IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received September 27, 2020, accepted October 2, 2020, date of publication October 12, 2020, date of current version October 22, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3030258

Video-Based Human Motion Capture Data Retrieval via MotionSet Network

TINGXIN REN[1], WEI LI[1], ZIFEI JIANG[1], XUEQING LI[1], YAN HUANG[1], AND JINGLIANG PENG[2,3], (Member, IEEE)
[1]School of Software, Shandong University, Jinan 250101, China
[2]Shandong Provincial Key Laboratory of Network Based Intelligent Computing, University of Jinan, Jinan 250022, China
[3]School of Information Science and Engineering, University of Jinan, Jinan 250022, China

Corresponding author: Yan Huang (yan.h@sdu.edu.cn)

This work was supported by the Fundamental Research Funds of Shandong University.

ABSTRACT Content-based human motion capture (MoCap) data retrieval facilitates reusing motion data that have already been captured and stored in a database. For a MoCap data retrieval system to get practically deployed, both high precision and natural interface are demanded. Targeting both, we propose a video-based human MoCap data retrieval solution in this work. It lets users to specify a query via a video clip, addresses the representational gap between video and MoCap clips and extracts discriminative motion features for precise retrieval. Specifically, the proposed scheme firstly converts each video clip or MoCap clip at a certain viewpoint to a binary silhouette sequence. Regarding a video or MoCap clip as a set of silhouette images, the proposed scheme uses a convolutional neural network, named MotionSet, to extract the discriminative motion feature of the clip. The extracted motion features are used to match a query to repository MoCap clips for the retrieval. Besides the algorithmic solution, we also contribute a human MoCap dataset and a human motion video dataset in couple that contain various action classes. Experiments show that our proposed scheme achieves an increase of around 0.25 in average MAP and costs about 1/26 time for online retrieval, when compared with the benchmark algorithm.

INDEX TERMS MotionSet, motion capture data retrieval, convolutional neural network, deep learning.

I. INTRODUCTION

There is a growing demand for motion capture (MoCap) technology in many fields including interactive virtual reality, film production, animation and so forth. However, capturing motions when needed is often not practical as motion capture systems are expensive and the capture processes are complex in general [1]. It is often desirable to retrieve and reuse motion clips that have been captured before and stored in databases. Straightforwardly, the retrieval may be done based on text labels of motion clips. However, it may be hard to fully characterize a motion segment of certain complexity by text labels. Further, different text labels may be used to describe the same motion, e.g., 'leaping' and 'jumping', 'jogging' and 'running slowly'. Semantic analysis of text labels is then required for precise motion retrieval, which itself is a challenging task. This has motivated intensive research on content-based retrieval of MoCap data instead, and we take this line as well in this work.

Content-based human MoCap data retrieval has drawn lots of research attention in recent years with many good algorithms proposed. In these algorithms, various modalities of query have been used, which include MoCap clip [1]–[11], hand-drawn sketch [12]–[14], puppet motion [15], [16], Kinect skeleton motion [5] and video clip [17]–[19]. MoCap clips are themselves hard to acquire. Hand-drawn sketches vary much across users in quality and style of drawing. Kinect skeletons have to be captured within a limited range of distance and view angle. Puppets are hard to pose for complex motions. Besides, Kinect and puppet devices may not be readily available. By contrast, video clips provide a means for natural and convenient specification of queries. For instance, the user may act out a motion of his or her interest in front of a video camera that readily comes with a laptop computer or mobile computing device. As such, we adopt the video modality of query in this work.

In this work, we propose a video-based human MoCap data retrieval scheme, which takes as input a video clip and retrieves similar MoCap clips from the repository. The key contribution of this work is a novel holistic scheme for

The associate editor coordinating the review of this manuscript and approving it for publication was Jingchang Huang.

cross-modality motion data retrieval. As the novel core component, it extracts discriminative cross-modality motion features, which is achieved by converting the original video or MoCap clips to binary silhouette sets and extracting the motion features by a MotionSet network. The proposed motion feature representation extends naturally for intra-modality retrieval as well. Besides, this work contributes a human MoCap dataset and a human motion video dataset in couple.

It should be noted that, in this work, we presume that the motion in a query video as a whole is to be searched. Should motions in a long video sequence be searched in a finer granularity, existing algorithms [20]–[23] may be used first to segment the raw video sequence into smaller clips that are then used as queries. Similarly, should motions in a long MoCap sequence be described in a finer granularity, segmentation may also be conducted [24].

The rest of this article is structured as follows. Related work is briefly reviewed in Sec. II. Overview of the proposed scheme is presented in Sec. III-A, with details of components provided in Sec. III. Experimental results are given in Sec. IV and conclusion is drawn in Sec. V.

II. RELATED WORK

For content-based MoCap data retrieval, various modalities of query have been used, such as MoCap clip [1]–[11], hand-drawn sketch [12]–[14], puppet motion [15], [16], Kinect skeleton motion [5] and video clip [17]–[19]. Generally, all the aforementioned input modalities have their pros and cons. With all factors taken into account, video-based MoCap data retrieval shows its advantages of affordability and user-friendliness. Using a commodity video camera, it is easy to specify the query by recording the user's performance.

A. VIDEO-BASED MoCap DATA RETRIEVAL

We briefly review the video-based MoCap data retrieval methods [17], [19], while a comprehensive survey of the existing MoCap data retrieval algorithms can be found in the reference [25].

Gupta *et al.* [19] make frame-by-frame alignment of the query video clip to a portion of a longer MoCap sequence using sub-sequence local normalization dynamic time warping (SLNDTW). They use the dense trajectories [26], [27] for video and MoCap data description, which is effective but time-consuming. Given a query video, Jiang *et al.* [17] reconstruct a skeleton animation by a deep learning based 3D pose estimation algorithm [28], compute a handcrafted motion signature for the skeleton animation and compare it with the motion signatures of the repository MoCap clips to get the result. The performance of retrieval is highly dependent on the quality of reconstructed 3D poses and the description power of handcrafted motion signatures.

It is worth noting that Gupta *et al.* [18] also match a video input to MoCap clips but their focus is on cross-view action recognition. For the motion description, they also adopt the time-consuming dense trajectory features [26], [27].

B. VIDEO-BASED ACTION RECOGNITION

Another field closely related to video-based MoCap data retrieval is video-based action recognition. For both fields, the core part is to effectively describe motions in videos. Recognizing human actions from video clips is an important research topic and, in this context, many methods have been proposed to describe human motions in video clips. The classical motion energy image (MEI) and motion history image (MHI) representations are proposed in the reference [29]. MEI represents where motion has occurred in an image sequence, while MHI is a scalar-valued image where intensity is a function of the motion's recency. These two descriptors usually perform well for actions that are performed in a fixed space without large movements. The space-time shape feature is proposed in the reference [30] for human action recognition. The template-based method by a maximum average correlation height (MACH) filter is proposed in the reference [31] for recognizing human actions. Besides, gait energy image (GEI) targeting human walking representation is proposed in the reference [32] for individual gait recognition.

