Generic Video-Based Motion Capture Data Retrieval

Zifei Jiang, Zhen Li, Wei Li, Xueqing Li, Jingliang Peng
School of Software, Shandong University, Shandong, China
E-mails: {jiangzifei, jenslee}@mail.sdu.edu.cn, {wli, xqli, jpeng}@sdu.edu.cn

***Abstract*—In this work we propose a novel and generic scheme for retrieval of motion capture (MoCap) data given a video query. We reconstruct skeleton animations from video clips by a convolutional neural network for 3-dimensional human pose estimation to narrow the gap between videos and MoCap data. A statistical motion signature is computed to extract both morphological and kinematic characteristics from the skeleton animations and the MoCap sequences. This as well ensures that the proposed scheme works on MoCap data with arbitrary skeleton structures. The retrieval is achieved by computing and sorting the distances between the motion signature of the query and those of the MoCap sequences which are pre-computed and stored in the MoCap database. For experimental evaluation, we respectively record a video dataset and capture a MoCap dataset with different performers, and conduct video-based MoCap data retrieval on them. Experimental results demonstrate the effectiveness of the proposed scheme.**

I. Introduction

Three-dimensional (3D) human animations are more and more widely used in virtual and augmented reality, social networking or video processing. A common application scenario is that a user upload a selfie video of a certain action and the system generates a 3D skeleton animation corresponding to the video. One method to solve the problem is estimating 3D pose from each frame and concatenating all the poses into a skeleton animation. However, the quality of the animation is limited by the performance of the estimating algorithm, the frame rate of the video and even the expertise of the actor/actress. A complete pre-built motion capture (MoCap) database which provides high frame rate, high precision and professionally performed 3D skeleton animation [1] to model the motion can meet the demand of the application scenario. Therefore, an effective and efficient retrieval method to find the most similar MoCap sequence in the database to the video is an effective way to achieve the application task.

MoCap techniques have been used to acquire realistic motion sequences since the late 1970s for computer animation and many other purposes [2]. Due to the complex procedure for capturing [3], building a complete MoCap database beforehand is often necessary to practical applications. The existing MoCap databases often use text labels to classify different kinds of MoCap sequences, but text labels are simply a rough classification and are difficult to describe the details of each MoCap sequence. Hence, many content-based MoCap data retrieval methods have been proposed in recent years. Most of the methods use a MoCap sequence as a query, (*e.g.*, [4]), and the query has the same skeleton structure as the MoCap sequences in the database. The similarity is measured by the features extracted from the joints' configurations. Only a few methods focus on other forms of query, such as hand-drawn sketches (*e.g.*, [5], [6]), Kinect-sensed motions (*e.g.*, [7], [8]) and video clips (*e.g.*, [9], [10]). This is primarily because of the representational gap between the MoCap sequences and other forms of motion data. Although good performance has been achieved using MoCap sequences as queries, it is often desirable to use other forms of query that may be easily acquired in a more natural way, one of which is monocular video.

In this work, we propose a generic and easy-to-use scheme for MoCap data retrieval. With the proposed scheme, the user simply acts the motion in front of a video camera, and the proposed scheme automatically retrieves sequences from the MoCap database containing similar motion to the query video. To the best of our knowledge, only few works [9], [10] on video-interfaced MoCap data retrieval have been published. They usually project 3D geometries to 2D planes and use projected 2D images for motion matching, with inevitable loss of information. Compared with them, the proposed scheme distinguishes in that it reconstructs truly 3D motions from 2D videos, characterizes 3D motions using an effective generic statistical descriptor, and matches motions based on the similarity of their descriptors, all of which contribute to the outstanding performance of the proposed scheme.

II. Related Work

In recent years, both human pose estimation and MoCap data retrieval have been considerably studied. In this section, we discuss related works on the two topics respectively.

A. 3D Pose Estimation

In order to reconstruct an accurate skeleton animation, it is crucial to get an precise estimation of the 3D human pose in each frame.

The two-dimensional (2D) pose estimation has been studied for years and the experiment results are remarkably accurate. A lot of research estimates corresponding 3D poses on the basis of the state-of-the-art 2D pose estimations [11], [12]. Zhou *et al.* [13], [14] use a CNN to extract 2D heatmaps from 2D poses to reconstruct a 3D pose sequence from a video clip. There are methods [15], [16], [17] that develop neural networks to find plausible 3D poses from estimated 2D poses. There are also methods [18], [19], [20] that estimate a series of statistical parameters of a 3D human model from 2D images directly. They project different features to a plane

which enable the algorithms to be trained on a database with only 2D ground truth. Although the algorithms achieve decent results, the disadvantages of the schemes are obvious. The reconstruction from 2D to 3D is more like a statistical process and is not supported by any image processing theory. The accuracy of the estimated results are influenced by both 2D and 3D steps, and the latter one brings ambiguity to the scheme because of lacking of the depth information.

Another class of methods adopts the strategy to directly learn the 3D poses from monocular images [21], [22], [23]. Li and Chan [24] use a multi-task framework jointly trains pose regression and body part detectors. Tekin *et al.* [25] train an overcomplete auto-encoder to learn a high-dimensional latent pose representation and account for joint dependencies. These methods are usually trained on fully annotated datasets, which restrict the effectiveness of them on large-scale 2D pose datasets.

