a}^v$ and $\mathbf{m}^v$. Since $\mathbf{a}^u$ is from the same category as $\mathbf{a}^v$ and the motion uncertainty code $\mathbf{z}$ is provided by $\mathbf{m}^v$, the reconstructed video feature $\hat{\mathbf{v}}$ (see Eqn. (5)) is expected to approach $\bar{\mathbf{v}}$. Therefore, we apply a L_2 reconstruction loss to enforce $\bar{\mathbf{v}}$ and $\hat{\mathbf{v}}$ to be close, which is formulated as:

$$\mathcal{L}_{re} = \|\bar{\mathbf{v}} - \hat{\mathbf{v}}\|_2^2. \quad (6)$$

Method	ActivityNet				THUMOS'14			
	mAP@10	mAP@20	mAP@50	mAP@100	mAP@10	mAP@20	mAP@50	mAP@100
CMDN (Peng, Huang, and Qi 2016)	0.289	0.280	0.269	0.257	0.518	0.513	0.508	0.504
DSPE (Wang, Li, and Lazebnik 2016)	0.281	0.273	0.261	0.249	0.507	0.505	0.501	0.498
JFSSL (Wang et al. 2016)	0.277	0.268	0.256	0.244	0.476	0.473	0.469	0.465
ACMR (Wang et al. 2017)	0.294	0.288	0.273	0.259	0.526	0.522	0.514	0.505
CCL (Peng et al. 2018)	0.287	0.279	0.267	0.256	0.512	0.509	0.506	0.502
DSCMR (Zhen et al. 2019)	0.297	0.292	0.281	0.269	0.625	0.623	0.622	0.621
SDML (Hu et al. 2019)	0.304	0.301	0.289	0.279	0.648	0.647	0.646	0.645
BPBC (Xu et al. 2017)	0.295	0.287	0.275	0.258	0.514	0.511	0.507	0.503
APIVR (Xu et al. 2020)	0.308	0.298	0.283	0.269	0.655	0.653	0.651	0.649
MAP-IVR (Appearance)	0.304	0.297	0.284	0.273	0.643	0.641	0.637	0.635
MAP-IVR (Video)	0.323	0.313	0.296	0.282	0.691	0.689	0.682	0.677
MAP-IVR (Comb)	**0.357**	**0.346**	**0.329**	**0.314**	**0.721**	**0.719**	**0.717**	**0.714**

Table 1: Comparison with existing methods on ActivityNet and THUMOS'14. Best results are denoted in boldface.

Finally, we collect the orthogonal loss $\mathcal{L}_{orth}$, classification loss $\mathcal{L}_{class}$, KL divergence loss $\mathcal{L}_{KL}$, and reconstruction loss $\mathcal{L}_{re}$ as our total training loss:

$$\mathcal{L}_{total} = \lambda_o \mathcal{L}_{orth} + \mathcal{L}_{class} + \mathcal{L}_{KL} + \mathcal{L}_{re}. \tag{7}$$

Retrieval

During testing, the comparison between images and videos can be performed in both appearance feature space and video feature space with our model.

Appearance Feature Space Given an image with feature $\mathbf{u}$ and a video with feature $\bar{\mathbf{v}}$, we apply appearance encoders E_u^{ap} and E_v^{ap} to obtain image appearance feature $\mathbf{a}^u$ and video appearance feature $\mathbf{a}^v$, respectively. Then, we calculate their distance by

$$S_A = 1 - \cos(\mathbf{a}^u, \mathbf{a}^v). \tag{8}$$

Video Feature Space Given an image with feature $\mathbf{u}$, after obtaining its appearance feature $\mathbf{a}^u$, we sample motion uncertainty code from $\mathcal{N}(\mathbf{0}, \mathbf{1})$ for h times, and concatenate each motion uncertainty code with $\mathbf{a}^u$, leading to h translated video features $\{\hat{\mathbf{v}}_i | i = 1 \ldots h\}$ containing different motion information. For the comparison between this image and a video with feature $\bar{\mathbf{v}}$, we need to find out the translated video feature $\hat{\mathbf{v}}_i$ which is closest to $\bar{\mathbf{v}}$. Formally, we calculate the distance as

$$S_V = \min_{i=1}^{h}(1 - \cos(\bar{\mathbf{v}}, \hat{\mathbf{v}}_i)). \tag{9}$$

To take full advantage of two spaces, we perform retrieval based on the weighted average of S_A and S_V:

$$S_{all} = (1 - \lambda_v) S_A + \lambda_v S_V, \tag{10}$$

where λ_v is a hyper-parameter to balance two feature spaces.

