Person Tube Retrieval via Language Description

Hehe Fan,[1,2*] **Yi Yang** [2]
[1]Baidu Research
[2]ReLER, CAI, University of Technology Sydney
hehe.fan@student.uts.edu.au, yi.yang@uts.edu.au

Abstract

This paper focuses on the problem of person tube (a sequence of bounding boxes which encloses a person in a video) retrieval using a natural language query. Different from images in person re-identification (re-ID) or person search, besides appearance, person tube contains abundant action and information. We exploit a 2D and a 3D residual networks (ResNets) to extract the appearance and action representation, respectively. To transform tubes and descriptions into a shared latent space where data from the two different modalities can be compared directly, we propose a Multi-Scale Structure Preservation (MSSP) approach. MSSP splits a person tube into several element-tubes on average, whose features are extracted by the two ResNets. Any number of consecutive element-tubes forms a sub-tube. MSSP considers the following constraints for sub-tubes and descriptions in the shared space. 1) Bidirectional ranking. Matching sub-tubes (resp. descriptions) should get ranked higher than incorrect ones for each description (resp. sub-tube). 2) External structure preservation. Sub-tubes (resp. descriptions) from different persons should stay away from each other. 3) Internal structure preservation. Sub-tubes (resp. descriptions) from the same person should be close to each other. Experimental results on person tube retrieval via language description and other two related tasks demonstrate the efficacy of MSSP.

Introduction

Person retrieval in large-scale video databases has important applications in activity analysis and video surveillance. One of the image-based person retrieval tasks is known as person re-identification (re-ID) in computer vision (Zeng et al. 2018; Fan et al. 2018; Yang, Wang, and Tao 2018; Sun and Zheng 2019; Ding et al. 2019; Yang et al. 2017; Miao et al. 2019; Quan et al. 2019), which aims at spotting a person of interest in other cameras by a query image. However, these methods require an example image of the person of interest, which is a limitation in practice. In terms of usability, it would be more desirable if the person could be searched using more easily available natural language descriptions.

The task of person search with natural language description (Li et al. 2017b) then replaces image queries with natural language descriptions for re-ID. Given the textual description of a person, this task aims at ranking all the images in the gallery and then retrieving the most relevant image corresponding to the queried description. As the work of (Li et al. 2017b) focuses on re-ID, images are well cropped for all persons and descriptions are only about person appearances. The temporal and scene information is missing in both visual modality and textual modality.

Person tube retrieval via natural language queries is a more general and practical task for person retrieval. A person tube is a sequence of bounding boxes which encloses a person in video. We illustrate a few of examples of person tube retrieval via language description in Fig. 1. Compared with re-ID, these persons are not limited to pedestrians. They can make a variety of actions in different scenes. Corresponding to these various person tubes, each query contains abundant information including appearance, action and the scene (Qi et al. 2019) around the person. Different from spatio-temporal person retrieval in videos (Yamaguchi et al. 2017), which includes person detection and tracking, person tube retrieval is consistent with re-ID and only focuses on the retrieval part. People in videos are well cropped and tracked, forming the so-called person tube. We respectively exploit a 2D residual networks (ResNet) (He et al. 2016) (pretrained on the large-scale image dataset ImageNet (Russakovsky et al. 2015)) to extract the appearance and scene representation, and a 3D ResNet (Hara, Kataoka, and Satoh 2018) (pretrained on the large-scale action dataset Kinetics (Carreira and Zisserman 2017)) to extract the action representation.

To learn a joint embedding for person tube representations and description representations into a shared latent space, we propose a Multi-Scale Structure Preservation (MSSP) approach. MSSP splits a person tube into several element-tubes on average, *e.g.*, 16 frames per element-tube. We use the two ResNets to extract the representation for each element-tube. Any number of consecutive element-tubes forms a sub-tube. The person tube can be seen as a visual group, which consists of several different scale sub-tubes. Similarly, the descriptions about the same person are

*This work was done when Hehe Fan was an intern at Baidu Research.

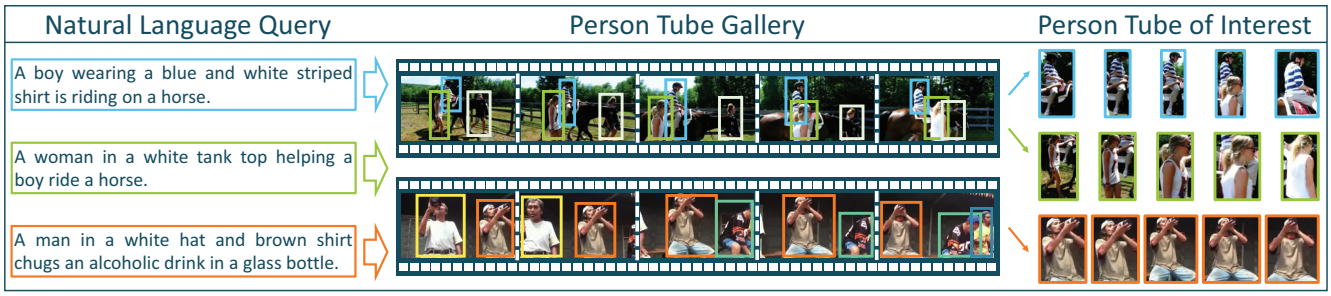


Figure 1: Examples of person tube retrieval via natural language queries. A person tube is a sequence of bounding boxes which encloses a person in a video. Given a language query, the system aims to retrieve through the person tube database (gallery) and outputs the target person tube.

considered as a textual group. In order to make the representations from the vision modality and the language modality can be compared directly, MSSP tries to optimize the following constraints.