Methods [22], [26], [27], [28] have also been proposed to predict 3D poses from images in the wild. Mehta *et al.* [22] adopt a transfer learning method and Mehta *et al.* [26] use kinematic skeleton fitting to achieve real-time 3D pose estimation. Zhou *et al.* [27] propose an end-to-end learning method. The mothed uses mixed 2D and 3D labels in a unified CNN and realize weakly-supervised transfer learning. Yang *et al.* [28] is a complementary to [27] by introducing an adversarial learning framework.

B. MoCap Data Retrieval

Some earlier algorithms (*e.g.*, [29], [30], [31]) directly compute the difference of 3D coordinates or generalized coordinates of the joints to make comparison of postures in motion sequences. Wang and Yeh [32] compute the differences of a set of joint coordinates between two martial art sequences and perform weighted accumulative addition operation to acquire the distance of the two. So and Baciu [33] calculate the change of directions of corresponding body parts between key postures. Miura *et al.* [34] investigate several kinematic parameters of the joints and present an effective parameter combination.

A series of methods use a hierarchical structure to analyse movement of different human body parts. Liu *et al.* [35] propose the searching of MoCap data with a motion index tree. Joint nodes of the body are hierarchically divided into 5 levels from the root (pelvis) to the limbs and head. Key frames of a motion are extracted to build the hierarchical motion index tree of clusters of motions. For a given query, comparisons are made in a hierarchical manner and a comparison between the input motion and motions in the closest leaf cluster finally completes the retrieval. Deng *et al.* [36] divide human body instead of joint nodes into hierarchical meaningful parts. Then each motion is segmented into 18 postures and an adaptive K-means algorithm is performed on each part of the body separately to build pattern index lists for each motion. An extended Knuth-Morris-Pratt (KMP) string match method is used for matching.

Some research extracts mathematical features from the movement of the joints. Tang *et al.* [37] calculate the joint relative distances of any pair of the joints and its symmetric pair and average the distances of each pair in a whole motion sequence with a Boolean characteristic. The weighted averages form a feature vector. Tang and Leung [38] calculate the variance of joint relative distance, which in part reflects motions of joints, to identify the similarity of MoCap data. They construct feature vectors of motion samples and use a linear regression model to get an optimal subset. This kind of single feature description can not completely describe the motion because they neglect the influence of some factors, such as joint rotation and model translation.

Most of the existing MoCap retrieval algorithms focus on the data with the same skeleton structure except Lv *et al.* [39]. They compute statistical motion signatures to extract morphological and kinematic characteristics of heterogeneous MoCap sequences.

To the best of our knowledge, only few works [9], [10] implement video-based MoCap data retrieval. Both works render a MoCap sequence with a roughly approximate "ball-and-cylinder" model. In this way, the MoCap sequence is projected to a group of 2D images with different view directions. Then they adopt some techniques for content-based video retrieval to complete the motion retrieval task. Although the algorithms have achieved good retrieval performance, the procedure of rendering is just an approximation with inevitable loss of information, and the multi-view rendering is time-consuming.

III. Method

A. Overview

In this work we propose a novel scheme for generic video-based MoCap data retrieval. We compute offline the motion signature [39] of each MoCap sequence in the database and store it with the raw MoCap sequence. For a video clip as a given query, we firstly compute the bounding box of the human in the video clip, then we adopt the state-of-the-art architecture [27] to reconstruct the skeleton animation from it. We compute the motion signature of the reconstructed skeleton animation and compare it with all those in the database. A list of MoCap sequences are returned as the result, which are ordered according to their similarities to the query in motion signature, from the highest to the lowest.

The flowchart of the proposed scheme is shown in Fig. 1. Three key components of the flowchart include: 1) reconstructing skeleton animation from the video clip, 2) computing motion signatures, and 3) measuring the distance between different signatures. Details of them are presented in the following sub-sections.

B. Skeleton animation reconstruction

In order to make the videos and the MoCap sequences comparable, a skeleton animation is reconstructed from each video clip which provides a series of body joints positions over time.

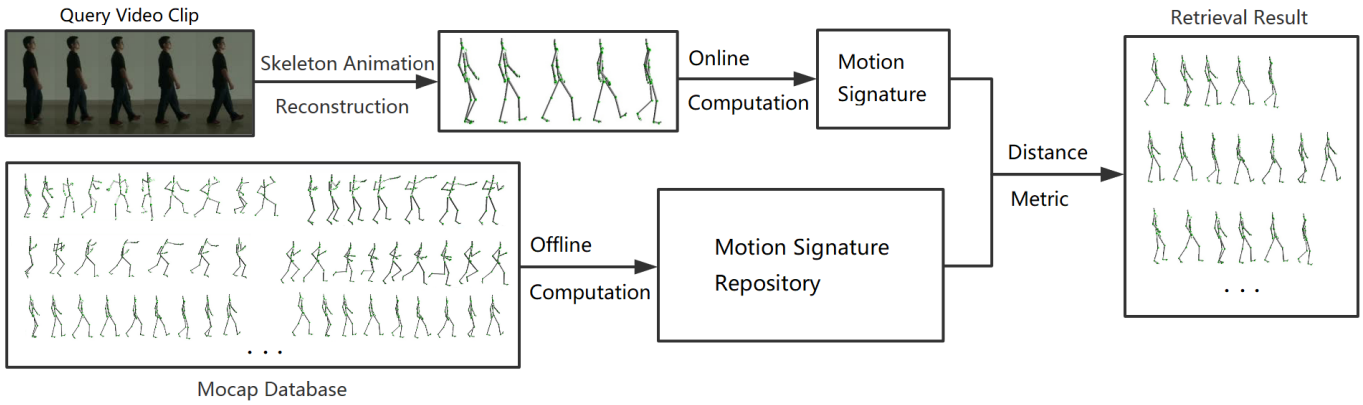


Fig. 1. Flowchart of the proposed method. We reconstruct a skeleton animation from an input video clip. We compute a motion signature for the reconstructed skeleton animation and any MoCap sequence in the database. A subset of the MoCap sequences whose motion signatures are the closest to the querys is returned as the result.