Experiment

In this section, we will introduce the datasets, implementation details, and evaluation metrics in Section . Then, we will compare our model with state-of-the-art methods in Section . To verify the effectiveness of our method, we will provide extensive ablation studies in Section . Besides, we will investigate hyper-parameters in our model in Section . Finally, in Section , we will visualize and analyse the retrieved results in two spaces.

Experiment Setup

Dataset Since previous AIVR method (Xu et al. 2020) does not release its used datasets, we construct the datasets in a similar way to them, based on two public activity video datasets THUMOS'14 (Jiang et al. 2014) [1] and ActivityNet (Heilbron et al. 2015) [2].

For THUMOS'14, we use 200 validation videos and 213 test videos from 20 different sports activities, because validation and test videos contain temporal annotations of actions. We merge similar activity categories: "Cricket Bowling" and "Cricket Shot", "Cliff Diving" and "Diving", resulting in 18 remaining activity categories. For ActivityNet, due to the limit of GPU memory and speed, we only use the validation videos. As some video links have expired, we eventually obtain 4727 videos from 200 activity categories. In order to ensure that each video only belongs to one activity category, we first divide each long video into multiple short videos according to the temporal annotation. Then, we sample a fixed number of consecutive key frames from each short video as a video clip. Following Xu et al. (2020), we set the number of key frames in each video clip as 768 for all datasets, which is large enough to cover at least one activity instance. To construct image-video pairs, for each video clip, we randomly sample a frame as its paired image.

Based on the collected image-video pairs, we can split each dataset into training and test set. For THUMOS'14, we obtain 7028 image-video pairs in total, and then divide them into 5614 training pairs and 1414 test pairs, where the test pairs exclude the validation videos because the validation set is used for finetuning R-C3D model. For ActivityNet, we obtain 4739 image-video pairs in total, which are divided into 3790 training pairs and 949 test pairs.

Implementation Details For image features, we employ VGG-16 (Simonyan and Zisserman 2015) pretrained on ImageNet (Deng et al. 2009) to extract the output from fc7 layer as the image features. For video clips in ActivityNet dataset, we apply R-C3D (Xu, Das, and Saenko 2017) model pretrained on the ActivityNet dataset (Heilbron et al. 2015) to extract activity proposals. To obtain better video features