- Inter-modality: bidirectional ranking constraint. This constraint aims at encouraging correct sub-tubes (resp. descriptions) to get ranked higher than incorrect ones for each training description (resp. sub-tube).
- Intra-modality: external-internal structure-preserving constraint. This constraint preserves neighborhood structure within visual or textual modality. Specifically, in the learned latent space, sub-tubes (resp. descriptions) belong to the same group should be close to each other, which is referred to as the internal structure preservation. Sub-tubes (resp. descriptions) from different groups should stay away from each other, which is referred to as the external structure preservation.

We evaluate our method on an ActivityNet-PTRLD datatset, demonstrating the efficacy of MSSP on person tube retrieval via language description. By slightly modifying MSSP, it can also be applied to image-based person retrieval and video retrieval, with language description. We evaluate variant MSSPs on the CUHK-PEDES datatset (Li et al. 2017b) and the MSR-VTT datatset (Xu et al. 2016). Experimental results show that MSSP also improves the accuracy on these two tasks.

Related Work

Person retrieval. To search possible criminal suspects from surveillance videos or person of interest from videos on the Internet is in urgent need. Given a query, person retrieval addresses the problem of searching the target person from image or video databases. Usually, person retrieval can be divided into image-query-based person-identification (re-ID) (Wang et al. 2016; Xiao et al. 2017; Zheng et al. 2015; Song et al. 2018; Chen et al. 2018b) and language-query-based person retrieval (Li et al. 2017b; Yamaguchi et al. 2017). As the image-query-based re-ID requires at least one photo of the target person, it has a limitation in practice. By contrast, language-query-based person retrieval is more flexible because free-form natural language description can be easily obtained. However, for the image-based person retrieval (Li et al. 2017b), the language query is limited to appearance, which may not sufficiently describe a person sometimes. Recently, video-based re-ID (Zheng et al. 2017a; Chen et al. 2018b) also attracts the community's attention. However, on one hand, as re-ID focuses on pedestrian persons, the videos are also about pedestrians. On the other hand, the existing video-based re-ID also requires at least one photo of the person of interest. In this paper, we focus on a more general and practical setting, *i.e.*, person tube retrieval via language description, which is language-query-based and not limited to pedestrians.

Video retrieval. Video retrieval (*e.g.*, multimedia event detection (Fan et al. 2017)), activity recognition (Qi et al. 2018) and video captioning (Qi, Wang, and an Jiebo Luo 2019) are also related to this paper. Among them, most related to this paper is language-query-based video retrieval, such as retrieving videos via complex textual queries (Lin et al. 2014), retrieving video segments using natural language queries (Hendricks et al. 2017) and retrieving complex compositional activities in videos using natural language descriptions (Liu et al. 2018). The task of person tube retrieval via natural language queries can also be considered as a kind of natural language video retrieval. Different with other problems, we focus on the person in videos.

Proposed Approach

In this section, we first describe how to extract person tube representation in Section . We then list several methods to extract textual descriptions in Section . At last, we show how MSSP jointly embeds tube representations and description representations into a shared latent space in Section .

Person Tube Representation

Person tube retrieval via language requires much more abundant information than image-based re-ID (Zheng, Yang, and Hauptmann 2016) or person search (Li et al. 2017b). For example, the natural language description for image-based person search can be "the woman is wearing a long, bright orange gown with a white belt at her waist", which only involves the person appearance information. However, for person tube retrieval, the textual query can be "a man in a blue workout suit throws a javelin along a track before

falling to the ground". The description includes not only appearance (*e.g.*, 'man' and 'blue workout suit'), but also action (*e.g.*, 'throw a javelin' and 'falling') and scene (*e.g.*, 'track' and 'ground'). Therefore, person tube retrieval requires more representations than image-based person search, which should contain the follows.

- Person representation. Similar to image-based re-ID or person search, it is about the appearance RGB feature of the cropped person. In this paper, we adopt a 2D residual networks (ResNet) (He et al. 2016) to extract the person representation. The network is pretrained on the large-scale image dataset ImageNet (Russakovsky et al. 2015).
- Scene representation. The frame corresponding to the cropped person provides the scene information. We use the same 2D ResNet as the person feature extraction to obtain the scene representation.
- Action representation. The action information lies in the consecutive cropped persons and video frames. We split a person tube and its corresponding video clip into multiple parts, with each part including 16 cropped persons and 16 video frames. Then, we apply a 3D ResNet (pretrained on the large-scale action dataset Kinetics (Carreira and Zisserman 2017)) to extract the action representation. The cropped person action representation and the video frame action representation are concatenated as the action representation of the entire part.

We illustrate the process of person tube feature extraction in Fig. 2. For simplicity, we refer to a combination of 16 cropped persons and its 16 corresponding video frames as an element-tube. The 16 frame-level person and scene features of an element-tube are averaged as a single feature, respectively, to represent the person and scene. They are then concatenated with the action representation, forming the person tube representation. Any number of successive element-tubes forming a person sub-tube. The sub-tube will be used in Section .

Textual Description Representation

GRU Description sentence is a kind of sequence data. To obtain the textual description representation, a natural method is to exploit a Recurrent Neural Network (RNN), *e.g.*, Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber 1997) or Gated Recurrent Unit (GRU) (Cho et al. 2014), to encode the textual description. Since empirical results show that there is not a clear winner between LSTMs and GRUs (Chung et al. 2014), we only use GRUs as the implementations of RNN for natural language encoding. Bidirectional GRUs are also evaluated in this paper.