We use a pre-trained Histogram of Oriented Gradient (HOG) + Linear Support Vector Machine (LSVM) model, which is based on the research of Dalal and Triggs [40], to detect human body in the video. Non-maximum suppression is adopted to ensure that only one bounding box is computed in each frame.

The procedure of the reconstruction can be viewed as a 3D pose estimator. A state-of-the-art architecture [27] is adopted to estimate 3D human pose of each frame in a video clip. The architecture includes a 2D pose estimation module and a depth regression module. The 2D pose estimation is a stacked hourglass network [41] in which a repeated bottom-up, top-down structure with intermediate supervision is used to improve the performance. The output of the network is J heat-maps. The peak location of the 2D probability distribution in each map represents a human joint, and J is the number of the joints to be estimated. The depth regression module combines the heat-maps and the intermediate feature representations generated by the 2D module as the input. Unlike [15], [16], [17] which use 2D pose coordinates as the only input, this kind of multi-level input provides more information for 3D pose recovery and avoid the inherent ambiguity to a certain extent. A set of residual modules is used to compute a $J \times 1$ vector as the output which denote the depth of the joints. Besides, weakly supervised learning of the depth regression module on images in the wild is achieved by a 3D geometric constraint induced loss. More detailed information may be found in the reference [27].

In the reconstruction stage, the absolute translation of the subject in video clips has been lost. Therefore, we translate the coordinate origin to the root joint in each frame of the reconstructed animations and the MoCap sequences for consistency.

C. Motion Signature

For the purpose of enhancing the flexibility and the generality of the proposed retrieval scheme, we do not restrict the morphological structure of the both kinds of motion sequences, the reconstructed skeleton animations and the MoCap sequences. Due to the diversity of the MoCap systems, the storage form of the data may be heterogeneous. A part of the data is stored as skeleton animations, and the other part is stored in the coordinates form of the capture markers. Even in the same storage form, the skeletal structure of skeleton animations or the sticking method of the markers may continue to increase the diversity. Most of the previously published algorithms avoid solving the heterogeneous problem because they utilize *a priori* knowledge of a skeletal structure that is consistently defined for all the MoCap sequences in a database. In this work, we adopt a statistical motion signature, which is proposed by Lv *et al.* [39], to extract features from the trajectories of the joints in the 3D Euclidean space without any assumption on the subject's morphological structure.

The motion signature used in this work describes both the morphological and the kinematic characteristics of a motion sequence. We first build a minimal motion spanning tree (MMST) of the joints that extracts the high-level morphological and kinematic characteristics from the motion. With the help of this MMST, then we extract low-level kinematic characteristics between separate joint pairs. A motion signature which describes a motion sequence consists of both high-level and low-level characteristics.

1) Minimal Motion Spanning Tree: We hope to connect the joints in a automatical way to resemble the morphological structure, which is similar to the *a priori* knowledge of the skeletal structure, and also reflects the overall kinematic characteristics of the motion. Specifically, a complete motion graph of the joints is built where each node corresponds to a joint, and each edge is weighted by the standard deviation of the spatial distance between two joints labeled over all the sequence. A sub-graph is the extracted from the complete map, which is explained as below. It is noteworthy that the nodes in the motion graph do not specify geometric attributes, and the motion graph is independent of the view.

In order to make the extracted sub-graph reflect the morphological structure of the subjects, the joints attached to the same rigid segment (for example, the skeleton in the

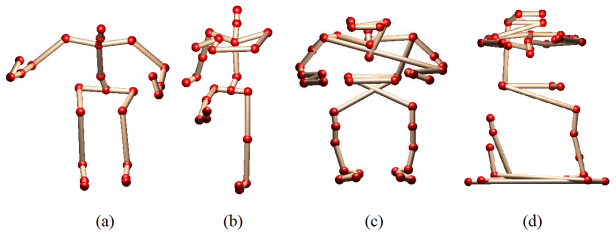


Fig. 2. Instantiated MMST examples for the (a) jump, (b) goose-step motions for joint based MoCap data, (c) jump and (d) squat motions for marker based MoCap data. The edge weights are ignored in the figure and the nodes in each MMST are positioned by the geometry of the subject in a certain frame.

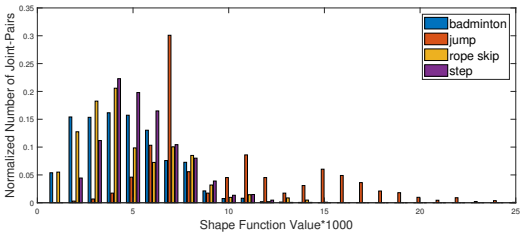


Fig. 3. Shape distribution histograms for MMSTs of 4 different types of motions.