The motion descriptors described above are mostly handcrafted. Their descriptive power is still limited. In recent years, deep learning techniques have been successfully employed to extract motion features in videos. Among various deep-learning-based video motion analysis algorithms, the GaitSet work [33] is of particular interest to us. Regarding a gait sequence as a set of silhouettes, it extracts frame-level features by a convolutional neural network (CNN) on each frame, and set-level features by set pooling the frame-level features. It is immune to permutation of frames, and can naturally integrate frames from videos filmed under different scenarios, adding to the robustness of gait recognition. As a result, the GaitSet work yields state-of-the-art performance for video-based gait recognition.

III. METHODOLOGY OF THE PROPOSED ALGORITHM

A. OVERALL FRAMEWORK

We propose a novel scheme to retrieve human MoCap clips from a database which contain motions similar to that specified in a query video clip. The flowchart of the proposed scheme is shown in Fig. 1. It is composed of two parts: offline pre-processing and online query.

In the offline pre-processing stage, we make a simple 3D avatar human model and animate it by each repository MoCap clip. For each frame of an animation, we render the posed model to several views and extract the binary human silhouettes of the rendered images. As a result, we obtain a set of silhouettes for each MoCap clip at each view. Next, we input each silhouette set to the MotionSet network to get

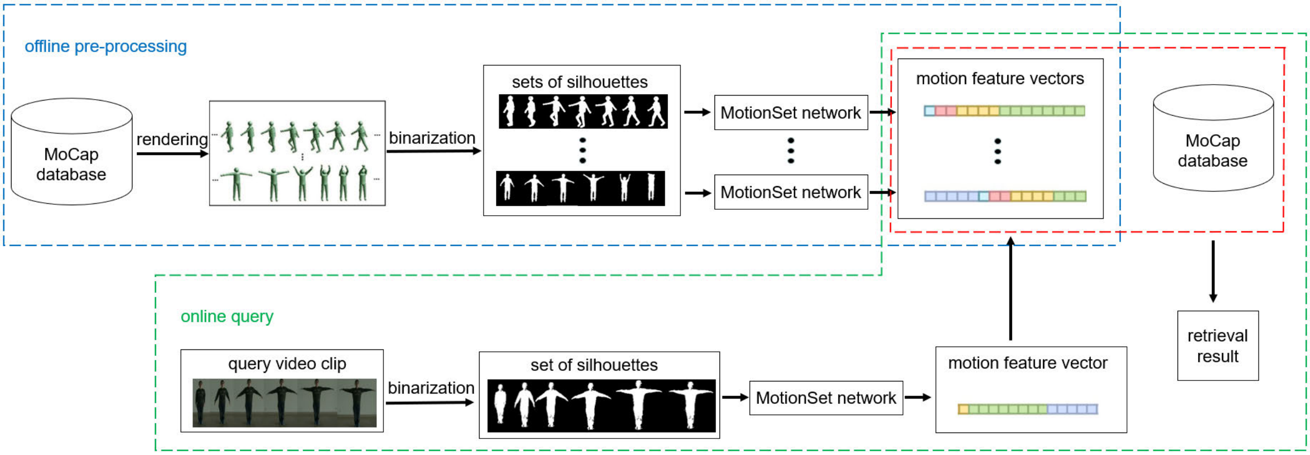


FIGURE 1. The flowchart of video-based human MoCap data retrieval via MotionSet network.

its motion feature. All these motion features are stored in association with their MoCap clips to facilitate online queries later.

In the online query stage, we extract the binary human silhouette in each query video frame by background subtraction and obtain a set of silhouettes for the query video clip. Next, we input this silhouette set to the MotionSet network to get the motion feature of the query. It is then compared with all the repository motion features to find the closest matches which give the final result of retrieval.

B. PROBLEM FORMULATION

A MoCap clip $\mathbf{M} = \{m_1, m_2, \ldots, m_n\}$ is converted to four sequences of binary silhouettes at four view angles, respectively, by $B_1(\mathbf{M}, \theta_i), \theta_i \in \{0, \pi/2, \pi, 3\pi/2\}$. A video clip $\mathbf{E} = \{e_1, e_2, \ldots, e_n\}$ is converted to a sequence of binary silhouettes by $B_2(\mathbf{E})$.

Regarding the silhouettes for a n-frame motion clip as a set, $\chi = \{x_i | i = 1, 2, \ldots, n\}$, we extract the motion feature f from χ by

$$f = H(S(F(\chi))) \qquad (1)$$

where F is a CNN that extracts frame-level features from each silhouette, S maps all the frame-level features of χ to a set-level feature through set pooling and a convolutional network, and H splices all the rows in the set-level feature map into a 1D motion feature vector.

For two motion clips, $\mathbf{C}_1$ and $\mathbf{C}_2$, with silhouettes sequences, χ_1 and χ_2, respectively. Their motion features are $f_1 = H(S(F(\chi_1)))$ and $f_2 = H(S(F(\chi_2)))$, respectively. The similarity between $\mathbf{C}_1$ and $\mathbf{C}_2$ is measured by the negated Euclidean distance between their motion features, *i.e.*, $L(\mathbf{C}_1, \mathbf{C}_2) = -||f_1 - f_2||_2$.

Details about the silhouette extraction and the motion feature extraction are provided in the following subsections, respectively.

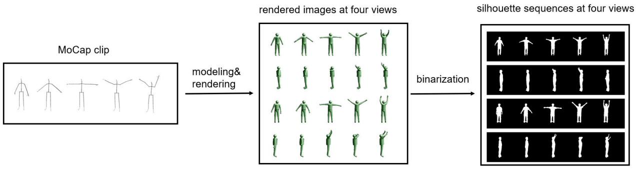

rendered images at four views
silhouette sequences at four views
MoCap clip
modeling& rendering
binarization

FIGURE 2. The process of converting a MoCap clip to four binary sihouette sequences.

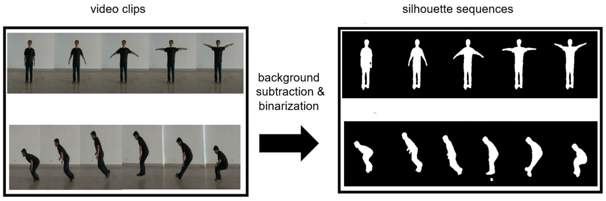


FIGURE 3. Examples of binary silhouette sequence for two video clips.

C. SILHOUETTE EXTRACTION

For each MoCap frame, we make a simple 3D avatar human model by fitting the head with a sphere and each of the other bones with a cylinder for the skeleton as posed in the current frame. Next, we render the posed 3D avatar model by orthographic projection onto four views (*i.e.*, front, back, left and right) and binarize each rendered image to get the avatar's silhouette. As a result, we obtain four binary silhouette sequences for each MoCap clip at four views, respectively. This process of converting a MoCap clip to four binary sihouette sequences is illustrated in Fig. 2.