CNN Convolutional Neural Networks (CNNs) have achieved success in sentence-level classification tasks (Kim 2014). We therefore adopt CNNs to encode natural language descriptions in this paper. Since sentences are one-dimensional data, 1D CNNs are used for sentence embedding. The 1D convolutional kernel can be seen as a window which slides along the description. We adopt multiple kernel sizes, to capture local sentence information with multiple scales.

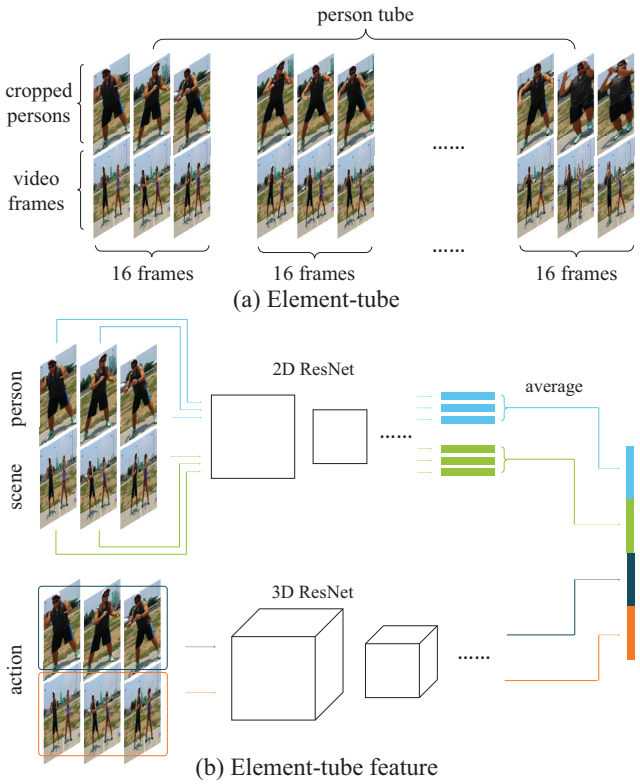


Figure 2: Illustration of element-tube and how to extract its representation. (a) Element-tube. We split a person tube and its corresponding video clip into multiple element-tubes. Each element-tube consists of 16 frames. (b) Element-tube feature extraction. For person and scene, we exploit a 2D ResNet to extract the feature for each cropped person and video frame. The cropped person features and video frame features are then averaged, respectively. For action, we use a 3D ResNet to extract the feature for the consecutive cropped persons and video frames. The person, scene and action features are concatenated as the representation of an element-tube.

HGLMM Apart from the above deep neural network methods, we also use the Fisher Vectors based on a hybrid Gaussian-Laplacian mixture model (HGLMM) (Klein et al. 2015) to represent sentences. The text is represented as a set of vectors obtained by the word2vec (Mikolov et al. 2013), which is then converted to the Fisher Vector.

Multi-scale Structure Preservation (MSSP)

For vision-language retrieval tasks, a core problem is how to measure the semantic similarity between visual data (*e.g.*, an image or video) and textual data (a sentence or phrase). A common solution is to learn a joint embedding for visual data and textual data into a common feature space where features from the two different views can be compared directly. We design the Multi-Scale Structure Preservation (MSSP) to learn linear transformations of visual and textual features to the shared space with a ranking loss.

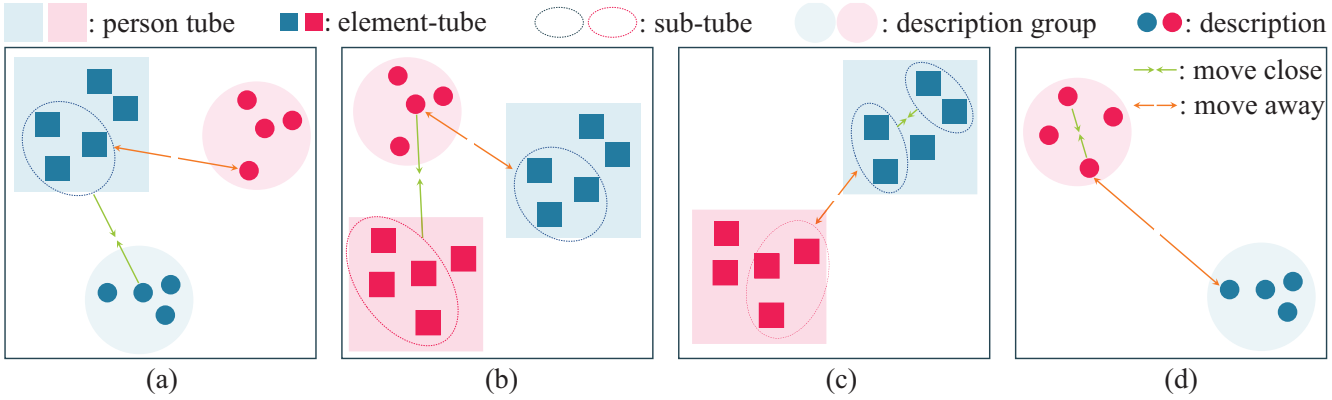


Figure 3: Illustration of how MSSP influences the distribution of visual and textual data in the joint feature space. (a) $l_{x\to y}$: encourage the related description to move close to the sub-tube and keep the unrelated description away. (b) $l_{y\to x}$: encourage the correct sub-tube to move close to the description and keep the incorrect sub-tube away. (c) $l_{x\to x}$: enforce sub-tubes from the same person to move closer and encourage sub-tubes from different person to stay away. (d) $l_{y\to y}$: enforce descriptions from the same person to move closer and encourage descriptions from different person to stay away.