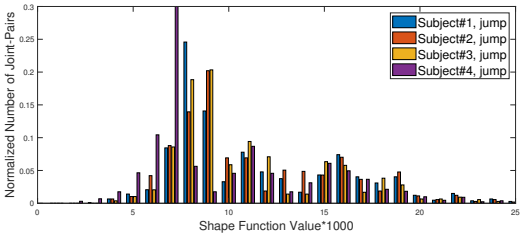


Fig. 4. Shape distribution histograms for MMSTs of the same type (*i.e.*, jump) of motions.

human skeleton) are kept connected. Joints on the same bone usually show less relative motion than those on different ones. Therefore, a minimal spanning tree is extracted from the motion graph, called MMST. The MMST also reflects the overall kinematic characteristics of motion sequence, because it roughly represents the configuration of kinematic correlation between joints, *i.e.*, each highly correlated joint pair is connected with one edge, while the less correlated joint pair is not.

It is noteworthy that the MMST is established to represent the high-level morphological and kinematic characteristics of the motion subject through the joint connection rather than the accurate morphological structure. When the skeleton structure is the same, MMSTS can differentiate movement to some extent. Given that, even if a skeletal structure has been provided with the original motion sequence, it won't be used as the basis of the motion description since the skeletal structure itself is not motion-discriminative and, further, semantic information about the joints may be lacking in the raw data.

In Fig. 2, we show exemplar MMSTs for motion sequences. For the convenience of illustration, the edge weights are ignored and the nodes in each MMST are positioned by the geometry of the subject in a certain frame (note that we do not utilize any geometric attributes in original data). As shown in Fig. 2(a) and (b), the MMSTs generated from joint based MoCap data for the jump and the goose-step motions topologically resemble the subject's skeletal structure quite well as the distribution of the joint is scattered and the subject fully exercises all the joints in these motions. We also show the effect of MMST on marker based MoCap data as (c) and (d), the MMSTs for the jump and the squat motions. The MMST also do well in (c), but do less well in (d) due to little relative motions among the markers on the shanks.

2) Motion Signature Composition: We construct a motion signature to describe both the high-level and the low-level morphological and kinematic characteristics of a MoCap sequence. Specifically, the high-level description is made by the description of the MMST and the low-level description is made by the description of the kinematic features between every marker pair.

Inspired by the shape distribution descriptors for 3D shape analysis [42], [43], we define a shape function of joint pairs on the graph and use the distribution of shape function values to statistically describe the characteristics of the MMST. For a pair of joints, j_i and j_j, $l_{i,j}$ denote the length of the shortest path between them (*i.e.*, the sum of edge weights on that path) in the MMST and L denote the sum of the edge weights in the whole MMST. To normalize the shortest distance and make the shape function scale invariant, we divide the length by L and define the corresponding shape function value between j_i and j_j as $s(j_i, j_j) = l_{i,j}/L$. Assuming that there are totally J joints, we compute the shape function values for all the $T = C(J, 2)$ joint pairs. The histogram of these shape function values constitutes the shape distribution descriptor of the MMST. Or, equivalently, we use an array of shape function values on all T joint pairs as the shape descriptor for the MMST.

We plot Fig. 3 and Fig. 4 to show the capability of the shape distribution histogram in distinguishing different motions. From Fig. 3, we can easily find that the histograms for different motion types shows low similarity, from Fig. 4, we find that histograms for same kind of motions shows high similarity.

However, on the basis of experiments, the shape distribution histogram itself does not achieve sufficient precision to classify the motion types. Therefore, more dimensions are needed to enhance the description and distinguishing capabilities of motion descriptors, as described below.

A set of quantities is used to describe the absolute and relative motion characteristics of each joint pair. The velocity (v_1) and acceleration (a_1) of centroid of joint pair in each frame are measured for describing the absolute motion, and the Euclidean distance (d), relative velocity (v_2), relative acceleration (a_2), relative angular velocity (v_3), and relative angular acceleration (a_3) between the pair of joints at each

frame are measured for describing the relative motion. The motion information of each joint pair in a F-frame motion sequence is then converted into seven curves: $v_1(t)$, $a_1(t)$, $d(t)$, $v_2(t)$, $a_2(t)$, $v_3(t)$ and $a_3(t)$, $t = 1, 2, \ldots, F$. The first three statistical moments are used to describe each curve, the arithmetic square root of the variance and the cube root of the skewness. By doing this, twenty-one quantities are extracted to describe the kinematic characteristics of the joint pair. These quantities are concatenated as the kinematic characteristics of the joint pair and represented by $k_i, i \in [1, 21]$.

Finally, the computation of the motion signature is achieved by combining the high-level and the low-level features of each joint pair and placing all the joint pairs' feature vectors together to form a 2D matrix.

Suppose there are a total of T joint pairs combined. For the p-th ($1 \leq p \leq T$) joint pair, s_p represent its shape function value and $k_{p,i}, 1 \leq i \leq 21$ the kinematic characteristics. The feature vector, $\mathbf{f}_p$, is defined as $\mathbf{f}_p = (s_p, k_{p,1}, k_{p,2}, \ldots, k_{p,21})$. The motion signature of the MoCap sequence is finally formed by putting the feature vectors of all T joint pairs together in ascending order of their first component (*i.e.*, shape function values).

D. Distance Metric

We compute and sort the distances between the query's motion signature and those in the database for retrieval. To deal with dimensionality inconsistency between motion signatures, we exploit the shape function values of the feature vectors to register two motion signatures.