The silhouette sequence representation for a video clip is similarly obtained but the binary silhouette for each video frame is obtained via background subtraction. Examples of binary silhouette sequence for two video clips are shown in Fig. 3.

D. MOTION FEATURE EXTRACTION

1) ARCHITECTURE OF MotionSet

In this section, we describe the neural network we use for motion feature extraction. Essentially, we adopt the network

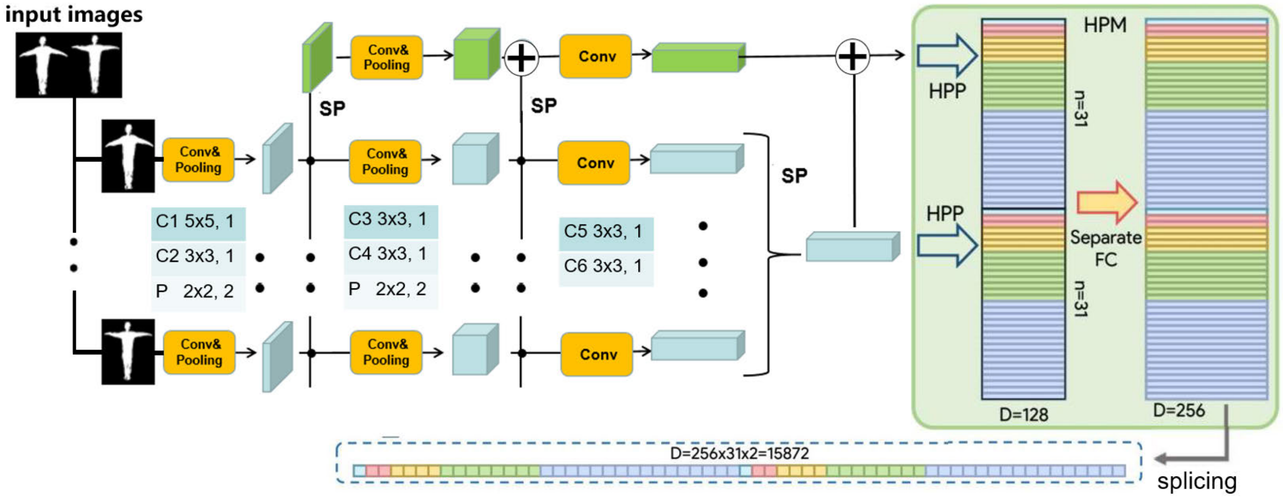


FIGURE 4. The architecture of MotionSet. 'SP' represents set pooling and 'HPM' represents horizontal pyramid mapping.

proposed for gait recognition by Chao *et al.* [33]. We use this network to describe a clip of arbitrary motion type but not gait only. Therefore, we call the network MotionSet in this work. The architecture of MotionSet is shown in Fig. 4.

As shown in Fig. 4, each silhouette frame is sent to a CNN branch to extract frame-level features at various stages. Further, at each stage, all the frame-level features are integrated by set pooling (SP) to extract the set-level feature. The set-level features are also processed by a CNN shown at the top. Regarding the network structure, the set-level branch and every frame-level branch have the same structure, except that set-pooled frame-level features are integrated into the data flow of the set-level branch as well. Finally, two feature maps result from the set-level branch and the frame-level branches, respectively, and these two feature maps are sent to a horizontal pyramid mapping (HPM) module to extract the motion feature.

2) SET POOLING

Set pooling aggregates motion information from a set of frame-level features, formulated as $z = S(V)$ with z being the set-level feature and $V = \{v_1, v_2, \ldots, v_n\}$ being the set of frame-level features. The mapping S should be permutation invariant. Statistics functions usually meet this requirement. Specifically, following Chao *et al.* [33], we use

$$S(\cdot) = 1\_1C(cat(max(\cdot) + mean(\cdot) + median(\cdot))) \quad (2)$$

where the statistics functions, max, $mean$ and $median$, are applied on the set dimension, cat means concatenate on the channel dimension and 1_1C means a 1×1 convolutional layer that weights the information extracted by the three pooling methods properly. Further, attention mechanism is uesd to improve the performance of set pooling, as illustrated in Fig. 5.

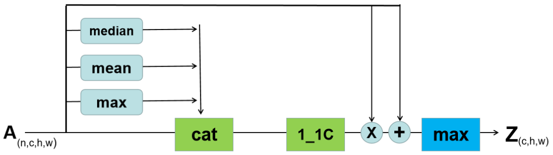


FIGURE 5. The structure of Set Pooling using attention. 'cat' and '1_1C' represent concatenate and 1 × 1 convolution, respectively. The multiplication and the addition are pointwise.

3) HORIZONTAL PYRAMID MAPPING

In order to capture features that are existent at various scales, we finally apply the horizontal pyramid mapping (HPM) module to map the extracted motion feature to a discriminative space. HPM was proposed by Chao *et al.* [33] which improves the horizontal pyramid pooling (HPP) [34] by replaying the 1×1 convolutional layer with fully connect layers (FC). As shown in Fig. 6, S scales are used for HPM. At each scale $s \in \{1, 2, \ldots, S\}$, the feature map extracted by SP is split to 2^{s-1} strips on the height dimension. On each strip $z_{s,t}$, $t \in \{1, 2, \ldots, 2^{s-1}\}$, global pooling is applied to get a 1D feature by $f'_{s,t} = maxpool(z_{s,t}) + avgpool(z_{s,t})$ where $maxpool$ and $avgpool$ mean global max pooling and global average pooling, respectively. Finally, an independent FC is applied on each $f'_{s,t}$, leading to the final motion feature f.

4) TRAINING AND TESTING

As explained above, the output of the MotionSet is a combination of $2 \times \sum_{s=1}^{S} 2^{s-1}$ feature vectors. Following Chao *et al.* [33], we use the corresponding features among different samples to compute the loss, and use batch all (BA+) triplet loss to train the network [35].

Given a video clip as the query, we input it to the MotionSet network to extract its motion feature, which is then compared with the motion features of all the MoCap clips in the repository. MoCap clips whose motion features are closest to

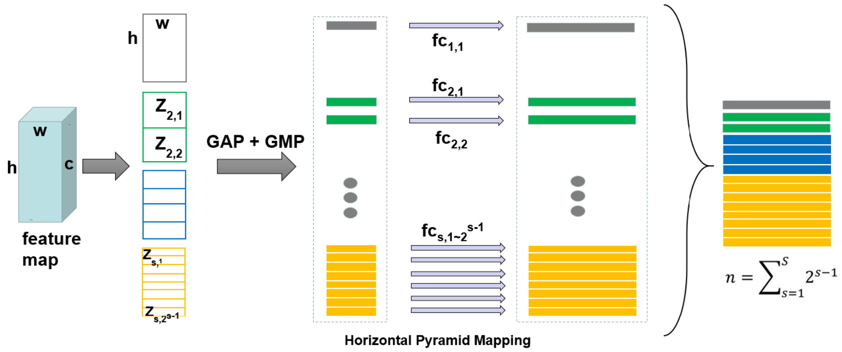


FIGURE 6. The flowchart of horizontal pyramid mapping. 'GAP' represents global average pooling and 'GMP' represents global max pooling.