Multi-Scale Person Sub-tube Suppose a person tube X contains M element-tubes $\{x_i, x_2, \cdots, x_M\}$. As we mentioned in Section , an element-tube consists of 16 consecutive cropped persons and 16 corresponding video frames. Based on the element-tube, we further define the person sub-tube. A person sub-tube is composed of any number of successive element-tubes. The person sub-tube feature can be represented as the average of element-tube features,

$$\sigma(X, a, b) = \frac{1}{b-a+1}\sum_{i=a}^{b} x_i, 1 \le a < b \le M. \quad (1)$$

When $a = b$, the sub-tube degrades to an element-tube. When $a = 1$ and $b = M$, the sub-tube equals to its corresponding person tube.

By Eq. 1, we can sample a large number of sub-tubes with multiple scales from a person tube. MSSP works based on these sub-tubes. During evaluation, we fix a to 1 and b to M.

Bidirectional Ranking To build up the relationship between the visual modality and textual modality, we exploit a ranking loss that applies a margin-based penalty to incorrect annotations that get ranked higher than correct ones for each training person sub-tube. Suppose the textual descriptions about a person tube is denoted as $Y = \{y_1, y_2, \cdots, y_N\}$, where y_i is a description and N is the number of descriptions. This loss can be formulated as follows,

$$\begin{aligned} l_{x\to y} = \max[0,\ &\cos(\sigma(X, a, b), \gamma(Y^+, i)) + \delta \\ &- \cos(\sigma(X, a, b), \gamma(Y^-, j))], \end{aligned} \quad (2)$$

where $x \to y$ indicates this loss is from visual modality to textual modality, $\delta > 0$ is the margin, $\cos(\cdot)$ is the cosine distance and $\gamma(\cdot)$ is a random access function which randomly selects a description from the set Y. The symbol $+$ and $-$ denote Y is related and unrelated to X, respectively.

The $l_{x\to y}$ loss function implements the constraint that $\cos(\sigma(X, a, b), \gamma(Y^+)) + \delta < \cos(\sigma(X, a, b), \gamma(Y^-)))$. This constraint encourages that the representation of person sub-tube should be close to that of the matching description, but stay away from that of the unrelated.

Similarly, we can establish the constraint from the textual modality to the visual modality as (Yamaguchi et al. 2017):

$$\begin{aligned} l_{y\to x} = \max[0,\ &\cos(\gamma(Y, i), \sigma(X^+, a, b)) + \delta \\ &- \cos(\gamma(Y, i), \sigma(X^-, c, d))]. \end{aligned} \quad (3)$$

The loss Eq. 2 and Eq. 3 compose the bidirectional ranking between the visual data and textual data.

External and Internal Structure Preservation Besides the inter-modality constraint, which aims at encouraging correct sub-tubes or descriptions to get ranked higher than incorrect ones, the intra-modality constraints should also be considered. Inspired by person re-ID, in which the learned discriminative representations of the same person should get closed in the feature space, we encourage the embedded person sub-tube representation of the same person to stay close to each other,

$$\begin{aligned} l_{x\to x} = \max[0,\ &\cos(\sigma(X_i, a, b), \sigma(X_i, c, d)) + \delta \\ &- \cos(\sigma(X_i, a, b), \sigma(X_j, e, f))], \end{aligned} \quad (4)$$

where X_i and X_j represent the i-th and j-th person, respectively. This loss preserves neighborhood structure within the visual modality. Specifically, minimizing the distance $\cos(\sigma(X_i, a, b), \sigma(X_i, c, d))$ is referred to the internal structure preservation. It enforces the sub-tubes of the same person tube to move closer. Meanwhile, maximizing the distance $\cos(\sigma(X_i, a, b), \sigma(X_j, e, f))$, which is referred to the external structure preservation, encourages sub-tubes to stay away from these of other persons.

Similarly, the textual modality can also be applied to the external-internal structure preservation,

$$\begin{aligned} l_{y\to y} = \max[0,\ &\cos(\gamma(Y_i, a), \gamma(Y_i, b)) + \delta \\ &- \cos(\gamma(Y_i, a), \gamma(Y_j, c))]. \end{aligned} \quad (5)$$

where Y_i and Y_j denote the description sets of the i-th and j-th persons, respectively. This loss encourages the descriptions belonging to the same person tube be close to each

other and the descriptions from different person tubes to stay away from each other.

Combining the loss Eq. 2, Eq. 3, Eq. 4 and Eq. 5 results the multi-scale external-internal structure preservation loss,

$$\mathcal{L} = \lambda_{xy} l_{x \to y} + \lambda_{xy} l_{y \to x} + \lambda_{xx} l_{x \to x} + \lambda_{yy} l_{y \to y}, \quad (6)$$

where λ_{xy}, λ_{yx}, λ_{xx} and λ_{yy} are non-negative hyper-parameters. We illustrate how this loss influences the distribution of visual and textual data in the joint feature space in Fig. 3.