1) Registration: Since the joint number is not restrict as a fixed value, the motion signatures may contains different numbers of feature vectors. For each feature vector in one motion signature, we need to find its corresponding one in the other motion signature. If the two motion sequences are similar, they will have similar MMSTs and corresponding joint pairs will have similar shape function values as well. Therefore, for a feature vector with shape function value s in one motion signature, we search for its correspondence from the feature vectors whose shape function value is the closest to s in the other motion signature. Since two motion sequences may be highly dissimilar in subject structure and/or motion type, there may be very few direct correspondences between their joint pairs. Therefore, a flexible correspondence rather than the best match correspondence is needed. We set up a window to achieve the match, and details are described in Sec. III-D2.

2) Distance Metric: Assume that there are two motion signatures, $\mathbf{S} = [\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_P]$ and $\mathbf{S}' = [\mathbf{f}'_1, \mathbf{f}'_2, \ldots, \mathbf{f}'_{P'}]$, where $\mathbf{f}_i$, $i \in [1, P]$ and $\mathbf{f}'_j$, $j \in [1, P']$ are 22-dimensional feature vectors and P and P' are the numbers of marker pairs in the corresponding motion sequences, respectively. We compute the symmetric distance between $\mathbf{S}$ and $\mathbf{S}'$ according to Eq. 1

$$D(\mathbf{S}, \mathbf{S}') = \sum_{l=1}^{22} \mathbf{W}[l] \times D_F(\mathbf{S}, \mathbf{S}', l) \tag{1}$$

where $\mathbf{W}$ is the weighting vector which contains the weight for each separate feature and $D_F(\mathbf{S}, \mathbf{S}', l)$, the symmetric distance between $\mathbf{S}$ and $\mathbf{S}'$ on the l-th feature, is defined as Eq. 2

$$D_F(\mathbf{S}, \mathbf{S}', l) = \begin{cases} D_H(h, h'), & l = 1 \\ \frac{1}{P} \sum_{i=1}^{P} |\mathbf{f}_i[l] - \mathbf{f'}_{c(i)}[l]| & \\ + \frac{1}{P'} \sum_{j=1}^{P'} |\mathbf{f'}_j[l] - \mathbf{f}_{c'(j)}[l]|, & l \in [2, 22] \end{cases} \tag{2}$$

In Eq. 2, h and h' are the normalized shape distribution histograms obtained from the shape function values in $\mathbf{S}$ and $\mathbf{S}'$, respectively, and $D_H(h, h')$ denotes the difference between h and h' with the reciprocal of their intersection. It should be noted that we compute the normalized shape distribution histogram for each motion sequence in the database just once and store it for later use. In Eq. 2, $\mathbf{f}'_{c(i)}$ (*resp.* $\mathbf{f}_{c'(j)}$) is the corresponding feature vector of $\mathbf{f}_i$ (*resp.* $\mathbf{f}'_j$) with $c(i)$ and $c'(j)$ defined as

$$\begin{aligned} c(i) &= \underset{j \in [j_i - w, j_i + w]}{argmin} \sum_{l=1}^{22} \mathbf{W}[l] \times |\mathbf{f}_i[l] - \mathbf{f'}_j[l]|, \\ c'(j) &= \underset{i \in [i_j - w, i_j + w]}{argmin} \sum_{l=1}^{22} \mathbf{W}[l] \times |\mathbf{f'}_j[l] - \mathbf{f}_i[l]| \end{aligned} \tag{3}$$

where j_i (*resp.* i_j) indexes the feature vector in $\mathbf{S}'$ (*resp.* $\mathbf{S}$) with the closest shape function value to $\mathbf{f}_i$ (*resp.* $\mathbf{f}'_j$). As formulated in Eq. 3, we set up a window, $[-w, w]$, around the closet feature vector to pick the best matching one to increase robustness of the scheme. The empirical parameter $w = 7$ is used in our scheme.

IV. Experiments

In this section, we conduct comprehensive experiments to evaluate the performance of our proposed generic video-based MoCap data retrieval algorithm. Specifically, we set one experiment to evaluate the performance of the proposed scheme for video-based MoCap data retrieval. Besides, we set another experiment to evaluate the performance of the proposed motion signature.

A. Databases and Performance Metrics

We exploit a human MoCap database with 12 daily action classes [10] and the detailed action classes can be found in Tab. I. Specifically, there are 240 clips in the MoCap database which are captured by five actors with various body shapes. Each actor or actress performs each motion 4 times, resulting in a total number of 240 clips in the MoCap database.

For the query video database, we also exploit the same 12 daily action classes. Three males and one female of various body shapes are employed to shoot these videos by a monocular camera. The videos are shot at four viewpoints (*i.e.*, Front, Back, Left and Right), and each class of the video query database consists of 20 video clips.

TABLE I
MAP STATISTICS OF NDVP [10] AND OUR PROPOSED METHOD.

Motion	MAP statistics	
(#) name	NDVP [10]	Ours
(1) phone	0.5292	09437
(2) jump	0.5019	0.8924
(3) punch	0.6684	0.5337
(4) bounce	0.6332	0.7892
(5) arm raise	0.5141	0.6565
(6) round walk	0.4723	0.5853
(7) side walk	0.4471	0.6849
(8) rope skip	0.2730	0.5634
(9) shoot	0.9861	0.8140
(10) badminton	0.2945	0.6412
(11) goose-step	0.7202	0.8894
(12) sit-down	0.7794	0.7263
average MAP	0.5683	0.7267

In this work, mean average precision (MAP), precision-recall curve (P-R curve) and precision at n (P@n) are exploited to evaluate the performance of the proposed video-based MoCap data retrieval algorithm as often used in the general field of information retrieval.