TABLE 1. Anthropometric measures of the volunteers for motion data collection.

	Actress	Actor#1	Actor#2	Actor#3	Actor#4	Actor#5	Actor#6
height (cm)	153	172	176	183	175	173	177
weight (kg)	48	70	80	72	75	60	72

the query's are viewed as the most similar to the query and returned as the result.

IV. EXPERIMENTS

In this section, we conduct a comprehensive performance evaluation of the proposed video-based human MoCap data retrieval scheme. This is done by 1) comparing with a state-of-the-art method in video-based MoCap data retrieval, and 2) demonstrating the versatility and extensibility of the proposed scheme.

A. PLATFORM AND PERFORMANCE METRICS

The proposed scheme in this article was implemented with Python3.6. We run the code on a computer with NVIDIA 1080TI GPU and Ubuntu16.04 operating system.

In our experiments, the commonly used mean average precision (MAP), precision-recall curve (P-R curve), precision at n (P@n) and confusion matrix are employed for the performance evaluation.

B. DATASETS

We invited 7 volunteers, denoted as Actress, Actor#1, Actor#2, Actor#3, Actor#4, Actor#5 and Actor#6, for the motion data collection. They have significantly variant anthropometric measures, as shown in Tab. 1.

1) HUMAN MoCap DATASET

We captured human motions by the Vicon motion capture system to compose the MoCap dataset that we use for experiments. For the motion capture, 12 cameras are used and 53 markers are attached to the suit, mostly around key joints and ends of the body. The scene of motion capture and an actor wearing the suit with markers are shown in Fig. 7(a) and Fig. 7(b), respectively. Five volunteers, *i.e.*, Actress, Actor#1, Actor#2, Actor#3 and Actor#4 were employed for the motion capture. They were asked to perform actions of 20 types. Each person performed each type of action for 4 times. Hence, 400 MoCap clips were captured in total. Different MoCap clips may have different frame counts, depending on the durations of specific actions. The indexes and names of all the actions are listed in Tab. 2.

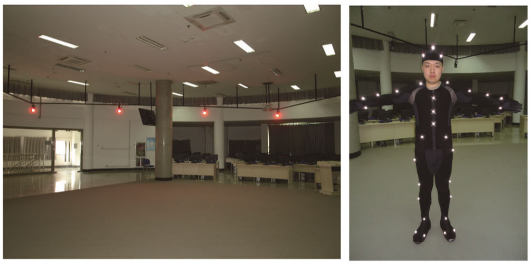

(a) Scene of Motion Capture
(b) Actor with markers

FIGURE 7. The scene of motion capture and an actor wearing the suit with markers.

2) HUMAN MOTION VIDEO DATASET

In company with the human MoCap dataset, we also compose a human motion video dataset. Four volunteers, *i.e.*, Actress, Actor#1, Actor#5 and Actor#6 were employed for the video recording. They were asked to act out the same 20 types of motions in the MoCap dataset (see Sec. IV-B1). Each person performed each type of motion for 5 times at each of 4 viewpoints (*i.e.*, front, back, left and right) in front of a monocular camera. Hence, 400 video clips were collected for each viewpoint and 1,600 video clips were collected in total. Different video clips may have different frame counts, depending on the durations of specific actions. Note that we use video clips from multiple viewpoints to train the Motion-

TABLE 2. MAP statistics of Jiang *et al.*'s method [17] and ours.

Action information		MAP statistics	
Action Index	Action Name	Jiang *et al.*'s method [17]	Ours
1	lateral raise	0.5850	0.9400
2	rope skipping	0.5600	0.9725
3	normal walking	0.4850	0.9800
4	racket waving	0.8400	1.0000
5	calisthenics movements	0.9000	0.9950
6	arms waving	0.6075	0.8775
7	basket shot	0.7200	1.0000
8	right hand waving	0.7525	0.8150
9	phone answering	0.9800	0.9975
10	parade step	0.7800	0.9975
11	sitting down	0.7175	0.9575
12	jump	0.6125	0.9975
13	bending down	0.6950	0.9950
14	punch in horse stance	0.8350	0.9350
15	basketball bouncing	0.9725	0.9975
16	front raise	0.6950	0.9500
17	rotating	0.7125	1.0000
18	walking in circles	0.5500	0.9975
19	side walking	0.3900	0.8450
20	normal running	0.5825	0.7525
Average MAP→		0.6986	0.9501

Set in order to accommodate an arbitrary viewpoint at which the query video may be shot online. Example frames from multiple video clips shot at various viewpoints are presented in Fig. 11.

C. VIDEO-BASED MoCap DATA RETRIEVAL

1) BENCHMARK ALGORITHM

As reviewed in Sec. II, two algorithms [17], [19] conduct video-based MoCap data retrieval. Gupta *et al.* [19] make fine-grained frame-by-frame alignment of the query video clip to a portion of each longer MoCap sequence, which is a time-consuming process. Similar to our algorithm, Jiang *et al.*'s method [17] also conducts whole-sequence similarity search without time-consuming frame alignment. As such, Jiang *et al.*'s method [17] is of the same type as ours, and we use it as the benchmark in performance evaluation.

2) EXPERIMENTAL RESULTS

Each of the 400 MoCap clips in our human MoCap dataset is converted to 4 binary silhouette sequences at 4 viewpoints, respectively, by the method in Sec. III-C. As a result, we obtain 1,600 synthetic silhouette sequences for all the MoCap clips. Each of the 1,600 video clips in our human motion video dataset is converted to 1 binary silhouette sequence by the method in Sec. III-C. As a result, we obtain 1,600 real silhouette sequences for all the video clips.