Experiments

Datasets

ActivityNet-PTRLD Dataset ActivityNet (Heilbron et al. 2015) is a large-scale video benchmark for human activity understanding. Yamaguchi et al (Yamaguchi et al. 2017) collected a part of short video clips from ActivityNet and cropped the person in the clips. Each of the cropped persons is then described by five annotators, resulting five different descriptions. The original dataset (Yamaguchi et al. 2017) is designed for spatio-temporal person retrieval in videos, where person detection and tracking are also involved. Persons in the dataset are therefore cropped at the rate of only one frame per second. We complement the dataset by cropping person in every frame, resulting the **ActivityNet-P**erson **T**ube **R**etrieval via **L**anguage **D**escription (ActivityNet-PTRLD) dataset. The dataset contains 6,068 persons with 1,578,080 bounding boxes and 30,340 annotations in total. 5,500 persons are for training, 284 for validation and 284 for evaluation.

CUHK-PEDES Dataset CUHK-PEDES (Li et al. 2017b) is a large-scale dataset for image-based person retrieval via language, including 40,206 images of 13,003 persons from existing person re-identification datasets and 80,412 descriptions. On average, each image contains two different textual descriptions. The training set has 34,054 images, 11,003 persons and 68,126 descriptions. The validation set has 3,078 images, 1,000 persons and 6,158 descriptions. The test set has 3,074 images, 1,000 persons and 6,156 descriptions. We use this dataset for the evaluation of image-based person search with natural language description.

MSR-VTT Dataset MSR-VTT (Xu et al. 2016) is a large-scale video benchmark for video captioning. The dataset includes 10,000 web video clips with 41.2 hours and 200,000 clip-sentence pairs in total. Each clip is annotated with about 20 natural sentences. In this paper, we use this dataset for the evaluation of video retrieval via language. We follow the same data split as 6,513, 2,990 and 497 clips in the training, testing and validation sets, respectively.

Implementaion

We exploit the 2D and 3D ResNet-152 to extract element-tube representation, resulting the $2,048 \times 4 = 8,192$ dimensional visual feature vector. The size of word embedding is set to 300. The hidden state size of GRU is set to 512. For the sentence-embedding CNN, we use three independent 1D CNN layers, with kernel (windows) sizes 3,4 and 5, respectively. The output channel size for all layers is set to 256. Each output of CNN layer is max-pooled to a 256-dimensional vector. The three 256-dimensional vectors are then concatenated to the $256 \times 3 = 768$ textual feature vector. For the HGLMM feature vector, we only use the first 1,000 dimension. We have implemented our algorithm using both PaddlePaddle and PyTorch, which have shown similar performance.

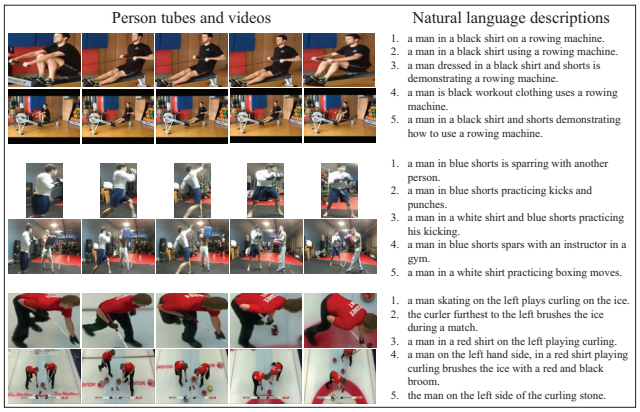


Figure 4: Examples of the ActivityNet-PTRL dataset. The person of interest is cropped in each frame. Each person is described by five different natural language annotations.

To project the visual feature and textual feature to the same number of dimensions, we use fully-connected (FC) layers to transform them. For visual modality, we append two FC layers, with 2,048 and 512 units, respectively. For textual modality, we append one FC layer with 512 units after the output of GRU or CNN. For HGLMM, we append two FC layers with the same number of units as the visual modality. Rectified Linear Unit (ReLU) is used between FC layers and batch normalization is used after the last FC layer.

Models are trained for 2,500 iterations, with batch size 1,500 and learning rate 0.01. Unless otherwise specified, λ_{xy}, λ_{yx}, λ_{xx} and λ_{yy} are set to 1.0, 2.0, 0.001 and 0.1, respectively.

Person Tube Retrieval via Language

Comparison with Other Methods We compare our method with two baselines, *i.e.*, contrastive cosine embedding and triplet cosine embedding, and three existing methods, *i.e.*, Canonical Correlation Analysis (CCA) (Hardoon, Szedmák, and Shawe-Taylor 2004), Deep Structure Preserving Embedding (DSPE) (Wang, Li, and Lazebnik 2016) and DSPE++ (Yamaguchi et al. 2017). We list the experimental results in Table 1. Our MSSP method with the HGLMM textual description embedding achieves the highest accuracy. We can make the following conclusions.

- Compared with the baselines, *i.e.*, contrastive and triplet cosine embedding, bidirectional ranking and structure preservation help method considerably increase accuracy. For example, with the HGLMM description embedding, MSSP outperforms triplet cosine embedding by 7.5% on the top-1 accuracy.

Table 1: Method comparison for person tube retrieval via language. "Bi-GRU" denotes bidirectional GRU.