[1] https://www.crcv.ucf.edu/THUMOS14/home.html
[2] http://activity-net.org/

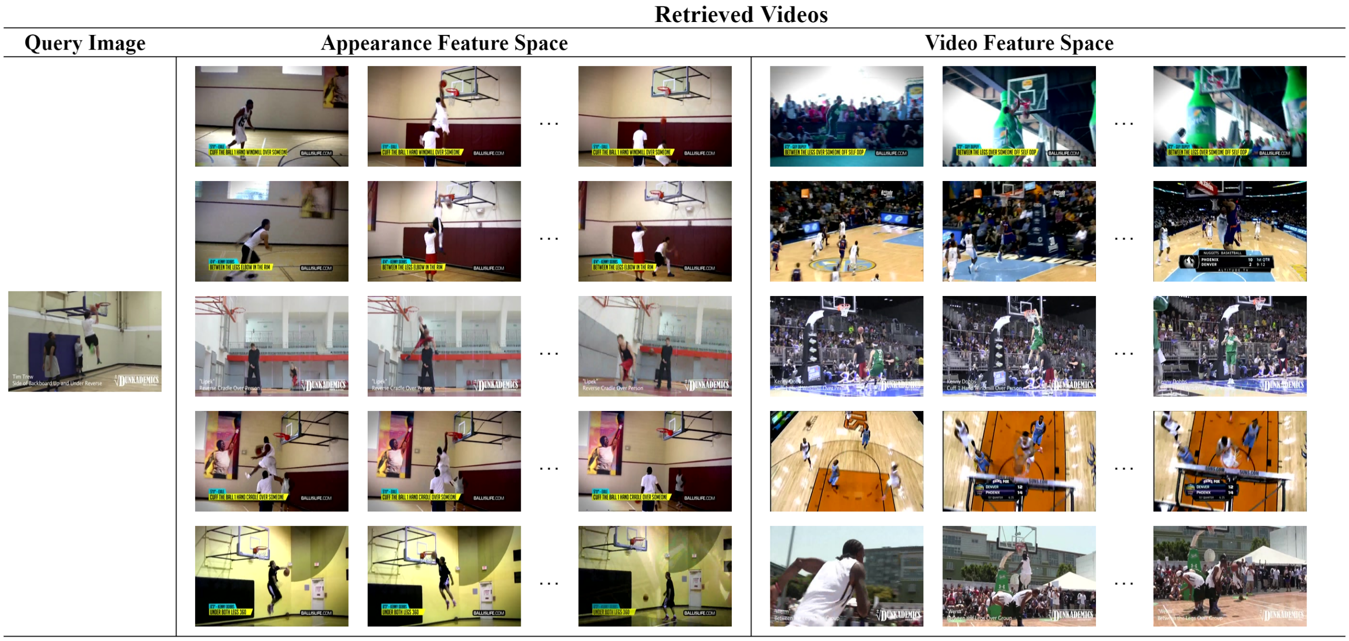

Figure 3: Visualization of the retrieved videos by our MAP-IVR approach. With a query image, we show the top 5 retrieved videos in each space (*i.e.*, appearance feature space and video feature space).

on THUMOS'14 dataset (Jiang et al. 2014), we finetune the pretrained R-C3D model with the validation videos of THUMOS'14, including 180 training videos and 20 held-out videos, to extract activity proposals of each video clip in THUMOS'14, which is different from Xu et al. (2020).

For all encoders and decoders used in our model, we employ three fully-connected layers with Batch Normalization and ReLU activations. The dimensionality of image feature and video feature is 4096. The dimensionality of appearance feature, motion feature, and motion uncertainty code is 1024. During training, we choose Adam (Kingma and Ba 2015) with learning rate 1×10^{-4} and set batch size as 32 for 60 epochs. Additionally, we set λ_o as 1. While retrieving, we set λ_v as 0.5. Besides, we sample 25 motion uncertainty codes $\mathbf{z}$ for the retrieval in video feature space (*i.e.*, $h = 25$). All the hyper-parameters are set via cross-validation. Our model is implemented by PyTorch1.4 (Paszke et al. 2019) on Ubuntu 16.04 and trained on a single GTX 1080Ti GPU. We set the random seed as 123. The significant test with different seeds will be described in Supplementary Material.

Evaluation Metrics For a fair comparison, we adopt the same evaluation metrics as Xu et al. (2020). In detail, we use mAP@K, *i.e.*, mean Aversion Precision based on top K retrieved results. In our paper, we report mAP@10, mAP@20, mAP@50, and mAP@100 on both datasets.

Comparison with Existing Methods

Since there are only a few methods especially targeting at image-to-video retrieval, we compare our proposed MAP-IVR approach with two types of state-of-the-art methods. One is general cross-modal retrieval models, including CMDN (Peng, Huang, and Qi 2016), DSPE (Wang, Li, and Lazebnik 2016), JFSSL (Wang et al. 2016), ACMR (Wang et al. 2017), CCL (Peng et al. 2018), DSCMR (Zhen et al. 2019), SDML (Hu et al. 2019). The other one is image-to-video retrieval methods: BPBC (Xu et al. 2017) for the Instance-based Image-to-Video Retrieval (IIVR) task and APIVR (Xu et al. 2020) for the Activity-based Image-to-Video Retrieval (AIVR) task. Note that although Araujo and Girod (2018) solved the IIVR task, its primary purpose is to improve video Fisher Vectors with bloom filters, which is unsuitable for our task. For each baseline method, we vary the dimension of feature used for retrieval in the range of $[64, 4096]$ and report the best results in Table 1. For our method, we report the retrieval results in appearance feature space, video feature space, and the combination of both.