For each query video clip, the fraction of relevant samples in the result set gives the precision, while the fraction of all relevant samples that has been returned gives the recall. If the result set has a size of n, the precision gives P@n. By varying n, we can obtain a P-R curve of this query. When $n = N$ with N being the size of the database, the average precision, AP, of this query can be computed by Eq. 4

$$AP = \frac{1}{R} \sum_{j=1}^{N} I_j \times \frac{R_j}{j} \tag{4}$$

where R is the number of relevant samples in the database, $I_j = 1$ if the jth ranking sample of the result set is relevant and $I_j = 0$ otherwise, and R_j is the number of relevant samples in the j top-ranking samples of the result set. P@n, P-R and MAP statistics for that class can be obtained by averaging the statistics of each query in a motion class.

B. Comparison with benchmark method

We compare our method with the NDVP method in work [10]. Note that we do not compare with the work [9] as it focuses on extracting similar (to the video query) sub-MoCap-sequences through effective but time-consuming frame-to-frame alignment, while our work targets at quick search of similar whole MoCap clips based on overall motion characteristics.

MAP statistics of the two methods are presented in Tab. I for comparison. From Tab. I we can observe that the proposed scheme outperforms NDVP [10] on most of motion classes.

In Fig. 5 and Fig. 6, we plot respectively for both methods its average P@n (n=5, 10, 15, 20) statistics, P-R curves and confusion matrix over all the action classes. NDVP shows excellent retrieval performance on shoot motion, but the overall performance is unstable. The figures again shows better performance of our method over NDVP [10].

C. Gap between query modalities

In this part, we examine the performance gap between two modalities of query: video clip and MoCap sequence. Specifically, we use MoCap sequences as queries, measure the corresponding retrieval performance, and compare it with that of the proposed video-based MoCap data retrieval as reported in Section IV-B. For this experiment, we get an MAP of 88.53%, and show P@n (n=5, 10, 15, 20) statistics, P-R curves and confusion matrix over all the action classes in Fig. 7. By comparison, we observe that there still exists a performance gap between the two query modalities. This is mainly due to the challenge in precise 3D skeleton animation reconstruction from a monocular 2D video clip.

V. CONCLUSIONS

We propose a novel generic video-based MoCap data retrieval scheme in this work. A CNN based 3D pose estimation approach is adopted to reconstruct skeleton animations from query video clips, which narrows the gap between these two data modalities: 2D video clip and 3D motion sequence. A statistical motion signature which consists of both high-level and low-level characteristics is computed for effective motion matching. The proposed scheme does not utilize *a priori* knowledge of skeletal structure and works on MoCap data in arbitrary skeleton structure. Experimental results demonstrate the promising performance of the proposed scheme.

Nevertheless, more precise 3D animation reconstruction methods are demanded to further reduce the performance gap between the two modalities of query.

ACKNOWLEDGMENT

This work is partially funded by the National Natural Science Foundation of China (Grants No. 61872398). The authors also thank Tingxin Ren for the help in running the 3D pose estimation code. Xueqing Li and Jingliang Peng are the corresponding authors.

REFERENCES

[1] F. W. Da Silva, L. Velho, P. R. Cavalcanti, and J. Gomes, "An architecture for motion capture based animation," in *Proceedings X Brazilian Symposium on Computer Graphics and Image Processing*, pp. 49–56, IEEE, 1997.

[2] N. Lv, Y. Huang, Z. Feng, and J. Peng, "A survey on motion capture data retrieval," *Applied Mechanics and Materials*, vol. 556-562, pp. 2944–2947, 2014.

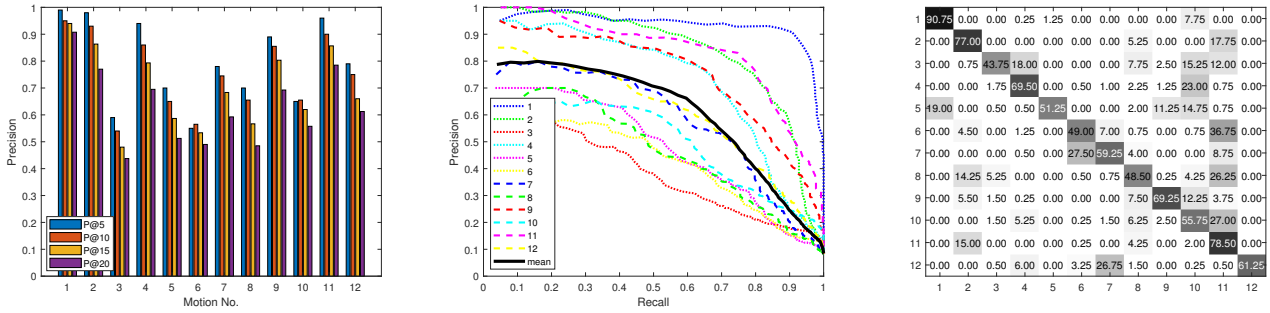
[3] Z. Jiang, Y. Huang, and J. Peng, "Recent advances in content-based motion capture data retrieval," *International Journal of Electrical Engineering*, vol. 25, no. 2, pp. 47–56, 2018.