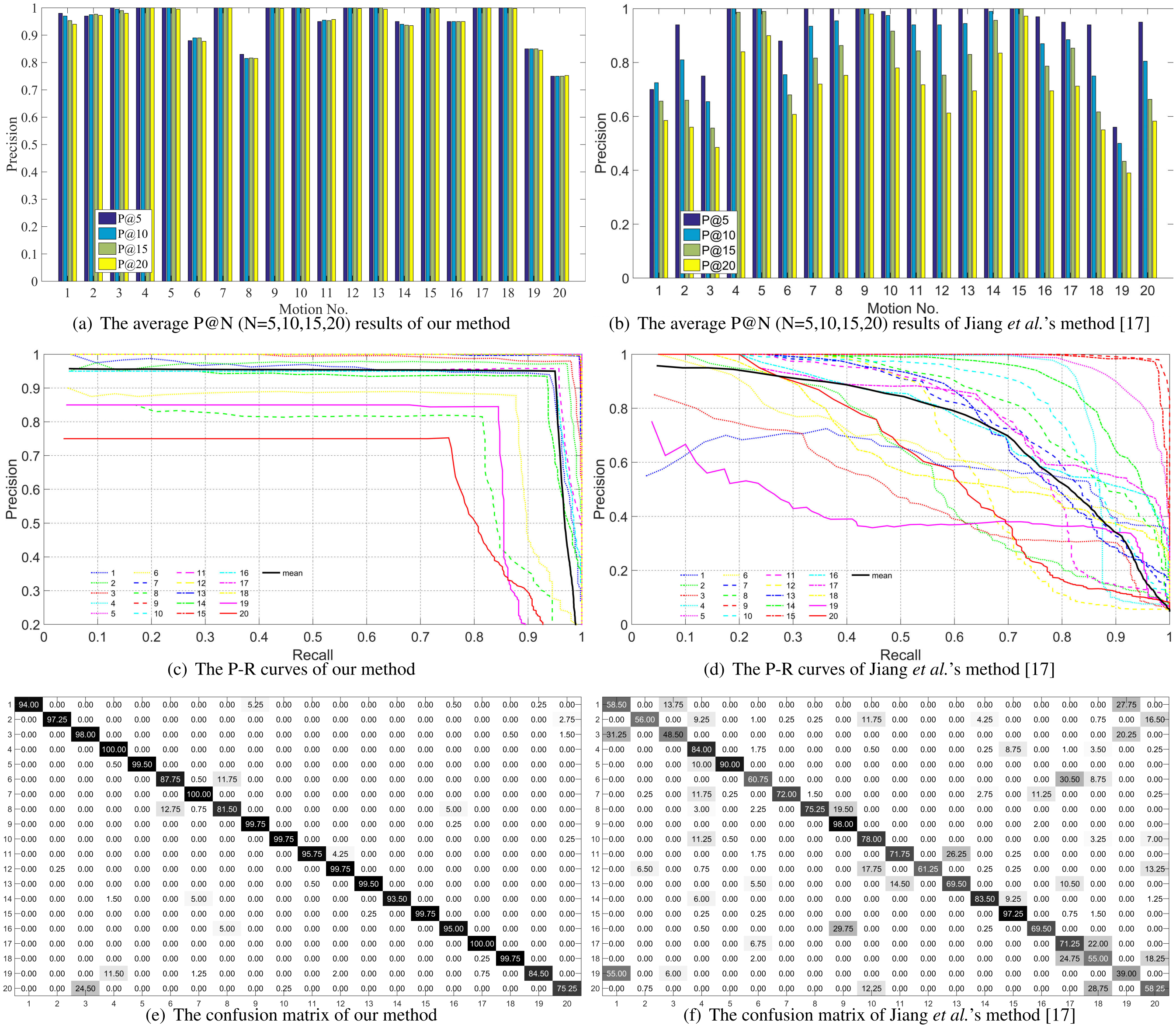
Firstly, we use the real silhouette sequences of two actors and all the synthetic silhouette sequences to train the MotionSet network. Next, we use the video clips of the rest two actors as queries and, for each of them, retrieve similar MoCap sequences from the human MoCap dataset.

For this experiment, the MAP statistics of Jiang *et al.*'s algorithm [17] and our algorithm are provided in Tab. 2, and the P@N ($n = 5, 10, 15, 20$) statistics, the P-R curves and the confusion matrices of the two algorithms are shown in Fig. 8. should be noted that the numerical results of Jiang *et al.*'s method in Tab. 2 are different from those reported in the original paper [17] as we are using different datasets in this work. From Tab. 2, we observe that our algorithm outperforms Jiang *et al.*'s algorithm [17] on most of the motion classes. As also shown in Tab. 2, the average MAPs of Jiang *et al.*'s algorithm [17] and ours are 0.6986 and 0.9501, respectively, showing a big advantage of our algorithm. From Fig. 8, we again observe significantly better performance of our algorithm.

Regarding the time efficiency, it takes 0.056s with our method and 1.5s with Jiang *et al.*'s method [17] on the average for each online retrieval on our datasets. As the motion features of the repository MoCap clips have been computed offline, the online retrieval task involves computing the motion feature of the query video (by the MotionSet network) and matching the query video's motion feature to those of the repository MoCap clips. Roughly speaking, the online retrieval time scales linearly with the number of MoCap clips in the repository dataset, if no hashing technique is applied.

D. VERSATILITY AND EXTENSIBILITY

The motion feature extracted using our proposed scheme (or proposed motion feature for brevity) is versatile as it may be used to characterize either a MoCap or a video clip. As such,



(a) The average P@N (N=5,10,15,20) results of our method

(b) The average P@N (N=5,10,15,20) results of Jiang *et al.*'s method [17]

(c) The P-R curves of our method

(d) The P-R curves of Jiang *et al.*'s method [17]

(e) The confusion matrix of our method

(f) The confusion matrix of Jiang *et al.*'s method [17]

FIGURE 8. The average P@N (N = 5, 10, 15, 20) results, P-R curves and confusion matrices of Jiang *et al.*'s method [17] and ours for video-based MoCap data retrieval.

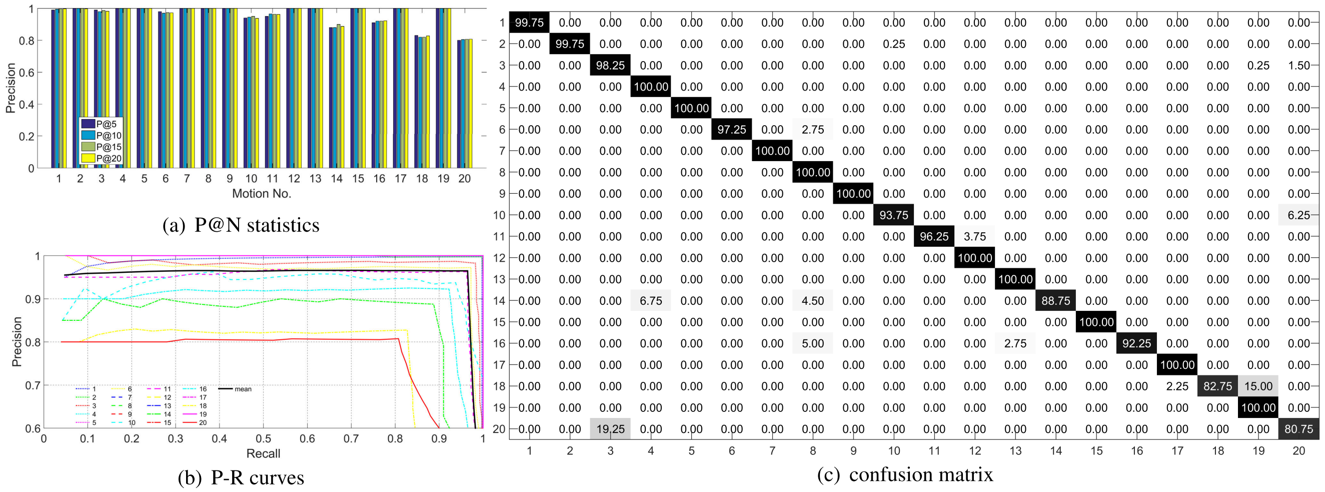
it naturally extends to support retrieval within either single modality as well. In order to demonstrate the performance of intra-modality retrieval with the proposed motion feature, we design two experiments.