We show our experimental results in Table 1. For both datasets, our method outperforms all the state-of-the-art methods by a large margin. Compared with APIVR (Xu et al. 2020), which focuses on the same task as ours, our method achieves an improvement of 4.9% on mAP@10 in ActivityNet, and 6.6% on mAP@10 in THUMOS'14 [3]. Besides, by comparing appearance feature space and video feature space for our method, the advantage of video feature space reflects the benefit of incorporating additional motion information. Moreover, combining two spaces can further boost the performance.

Ablation Study

In order to study the effectiveness of different components, we ablate different loss terms and observe the performance

[3] Note we fine-tune the R-C3D on THUMOS'14, which makes the results of our reproduced APIVR (Xu et al. 2020) much higher than their reported results.

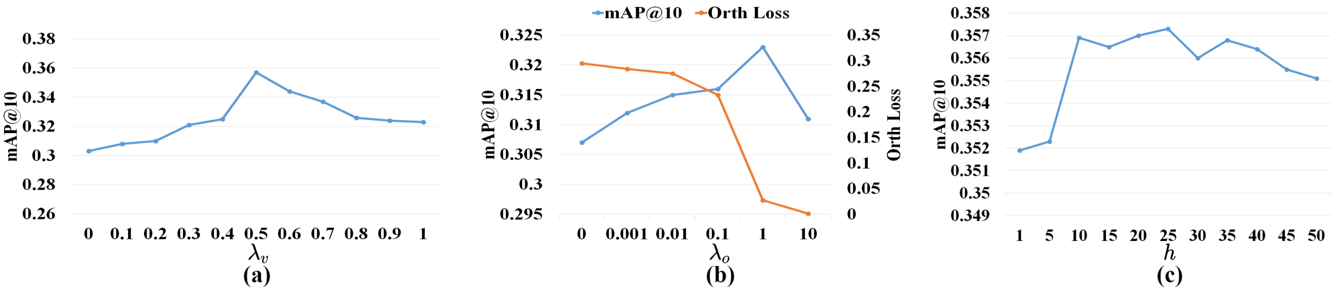

Figure 4: Analysis of different hyper-parameters. (a) The retrieval results with various combination ratios of two spaces. (b) The retrieval results in video feature space and values of orthogonal loss with different λ_o. (c) The retrieval results based on the combination of two spaces with different h motion uncertain codes in the testing stage.

	$\mathcal{L}_{class}$	$\mathcal{L}_{KL}$	$\mathcal{L}_{re}$	$\mathcal{L}_{orth}$	Comb	Ap	Vi
1	√	√	√	√	0.357	0.304	0.323
2	×	√	√	√	0.296	—	0.296
3	√	×	√	√	0.221	0.297	0.047
4	√	√	×	√	0.299	0.299	—
5	√	√	√	×	0.334	0.303	0.307
6	√	×	×	×	0.285	0.285	—

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

variance. All experiments are conducted on the ActivityNet dataset using the evaluation metric mAP@10, and results are reported in Table 2. After removing $\mathcal{L}_{class}$, there is an evident drop in the video feature space (“Vi” in row 1 *v.s.* “Vi” in row 2), which indicates the importance of aligning features in the appearance feature space. However, when only using $\mathcal{L}_{class}$, the results also decrease (“Comb” in row 1 *v.s.* “Comb” in row 6), implying that a single classification loss cannot guarantee the quality of feature disentanglement. The results become worse without using $\mathcal{L}_{re}$ (“Comb” in row 1 *v.s.* “Comb” in row 4), which verifies that video feature reconstruction module can help the feature disentanglement.