[4] M. Müller, T. Röder, and M. Clausen, "Efficient content-based retrieval of motion capture data," in *ACM Transactions on Graphics (ToG)*, vol. 24, pp. 677–685, ACM, 2005.

[5] Q. L. Li, W. D. Geng, T. Yu, X. J. Shen, N. Lau, and G. Yu, "Motionmaster: authoring and choreographing kung-fu motions by sketch drawings," in *Proceedings of the 2006 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pp. 233–241, Eurographics Association, 2006.

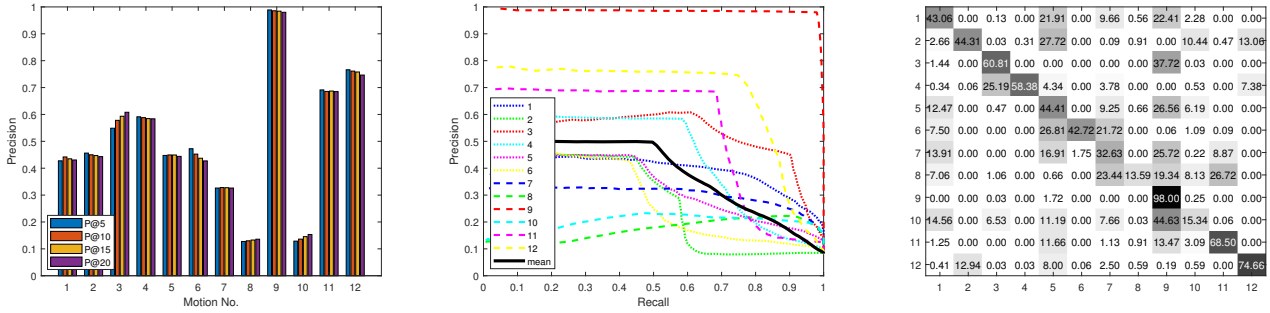
[6] M.-W. Chao, C.-H. Lin, J. Assa, and T.-Y. Lee, "Human motion retrieval from hand-drawn sketch," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 5, pp. 729–740, 2011.

[7] M. Kapadia, I.-k. Chiang, T. Thomas, N. I. Badler, J. T. Kider Jr, *et al.*, "Efficient motion retrieval in large motion databases," in *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, pp. 19–28, ACM, 2013.



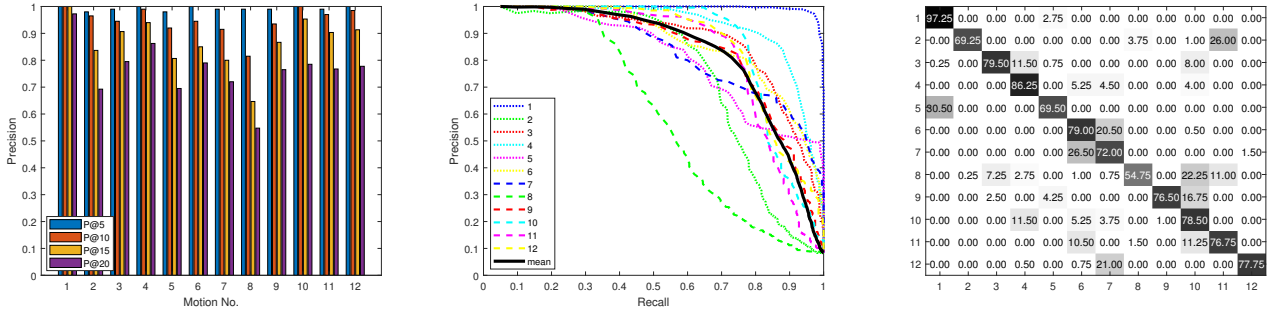
P@N P-R curves confusion matrix

Fig. 5. P@n statistics, P-R curves and confusion matrix of our scheme for video-based MoCap retrieval.



P@N P-R curves confusion matrix

Fig. 6. P@n statistics, P-R curves and confusion matrix of NDVP for video-based MoCap retrieval.



P@N P-R curves confusion matrix

Fig. 7. P@n statistics, P-R curves and confusion matrix of our scheme for MoCap retrieval.

[8] E. C.-H. Lin, "A research on 3d motion database management and query system based on kinect," in *Future Information Technology-II*, pp. 29–35, Springer, 2015.

[9] A. Gupta, J. He, J. Martinez, J. J. Little, and R. J. Woodham, "Efficient video-based retrieval of human motion with flexible alignment," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–9, IEEE, 2016.

[10] W. Li, Y. Huang, C.-C. J. Kuo, J. Peng, *et al.*, "Video-based human motion capture data retrieval via normalized motion energy image subspace projections," in *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 243–248, IEEE, 2017.

[11] C.-H. Chen and D. Ramanan, "3d human pose estimation= 2d pose estimation+ matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7035–7043, 2017.

[12] E. Jahangiri and A. L. Yuille, "Generating multiple diverse hypotheses for human 3d pose consistent with 2d joint detections," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 805–814, 2017.

[13] X. Zhou, M. Zhu, S. Leonardos, K. G. Derpanis, and K. Daniilidis, "Sparseness meets deepness: 3d human pose estimation from monocular video," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4966–4975, 2016.