Methods		textual feature	Accuracy (%)		
			Top-1	Top-5	Top 10
Baselines	Contrastive	1-layer GRU	14.4	37.7	51.4
		2-layer GRU	16.0	40.2	52.3
		2-layer Bi-GRU	18.2	48.7	60.3
		CNN	10.1	20.8	37.5
		HGLMM	21.9	54.0	69.7
	Triplet	1-layer GRU	20.5	55.7	70.2
		2-layer GRU	22.2	58.0	71.7
		2-layer Bi-GRU	28.3	63.4	75.1
		CNN	15.9	27.1	41.1
		HGLMM	33.8	71.3	84.2
CCA		HGLMM	34.6	72.7	85.6
DSPE		1-layer GRU	26.5	63.8	76.3
		2-layer GRU	28.2	64.7	78.3
		2-layer Bi-GRU	35.9	68.8	79.9
		CNN	25.0	56.5	70.7
		HGLMM	37.4	76.8	88.3
DSPE++		1-layer GRU	26.5	63.2	76.1
		2-layer GRU	27.8	64.6	77.9
		2-layer Bi-GRU	34.1	68.5	80.0
		CNN	24.2	56.1	70.9
		HGLMM	38.8	76.7	88.1
MSSP (Ours)		1-layer GRU	28.6	64.1	76.7
		2-layer GRU	29.4	65.6	78.8
		2-layer Bi-GRU	35.0	69.1	80.7
		CNN	24.9	58.5	72.2
		HGLMM	**41.3**	**77.4**	**89.3**

- The multi-scale person sub-tube and visual internal structure preservation further improve the structure preservation methods. For example, with the HGLMM description embedding, MSSP outperforms DSPE++ by 2.5%.
- Among the GRU, CNN and HGLMM textual description embedding methods, HGLMM achieves the best performance. Bidirectional structure can efficiently improve GRU for description embedding.

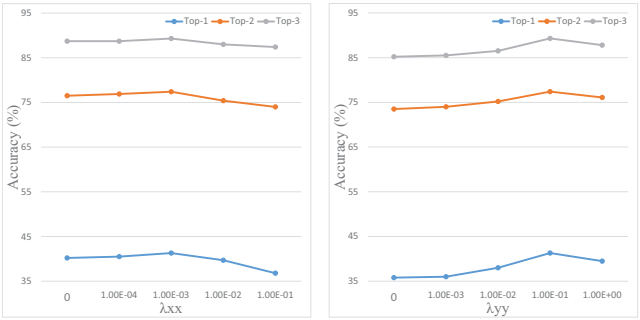
Ablation Study **1. How do visual representations influence the accuracy?** We evaluate the influence of visual representations by applying different combinations of them. The experimental results are listed in Table 2. Based on the person representaion, scene and action representations can further improve retrieval accuracy.

Specifically, comparing "person" and "person+scene", the "scene" representation improves 1.3% on top-1 accuracy. Similarly, the "person+action" combination outperforms the single "person" representation by 10.8% on top-1.

2. How do visual and textual structure preservation influence the accuracy? We respectively evaluate visual and textual structure preservation with the HGLMM textual description embedding. When exploring the visual structure preservation, *i.e.*, how the change of λ_{xx} affects the accuracy, λ_{xy}, λ_{xy} and λ_{yy} are fixed to 1.0, 2.0 and 0.1. When exploring the textual structure preservation, *i.e.*, how the change of λ_{yy} affects the accuracy, λ_{xy}, λ_{xy} and λ_{xx} are fixed to 1.0, 2.0 and 0.001.

Table 2: Influence of visual representations. The symbol "✓" denotes the representation is used. HGLMM is exploited for textual description embedding.

Visual representation			Accuracy(%)		
Person	Scene	Action	Top-1	Top-5	Top 10
✓			27.8	58.8	70.9
✓	✓		29.1	63.6	77.1
✓		✓	38.6	74.3	88.0
✓	✓	✓	41.3	77.4	89.3



(a) Visual preservation (b) Textual preservation

Figure 5: Influence of structure preservation. HGLMM is exploited for textual description embedding. (a) Accuracy change respect to λ_{xx}. (b) Accuracy change respect to λ_{yy}.

Experimental results are shown in Fig. 5. As λ_{xx} and λ_{yy} increase, the accuracy first increases and then decreases. Specifically, $\lambda_{xx} = 0.001$ and $\lambda_{yy} = 0.1$ achieve the highest accuracy. This indicates that appropriate structure preservation can improve the retrieval accuracy. However, excessive structure preservation damages the performance. The reason is that preserving the visual and textual too much decreases the diversity of the visual and textual representations. The detail of visual and textual data is lost and the representation ability of feature is impaired.

Retrieval Result Visualization We show some experimental results in Fig. 6. The examples respectively involve the case where the natural language query contains one person and two persons.

Our method can effectively exploit query information and output the reasonable person tubes. For the first example, all the retrieved person tubes are very similar to each other and our method successfully ranks the ground truth as top-1. For the second example, our method can comprehensively understand the query. Therefore, both of the top-2 video clips contain two persons and both of the target persons are in red.

The examples also demonstrate the efficacy of the "person + scene + action" representation. Without the "scene" representation, our method may not find the balance beam in the first example. Without the "person" representation, our method may not make a difference between the rank-1 and rank5 in the second example.