Hyper-parameter Analysis

By taking the ActivityNet dataset as an example, we analyse the hyper-parameters (*i.e.*, λ_v, λ_o, and h) used in our method, with the evaluation metric mAP@10. To explore the effectiveness of different spaces, we vary λ_v in the range of $[0, 1]$ and the results based on the combination of two spaces are shown in Figure 4 (a). When $\lambda_v = 0.5$, our model reaches the best results, which implies that the combination of two spaces has a positive impact on the retrieval performance. In order to learn the influence of the orthogonal loss $\mathcal{L}_{orth}$ (see in Eqn. 2) on feature disentanglement, we plot $\mathcal{L}_{orth}$ and the retrieval performance in video feature space by varying λ_o, which is shown in Figure 4 (b). As λ_o increases in a reasonable range $[0.001, 1]$, the final $\mathcal{L}_{orth}$ drops but mAP@10 arises, which proves that better disentanglement leads to better retrieval performance. Finally, we experiment with different h, *i.e.*, the number of sampled motion uncertain codes used in the testing stage. From Figure 4 (c), when $\lambda_v = 0.5$, we can find that the model can achieve satisfactory results based on the combination of two spaces when h is reasonably large (*e.g.*, $h > 10$). With h getting too large, some outlier points may be sampled and harm the retrieval performance in video feature space. We also take a study on the dimension of motion feature and appearance feature (*i.e.*, d) in Supplementary Material.

Visualization of Retrieved Videos

To better demonstrate the effectiveness of our method, we provide a retrieval example on THUMOS’14 in Figure 3. More examples will be shown in Supplementary Material. With a query image, we show top 5 retrieved videos in appearance feature space and video feature space. We conjecture that the former cares more about static appearance information (*e.g.*, main object and background) while the latter additionally focuses on dynamic motion information (*e.g.*, the trajectory of key points and variation of objects).

As shown in Figure 3, with a query image from category “Basketball Dunk”, the retrieved videos in appearance feature space contain the main objects such as people, basketball, and backboard. Besides, the background also resembles the query image. In video feature space, the retrieved videos share the same activity, but vary in shooting angle and background scene. According to this example, we can observe that results in video feature space focus more on the activity with the help of the motion information. Therefore, it is evident that our method can separate the motion and appearance information successfully. Furthermore, since the results in appearance feature space and video feature space are complementary, the combination of both can produce better results (see Table 1).

Conclusion

In this paper, we have studied the Activity-based Image-to-Video Retrieval (AIVR) from a new viewpoint. Specifically, we have proposed our MAP-IVR approach, which utilizes motion information to facilitate the AIVR task via video feature disentanglement and reconstruction. Comprehensive experiments on two benchmark datasets have demonstrated the effectiveness of our MAP-IVR approach.

Acknowledgments

The work is supported by the National Key R&D Program of China (2018AAA0100704) and is partially sponsored by National Natural Science Foundation of China (Grant No.61902247) and the Shanghai Science and Technology RD Program of China (20511100300). This work is also sponsored by Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102).

References

Araujo, A.; and Girod, B. 2018. Large-Scale Video Retrieval Using Image Queries. *IEEE Transactions on Circuits and Systems for Video Technology* 28(6): 1406–1420.

Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; and Li, F. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 248–255.

Denton, E. L.; and Birodkar, V. 2017. Unsupervised Learning of Disentangled Representations from Video. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 4414–4423.

Fan, H.; Zhu, L.; and Yang, Y. 2019. Cubic LSTMs for Video Prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 8263–8270.

Harris, Z. 1954. Distributional structure. *Word* 10(23): 146–162.

Heilbron, F. C.; Escorcia, V.; Ghanem, B.; and Niebles, J. C. 2015. ActivityNet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 961–970.

Hsieh, J.; Liu, B.; Huang, D.; Li, F.; and Niebles, J. C. 2018. Learning to Decompose and Disentangle Representations for Video Prediction. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 515–524.

Hu, P.; Zhen, L.; Peng, D.; and Liu, P. 2019. Scalable Deep Multimodal Learning for Cross-Modal Retrieval. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 635–644.

Jiang, Y.-G.; Liu, J.; Roshan Zamir, A.; Toderici, G.; Laptev, I.; Shah, M.; and Sukthankar, R. 2014. THUMOS Challenge: Action Recognition with a Large Number of Classes.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Kullback, S.; and Leibler, R. A. 1951. On Information and Sufficiency. *Ann. Math. Statist.* 22(1): 79–86. doi:10.1214/aoms/1177729694.

Lin, X.; Campos, V.; Giro-i Nieto, X.; Torres, J.; and Ferrer, C. C. 2017. Disentangling Motion, Foreground and Background Features in Videos. In *CVPR 2017 Workshop: Brave new ideas for motion representations in videos II.*

Ohnishi, K.; Yamamoto, S.; Ushiku, Y.; and Harada, T. 2018. Hierarchical Video Generation From Orthogonal Information: Optical Flow and Texture. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2387–2394.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 8026–8037.