In the first experiment, we use the human MoCap dataset and conduct MoCap-to-MoCap data retrieval. The MoCap clips of three performers form the gallery and their silhouette sequences are used to train the MotionSet network. The MoCap clips of the rest two performers are used as queries. In the second experiment, we use the human motion video dataset and conduct video-to-video data retrieval. The video clips of two performers form the gallery and their silhouette sequences are used to train the MotionSet network. The video clips of the rest two performers are used as queries. The results of these two experiments are plotted in Fig. 9 and Fig. 10, respectively, showing the outstanding performance of intra-modality retrieval by the proposed motion feature.

E. DISCUSSION

In order to investigate the performance of our proposed scheme on repetitive motions and combined motions, we further design two experiments as detailed below.

In both experiments, we use the same MotionSet model and the same query video dataset as for the experiment in Sec. IV-C2, but different repository MoCap datasets. For the first experiment, we extend each clip in our original MoCap dataset to two cycles by duplication. For the second



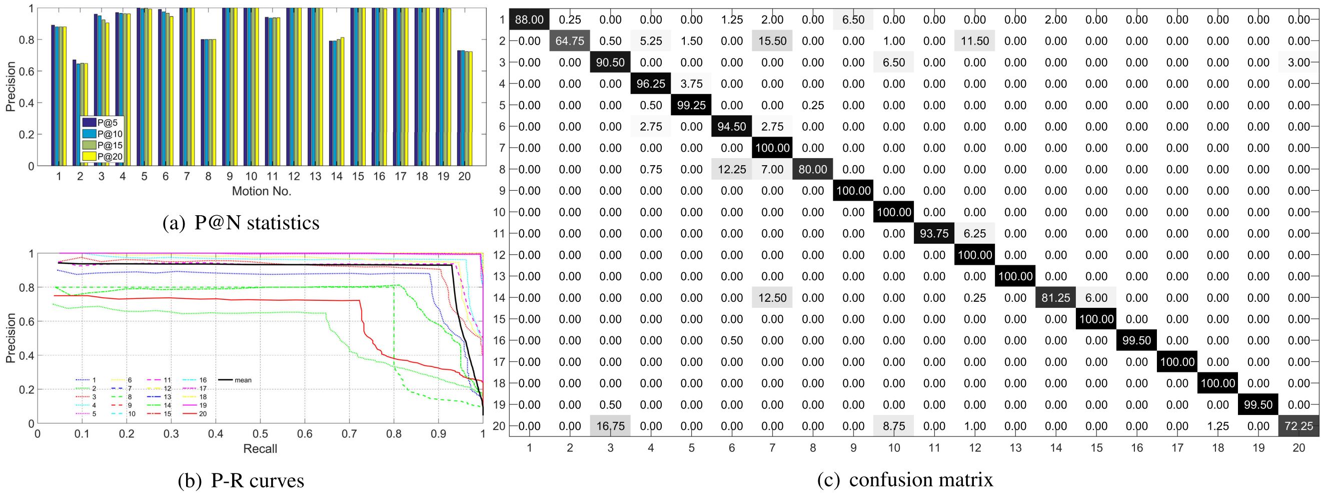
(a) P@N statistics

(b) P-R curves

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	99.75	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	99.75	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.00	0.00	98.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.25	1.50
4	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.00	0.00	97.25	0.00	2.75	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
7	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	93.75	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	6.25
11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	96.25	3.75	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
14	0.00	0.00	0.00	6.75	0.00	0.00	0.00	4.50	0.00	0.00	0.00	0.00	0.00	88.75	0.00	0.00	0.00	0.00	0.00	0.00
15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00
16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	5.00	0.00	0.00	0.00	0.00	2.75	0.00	0.00	92.25	0.00	0.00	0.00	0.00
17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00
18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.25	82.75	15.00	0.00
19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00
20	0.00	0.00	19.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	80.75

(c) confusion matrix

FIGURE 9. **The average P@N (N = 5, 10, 15, 20) statistics, P-R curves and confusion matrix using the proposed motion feature for MoCap-to-MoCap data retrieval.**



(a) P@N statistics

(b) P-R curves

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	88.00	0.25	0.00	0.00	0.00	1.25	2.00	0.00	6.50	0.00	0.00	0.00	0.00	2.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	64.75	0.50	5.25	1.50	0.00	15.50	0.00	0.00	1.00	0.00	11.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.00	0.00	90.50	0.00	0.00	0.00	0.00	0.00	0.00	6.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.00
4	0.00	0.00	0.00	96.25	3.75	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.50	99.25	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	2.75	0.00	94.50	2.75	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
7	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8	0.00	0.00	0.00	0.75	0.00	12.25	7.00	80.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	93.75	6.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
14	0.00	0.00	0.00	0.00	0.00	0.00	12.50	0.00	0.00	0.00	0.00	0.25	0.00	81.25	6.00	0.00	0.00	0.00	0.00	0.00
15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00
16	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	99.50	0.00	0.00	0.00	0.00
17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00
18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00
19	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	99.50	0.00
20	0.00	0.00	16.75	0.00	0.00	0.00	0.00	0.00	0.00	8.75	0.00	1.00	0.00	0.00	0.00	0.00	0.00	1.25	0.00	72.25

(c) confusion matrix

FIGURE 10. **The average P@N (N = 5, 10, 15, 20) statistics, P-R curves and confusion matrix using the proposed motion feature for video-to-video data retrieval.**

experiment, we replace each volunteer's 'normal walking', 'normal running', and 'bending down' clips in our original MoCap dataset by 'walking to running then bending' clips, each formed by concatenating the original 'normal walking', 'normal running' and 'bending down' clips, one instance per each. As before, the motion features of all the updated MoCap clips are then computed and stored offline.

For the first experiment, the average MAP is 0.9425, showing the robustness of our scheme against pure cycle variance. For the second experiment, the average MAPs for 'normal walking' and 'normal running' are as high as 0.9800 and 1.0000, respectively, while the MAP for 'bending down' is just 0.4400. The reason is that 'normal walking' and 'normal running' are similar to each other and together account for the major portion of each 'walking to running then bending' clip. As a result, query walking or running video clips are well matched but bending down clips are not.

In general, since our work focuses on holistic motion retrieval, a sub-motion in a long clip may not be identified especially when it has a short duration. Therefore, for our scheme to get applied for sub-motion retrieval, it is advised to firstly segment long MoCap or video clips to short ones containing elementary motions.