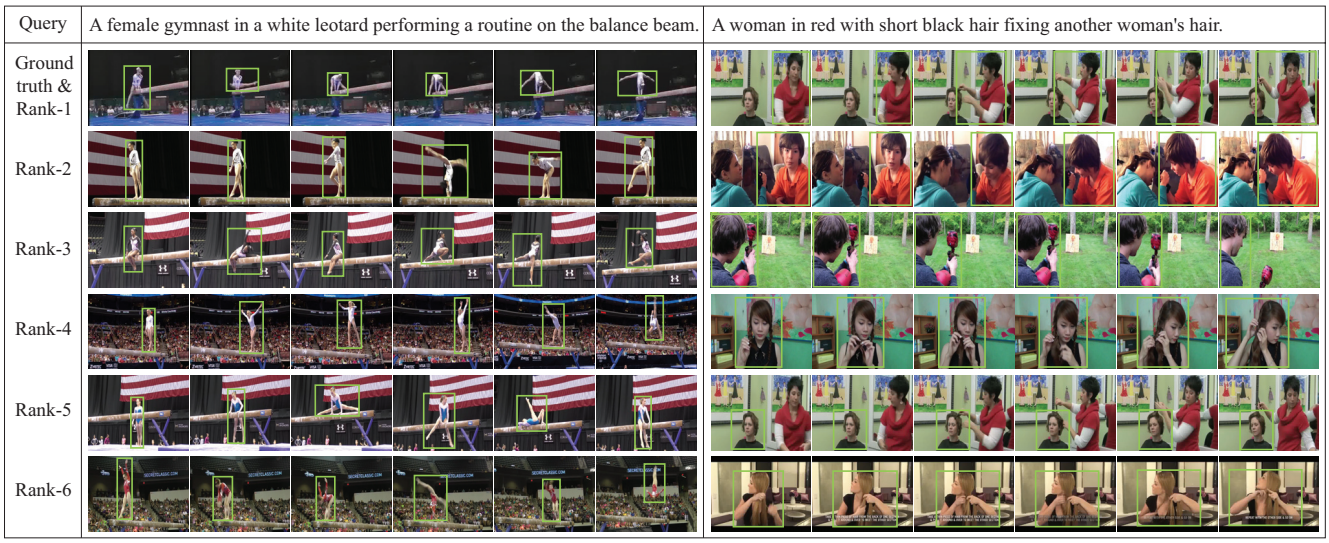


Figure 6: Visualization of person tube retrieval results.

Table 3: Method comparison for image-based person search via language on CUHK-PEDES.

Methods		Top-1	Top-5	Top 10
VGG-16	Neural Talk (Vinyals et al. 2015)	13.66	-	41.72
	CNN-RNN (Reed et al. 2016)	8.07	-	32.47
	GNA-RNN (Li et al. 2017b)	19.05	-	53.64
	IATV (Li et al. 2017a)	25.94	-	60.48
	Dual Path (Zheng et al. 2017b)	32.15	54.42	64.30
	MSSP-HGLMM (Ours)	**35.90**	**59.88**	**70.95**
ResNet-50	Dual Path (Zheng et al. 2017b)	44.40	66.26	75.07
	GLA (Chen et al. 2018a)	43.58	66.93	76.26
	MSSP-HGLMM (Ours)	**44.92**	**70.38**	**77.10**

Image-based Person Search via Language

To evaluate MSSP on the task of image-based person search, we replace the element-tube in ActivityNet-PTRL with the cropped person in CUHK-PEDES. The cropped images about the same person form a person group, which is corresponding to a person tube. During training, we randomly select multiple cropped images from a person group and average their features, to form the representation of the multi-scale sub-group. Since the image-based person retrieval focused on the person appearance, we only use the 2D CNNs to extract the appearance representation.

Experimental results are listed in Table 3. Our method also effectively improves the retrieval accuracy. For example, when using VGG-16 feature, our method outperforms Dual Path by 3.75% on top-1. When using ResNet-50 feature, our method outperforms GLA by 1.34% on top-1.

Video Retrieval via Language

At last, we apply MSSP to the task of video retrieval via language on the MSR-VTT dataset. Similar to person tube, we split each video clip into multiple element-clips on average, with each 16 frames. For the video representation, we use the pretrained 2D and 3D ResNet152s to extract the appearance and action representations. Any number of element-clips form the sub-clip. We replace the element-tube in ActivityNet-PTRL with the sub-clip in MSR-VTT.

Table 4: Method comparison for video retrieval via language on MSR-VTT.

Methods	textual feature	Accuracy (%)		
		Top-1	Top-5	Top 10
Contrastive	2-layer Bi-GRU	23.4	65.4	79.0
	HGLMM	25.0	65.8	79.3
Triplet	2-layer Bi-GRU	26.1	66.6	80.1
	HGLMM	29.3	69.2	82.2
CCA	HGLMM	29.6	69.0	82.4
DSPE	2-layer Bi-GRU	29.5	69.8	82.7
	HGLMM	33.1	71.4	84.5
DSPE++	2-layer Bi-GRU	29.2	69.2	82.1
	HGLMM	32.7	71.4	84.2
MSSP (Ours)	2-layer Bi-GRU	30.2	70.0	83.1
	HGLMM	**33.1**	**72.2**	**84.8**

Experimental results are listed in Table 4. Compared with the contrastive and triplet cosine embedding methods, our method considerably improves the retrieval accuracy. For example, with HGLMM description encoding, our method outperforms contrastive cosine embedding by 8.1% on top-1. Compared with DSPE and DSPE++, MSSP also achieves a little improvement.

Conclusion

To transform tubes and descriptions into a shared latent space, we propose a Multi-Scale Structure Preservation (MSSP) approach for person tube retrieval via language description. We conduct experiment on person tube retrieval via language, image-based person search via language and video retrieval via language, demonstrating that MSSP can efficiently improve the accuracy on these tasks.

References

Carreira, J., and Zisserman, A. 2017. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*.

Chen, D.; Li, H.; Liu, X.; Shen, Y.; Shao, J.; Yuan, Z.; and Wang, X. 2018a. Improving deep visual representation for person re-identification by global and local image-language association. In *ECCV*.

Chen, D.; Li, H.; Xiao, T.; Yi, S.; and Wang, X. 2018b. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In *CVPR*.

Cho, K.; van Merrienboer, B.; Gülçehre, Ç.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP*.