Peng, Y.; Huang, X.; and Qi, J. 2016. Cross-Media Shared Representation by Hierarchical Learning with Multiple Deep Networks. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 3846–3853.

Peng, Y.; Qi, J.; Huang, X.; and Yuan, Y. 2018. CCL: Cross-modal Correlation Learning With Multigrained Fusion by Hierarchical Network. *IEEE Transactions on Multimedia* 20(2): 405–420.

Saito, M.; Matsumoto, E.; and Saito, S. 2017. Temporal Generative Adversarial Nets with Singular Value Clipping. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2849–2858.

Shukla, A.; Bhagat, S.; Uppal, S.; Anand, S.; and Turaga, P. K. 2019. PrOSe: Product of Orthogonal Spheres Parameterization for Disentangled Representation Learning. In *Proceedings of the British Machine Vision Conference (BMVC)*, 88.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Sivic, J.; Schaffalitzky, F.; and Zisserman, A. 2004. Object Level Grouping for Video Shots. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 85–98.

Sivic, J.; and Zisserman, A. 2003. Video Google: A Text Retrieval Approach to Object Matching in Videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 1470–1477.

Soran, B.; Farhadi, A.; and Shapiro, L. G. 2015. Generating Notifications for Missing Actions: Don't Forget to Turn the Lights Off! In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 4669–4677.

Tulyakov, S.; Liu, M.; Yang, X.; and Kautz, J. 2018. MoCoGAN: Decomposing Motion and Content for Video Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1526–1535.

Villegas, R.; Yang, J.; Hong, S.; Lin, X.; and Lee, H. 2017. Decomposing Motion and Content for Natural Video Sequence Prediction. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Wang, B.; Yang, Y.; Xu, X.; Hanjalic, A.; and Shen, H. T. 2017. Adversarial Cross-Modal Retrieval. In *Proceedings of the ACM on Multimedia Conference (ACM MM)*, 154–162.

Wang, K.; He, R.; Wang, L.; Wang, W.; and Tan, T. 2016. Joint Feature Selection and Subspace Learning for Cross-Modal Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38(10): 2010–2023.

Wang, L.; Li, Y.; and Lazebnik, S. 2016. Learning Deep Structure-Preserving Image-Text Embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5005–5013.

Wang, Y.; Bilinski, P.; Brémond, F.; and Dantcheva, A. 2020. G3AN: Disentangling Appearance and Motion for Video Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5263–5272.

Wang, Y.; Lin, X.; Wu, L.; Zhang, W.; Zhang, Q.; and Huang, X. 2015. Robust Subspace Clustering for Multi-View Data by Exploiting Correlation Consensus. *IEEE Trans. Image Process.* 24(11): 3939–3949.

Xu, H.; Das, A.; and Saenko, K. 2017. R-C3D: Region Convolutional 3D Network for Temporal Activity Detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 5794–5803.

Xu, R.; Niu, L.; Zhang, J.; and Zhang, L. 2020. A Proposal-Based Approach for Activity Image-to-Video Retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 12524–12531.

Xu, R.; Yang, Y.; Shen, F.; Xie, N.; and Shen, H. T. 2017. Efficient Binary Coding for Subspace-based Query-by-Image Video Retrieval. In *Proceedings of the ACM on Multimedia Conference (ACM MM)*, 1354–1362.

Yu, T.; Wang, Z.; and Yuan, J. 2017. Compressive Quantization for Fast Object Instance Search in Videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 726–735.

Zhang, C.; Lin, Y.; Zhu, L.; Liu, A.; Zhang, Z.; and Huang, F. 2019. CNN-VWII: An efficient approach for large-scale video retrieval by image queries. *Pattern Recognit. Lett.* 123: 82–88.

Zhen, L.; Hu, P.; Wang, X.; and Peng, D. 2019. Deep Supervised Cross-Modal Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10394–10403.

Zhu, C.; and Satoh, S. 2012. Large vocabulary quantization for searching instances from videos. In *Proceedings of the International Conference on Multimedia Retrieval (ICMR)*, 52.