Regarding the feature description of a MoCap clip, the proposed method presumes that the captured subject has an explicitly specified skeletal structure such that a 3D avatar may be constructed and rendered, enabling the extraction of binary silhouettes and the application of MotionSet for feature extraction. As such, the proposed method applies to both full body and half body MoCap clips, as both are well structured. For less organized markers like those on a human face for expression capture, our method is not well suited.

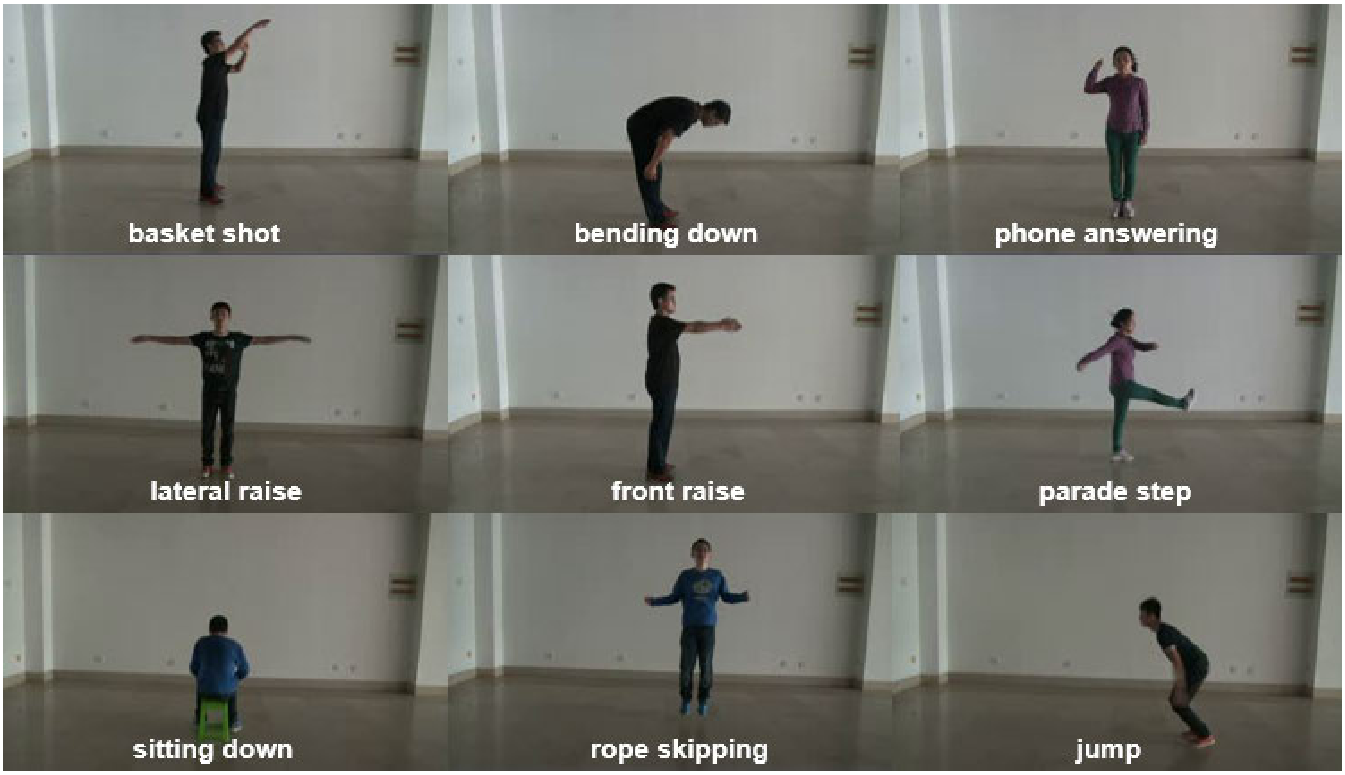


FIGURE 11. Frame samples of various motions and at various viewpoints from the human motion video dataset.

V. CONCLUSION

A novel video-based MoCap data retrieval scheme is proposed in this work. In order to make two different modalities comparable, we propose a systematic approach to extract motion features for both video and MoCap clips, and conduct the retrieval by comparing the video query's motion feature with those of the repository MoCap sequences. The motion feature extraction works in two major steps. Firstly, every motion clip (either a video or a MoCap clip) is converted to a sequence of binary silhouettes. Secondly, regarding a sequence of silhouettes as a set, we use a CNN named MotionSet to extract the discriminative motion feature. Besides the algorithmic solution, we also contribute a MoCap dataset and a video dataset in couple that contain motion clips of 20 action classes. Experiments show that the proposed scheme outperforms the state-of-the-art method of the same type by large margins (*i.e.*, an increase of about 0.25 for average MAP and a reduction to 1/26 for online retrieval time) and extends naturally for video-to-video data retrieval and MoCap-to-MoCap data retrieval.

In this work, we view frames in a motion sequence as a set. Nevertheless, the proposed scheme may not work precisely if the temporal order of frames is crucial to discriminate certain motions. In the future, we will try to incorporate temporal information into the CNN data flow as well to capture both spatial and temporal characteristics of motion clips. Further, the proposed scheme focuses on holistic motion but not sub-motion retrieval. Sub-motion retrieval is planned for our future research as well, which enables more flexible retrieval of long and complex motions.

ACKNOWLEDGMENT

(Tingxin Ren and Wei Li contributed equally to this work.)

REFERENCES

[1] M. Müller, T. Röder, and M. Clausen, "Efficient content-based retrieval of motion capture data," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 677–685, Jul. 2005.

[2] C. Sun, I. Junejo, and H. Foroosh, "Motion retrieval using low-rank subspace decomposition of motion volume," *Comput. Graph. Forum*, vol. 30, no. 7, pp. 1953–1962, Sep. 2011.

[3] F. Liu, Y. Zhuang, F. Wu, and Y. Pan, "3D motion retrieval with motion index tree," *Comput. Vis. Image Understand.*, vol. 92, nos. 2–3, pp. 265–284, Nov. 2003.

[4] Z. Deng, Q. Gu, and Q. Li, "Perceptually consistent example-based human motion retrieval," in *Proc. Symp. Interact. 3D Graph. Games I3D*, 2009, pp. 191–198.

[5] M. Kapadia, I.-K. Chiang, T. Thomas, N. I. Badler, and J. T. Kider, "Efficient motion retrieval in large motion databases," in *Proc. ACM SIGGRAPH Symp. Interact. 3D Graph. Games I3D*, 2013, pp. 19–28.

[6] P. Wang, R. W. H. Lau, Z. Pan, J. Wang, and H. Song, "An eigen-based motion retrieval method for real-time animation," *Comput. Graph.*, vol. 38, pp. 255–267, Feb. 2014.

[7] Q. Xiao, Y. Wang, and H. Wang, "Motion retrieval using weighted graph matching," *Soft Comput.*, vol. 19, no. 1, pp. 133–144, Jan. 2015.

[8] Q. Xiao and W. Yuan, "Motion retrieval based on dynamic Bayesian network and canonical time warping," in *Proc. 8th Int. Symp. Comput. Intell. Design (ISCID)*, Dec. 2015, pp. 267–280.