Chung, J.; Gülçehre, Ç.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv* abs/1412.3555.

Ding, Y.; Fan, H.; Xu, M.; and Yang, Y. 2019. Adaptive exploration for unsupervised person re-identification. *ACM Transactions on Multimedia Computing, Communications, and Applications TOMCCAP*.

Fan, H.; Chang, X.; Cheng, D.; Yang, Y.; Xu, D.; and Hauptmann, A. G. 2017. Complex event detection by identifying reliable shots from untrimmed videos. In *ICCV*.

Fan, H.; Zheng, L.; Yan, C.; and Yang, Y. 2018. Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications TOMCCAP* 14(4).

Hara, K.; Kataoka, H.; and Satoh, Y. 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *CVPR*.

Hardoon, D. R.; Szedmák, S.; and Shawe-Taylor, J. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation* 16(12):2639–2664.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Heilbron, F. C.; Escorcia, V.; Ghanem, B.; and Niebles, J. C. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*.

Hendricks, L. A.; Wang, O.; Shechtman, E.; Sivic, J.; Darrell, T.; and Russell, B. 2017. Localizing moments in video with natural language. In *ICCV*.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural Computation* 9(8):1735–1780.

Kim, Y. 2014. Convolutional neural networks for sentence classification. In *EMNLP*.

Klein, B.; Lev, G.; Sadeh, G.; and Wolf, L. 2015. Associating neural word embeddings with deep image representations using fisher vectors. In *CVPR*.

Li, S.; Xiao, T.; Li, H.; Yang, W.; and Wang, X. 2017a. Identity-aware textual-visual matching with latent co-attention. In *ICCV*.

Li, S.; Xiao, T.; Li, H.; Zhou, B.; Yue, D.; and Wang, X. 2017b. Person search with natural language description. In *CVPR*.

Lin, D.; Fidler, S.; Kong, C.; and Urtasun, R. 2014. Visual semantic search: Retrieving videos via complex textual queries. In *CVPR*.

Liu, B.; Yeung, S.; Chou, E.; Huang, D.; Fei-Fei, L.; and Niebles, J. C. 2018. Temporal modular networks for retrieving complex compositional activities in videos. In *ECCV*.

Miao, J.; Wu, Y.; Liu, P.; Ding, Y.; and Yang, Y. 2019. Pose-guided feature alignment for occluded person re-identification. In *ICCV*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*.

Qi, M.; Qin, J.; Li, A.; Wang, Y.; Luo, J.; and Gool, L. V. 2018. stagnet: An attentive semantic RNN for group activity recognition. In *ECCV*.

Qi, M.; Li, W.; Yang, Z.; Wang, Y.; and Luo, J. 2019. Attentive relational networks for mapping images to scene graphs. In *CVPR*.

Qi, M.; Wang, Y.; and an Jiebo Luo, A. L. 2019. Sports video captioning via attentive motion representation and group relationship modeling. *IEEE Transactions on Circuits and Systems for Video Technology*.

Quan, R.; Dong, X.; Wu, Y.; Zhu, L.; and Yang, Y. 2019. Auto-reid: Searching for a part-aware convnet for person re-identification. In *ICCV*.

Reed, S. E.; Akata, Z.; Lee, H.; and Schiele, B. 2016. Learning deep representations of fine-grained visual descriptions. In *CVPR*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. S.; Berg, A. C.; and Li, F. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115(3):211–252.

Song, C.; Huang, Y.; Ouyang, W.; and Wang, L. 2018. Mask-guided contrastive attention model for person re-identification. In *CVPR*.

Sun, X., and Zheng, L. 2019. Dissecting person re-identification from the viewpoint of viewpoint. In *CVPR*.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *CVPR*.

Wang, H.; Gong, S.; Zhu, X.; and Xiang, T. 2016. Human-in-the-loop person re-identification. In *ECCV*.

Wang, L.; Li, Y.; and Lazebnik, S. 2016. Learning deep structure-preserving image-text embeddings. In *CVPR*.

Xiao, T.; Li, S.; Wang, B.; Lin, L.; and Wang, X. 2017. Joint detection and identification feature learning for person search. In *CVPR*.

Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 2016. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR*.

Yamaguchi, M.; Saito, K.; Ushiku, Y.; and Harada, T. 2017. Spatio-temporal person retrieval via natural language queries. In *ICCV*.

Yang, X.; Wang, M.; Hong, R.; Tian, Q.; and Rui, Y. 2017. Enhancing person re-identification in a self-trained subspace. *TOMCCAP* 13(3):27:1–27:23.

Yang, X.; Wang, M.; and Tao, D. 2018. Person re-identification with metric learning using privileged information. *IEEE Trans. Image Processing*.

Zeng, Z.; Li, Z.; Cheng, D.; Zhang, H.; Zhan, K.; and Yang, Y. 2018. Two-stream multirate recurrent neural network for video-based pedestrian reidentification. *IEEE Trans. Industrial Informatics* 14(7):3179–3186.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable person re-identification: A benchmark. In *ICCV*.

Zheng, L.; Zhang, H.; Sun, S.; Chandraker, M.; Yang, Y.; and Tian, Q. 2017a. Person re-identification in the wild. In *CVPR*.

Zheng, Z.; Zheng, L.; Garrett, M.; Yang, Y.; and Shen, Y. 2017b. Dual-path convolutional image-text embedding. *arXiv* abs/1711.05535.

Zheng, L.; Yang, Y.; and Hauptmann, A. G. 2016. Person re-identification: Past, present and future. *arXiv* abs/1610.02984.