[9] T. Qi, Y. Feng, J. Xiao, Y. Zhuang, X. Yang, and J. Zhang, "A semantic feature for human motion retrieval," *Comput. Animation Virtual Worlds*, vol. 24, nos. 3–4, pp. 399–407, May 2013.

[10] Z. Wang, Y. Feng, T. Qi, X. Yang, and J. J. Zhang, "Adaptive multi-view feature selection for human motion retrieval," *Signal Process.*, vol. 120, pp. 691–701, Mar. 2016.

[11] N. Lv, Z. Jiang, Y. Huang, X. Meng, G. Meenakshisundaram, and J. Peng, "Generic content-based retrieval of marker-based motion capture data," *IEEE Trans. Vis. Comput. Graphics*, vol. 24, no. 6, pp. 1969–1982, Jun. 2018.

[12] M.-W. Chao, C.-H. Lin, J. Assa, and T.-Y. Lee, "Human motion retrieval from hand-drawn sketch," *IEEE Trans. Vis. Comput. Graphics*, vol. 18, no. 5, pp. 729–740, May 2012.

[13] M. G. Choi, K. Yang, T. Igarashi, J. Mitani, and J. Lee, "Retrieval and visualization of human motion data via stick figures," in *Computer Graphics Forum*, vol. 31, no. 7. Hoboken, NJ, USA: Wiley, 2012, pp. 2057–2065.

[14] J. Xiao, Z. Tang, Y. Feng, and Z. Xiao, "Sketch-based human motion retrieval via selected 2D geometric posture descriptor," *Signal Process.*, vol. 113, pp. 1–8, Aug. 2015.

[15] T.-C. Feng, P. Gunawardane, J. Davis, and B. Jiang, "Motion capture data retrieval using an artist's doll," in *Proc. 19th Int. Conf. Pattern Recognit.*, Dec. 2008, pp. 1–4.

[16] N. Numaguchi, A. Nakazawa, T. Shiratori, and J. K. Hodgins, "A puppet interface for retrieval of motion capture data," in *Proc. ACM SIGGRAPH/Eurograph. Symp. Comput. Animation SCA*, 2011, pp. 157–166.

[17] Z. Jiang, Z. Li, W. Li, X. Li, and J. Peng, "Generic video-based motion capture data retrieval," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Nov. 2019, pp. 1950–1957.

[18] A. Gupta, J. Martinez, J. J. Little, and R. J. Woodham, "3D pose from motion for cross-view action recognition via non-linear circulant temporal encoding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2601–2608.

[19] A. Gupta, J. He, J. Martinez, J. J. Little, and R. J. Woodham, "Efficient video-based retrieval of human motion with flexible alignment," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–9.

[20] S. H. Abdulhussain, A. R. Ramli, B. M. Mahmmod, M. I. Saripan, S. A. R. Al-Haddad, T. Baker, W. N. Flayyih, and W. A. Jassim, "A fast feature extraction algorithm for image and video processing," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–8.

[21] S. H. Abdulhussain, S. A. R. Al-Haddad, M. I. Saripan, B. M. Mahmmod, and A. Hussien, "Fast temporal video segmentation based on krawtchouk-tchebichef moments," *IEEE Access*, vol. 8, pp. 72347–72359, 2020.

[22] A. Sasithradevi and S. Mohamed Mansoor Roomi, "A new pyramidal opponent color-shape model based video shot boundary detection," *J. Vis. Commun. Image Represent.*, vol. 67, Feb. 2020, Art. no. 102754.

[23] N. Lv, Z. Feng, and J. Peng, "Mutual information based video shot boundary detection," in *Proc. Int. Conf. Image Anal. Signal Process.*, Nov. 2012, pp. 1–5.

[24] N. Lv, Y. Huang, Z. Feng, and J. Peng, "A genetic algorithm approach to human motion capture data segmentation," *Comput. Animation Virtual Worlds*, vol. 25, nos. 3–4, pp. 281–290, May 2014.

[25] Z. Jiang, Y. Huang, and J. Peng, "Recent advances in content-based motion capture data retrieval," *Int. J. Electr. Eng.*, vol. 25, no. 2, pp. 47–56, 2018.

[26] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *Int. J. Comput. Vis.*, vol. 103, no. 1, pp. 60–79, May 2013.

[27] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3551–3558.

[28] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei, "Towards 3D human pose estimation in the wild: A weakly-supervised approach," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 398–407.

[29] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, Mar. 2001.

[30] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Proc. 10th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 2, Oct. 2005, pp. 1395–1402.

[31] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action MACH a spatio-temporal maximum average correlation height filter for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

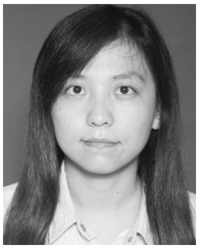
[32] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 2, pp. 316–322, Feb. 2006.

[33] H. Chao, Y. He, J. Zhang, and J. Feng, "GaitSet: Regarding gait as a set for cross-view gait recognition," in *Proc. AAAI*, 2019, pp. 8126–8133.

[34] Y. Fu, Y. Wei, Y. Zhou, H. Shi, G. Huang, X. Wang, Z. Yao, and T. Huang, "Horizontal pyramid matching for person re-identification," *Proc. AAAI Conf. Artif. Intell.*, vol. 33, pp. 8295–8302, Jul. 2019.

[35] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, *arXiv:1703.07737*. [Online]. Available: http://arxiv.org/abs/1703.07737

TINGXIN REN received the bachelor's degree in digital media technology from Shandong University, in 2018, where he is currently pursuing the master's degree with the School of Software. His current research interests include computer vision, image/video analysis, and gait recognition.



WEI LI received the Ph.D. degree in computer science and technology from Shandong University, China, in 2018. She is currently doing her postdoctoral research with the School of Software, Shandong University. Her current research interests include computer vision, image/video analysis, and biometrics.

ZIFEI JIANG received the master's degree in software engineering from Shandong University, China, where he is currently pursuing the Ph.D. degree with the School of Computer Science and Technology. His main research interest includes motion data processing and analysis.

XUEQING LI received the Ph.D. degree in computer science and technology from Shandong University, China. He is currently a Professor with the School of Software, Shandong University.

YAN HUANG received the B.S. degree in computer science from Peking University, in 1997, and the M.S. and Ph.D. degrees in computer science from the University of California at Irvine, in 2003 and 2009, respectively. She is currently an Associate Professor with the School of Software, Shandong University, China. Her research interests include digital geometry processing and image/video analysis and understanding.

JINGLIANG PENG (Member, IEEE) received the B.S. and M.S. degrees in computer science from Peking University, in 1997 and 2000, respectively, and the Ph.D. degree in electrical engineering from the University of Southern California, in 2006. He is currently a Professor with the School of Information Science and Engineering, University of Jinan, China. His research interests include digital geometry processing, virtual/enhanced reality, and image/video analysis.