[14] X. Zhou, M. Zhu, G. Pavlakos, S. Leonardos, K. G. Derpanis, and K. Daniilidis, "Monocap: Monocular human motion capture using a cnn coupled with a geometric prior," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 4, pp. 901–914, 2018.

[15] J. Wu, T. Xue, J. J. Lim, Y. Tian, J. B. Tenenbaum, A. Torralba, and W. T. Freeman, "Single image 3d interpreter network," in *European Conference on Computer Vision*, pp. 365–382, Springer, 2016.

[16] F. Moreno-Noguer, "3d human pose estimation from a single image via distance matrix regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2823–2832, 2017.

[17] D. Tome, C. Russell, and L. Agapito, "Lifting from the deep: Convolutional 3d pose estimation from a single image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2500–2509, 2017.

[18] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it smpl: Automatic estimation of 3d human pose and shape from a single image," in *European Conference on Computer Vision*, pp. 561–578, Springer, 2016.

[19] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler, "Unite the people: Closing the loop between 3d and 2d human representations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6050–6059, 2017.

[20] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7122–7131, 2018.

[21] B. Tekin, P. Márquez-Neila, M. Salzmann, and P. Fua, "Learning to fuse 2d and 3d image cues for monocular body pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3941–3950, 2017.

[22] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt, "Monocular 3d human pose estimation in the wild using improved cnn supervision," in *2017 International Conference on 3D Vision (3DV)*, pp. 506–516, IEEE, 2017.

[23] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, "Coarse-to-fine volumetric prediction for single-image 3d human pose," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7025–7034, 2017.

[24] S. Li and A. B. Chan, "3d human pose estimation from monocular images with deep convolutional neural network," in *Asian Conference on Computer Vision*, pp. 332–347, Springer, 2014.

[25] B. Tekin, I. Katircioglu, M. Salzmann, V. Lepetit, and P. Fua, "Structured prediction of 3d human pose with deep neural networks," *arXiv preprint arXiv:1605.05180*, 2016.

[26] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt, "Vnect: Real-time 3d human pose estimation with a single rgb camera," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, p. 44, 2017.

[27] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei, "Towards 3d human pose estimation in the wild: a weakly-supervised approach," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 398–407, 2017.

[28] W. Yang, W. Ouyang, X. Wang, J. Ren, H. Li, and X. Wang, "3d human pose estimation in the wild by adversarial learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5255–5264, 2018.

[29] L. Kovar, M. Gleicher, and F. Pighin, "Motion graphs," in *ACM SIGGRAPH 2008 classes*, p. 51, ACM, 2008.

[30] J. Lee, J. Chai, P. S. Reitsma, J. K. Hodgins, and N. S. Pollard, "Interactive control of avatars animated with human motion data," in *ACM Transactions on Graphics (ToG)*, vol. 21, pp. 491–500, ACM, 2002.

[31] O. Arikan and D. A. Forsyth, "Interactive motion generation from examples," in *ACM Transactions on Graphics (TOG)*, vol. 21, pp. 483–490, ACM, 2002.

[32] C.-S. Wang, "3d motion retrieval for martial arts," *Tamsui Oxford Journal of Mathematical Sciences*, vol. 20, no. 2, pp. 327–337, 2004.

[33] C. K. So and G. Baciu, "Entropy-based motion extraction for motion capture animation," *Computer Animation and Virtual Worlds*, vol. 16, no. 3-4, pp. 225–235, 2005.

[34] K. Miura, H. Furukawa, and M. Shoji, "Similarity of human motion: congruity between perception and data," in *2006 IEEE International Conference on Systems, Man and Cybernetics*, vol. 2, pp. 1184–1189, IEEE, 2006.

[35] F. Liu, Y. Zhuang, F. Wu, and Y. Pan, "3d motion retrieval with motion index tree," *Computer Vision and Image Understanding*, vol. 92, no. 2-3, pp. 265–284, 2003.

[36] Z. Deng, Q. Gu, and Q. Li, "Perceptually consistent example-based human motion retrieval," in *Proceedings of the 2009 symposium on Interactive 3D graphics and games*, pp. 191–198, ACM, 2009.

[37] J. K. Tang, H. Leung, T. Komura, and H. P. Shum, "Emulating human perception of motion similarity," *Computer Animation and Virtual Worlds*, vol. 19, no. 3-4, pp. 211–221, 2008.

[38] J. K. Tang and H. Leung, "Retrieval of logically relevant 3d human motions by adaptive feature selection with graded relevance feedback," *Pattern Recognition Letters*, vol. 33, no. 4, pp. 420–430, 2012.

[39] N. Lv, Z. Jiang, Y. Huang, X. Meng, G. Meenakshisundaram, and J. Peng, "Generic content-based retrieval of marker-based motion capture data," *IEEE transactions on visualization and computer graphics*, vol. 24, no. 6, pp. 1969–1982, 2017.

[40] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *international Conference on computer vision & Pattern Recognition (CVPR'05)*, vol. 1, pp. 886–893, IEEE Computer Society, 2005.

[41] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European Conference on Computer Vision*, pp. 483–499, Springer, 2016.

[42] M. Ankerst, G. Kastenmüller, H.-P. Kriegel, and T. Seidl, "3d shape histograms for similarity search and classification in spatial databases," in *International Symposium on Spatial Databases*, pp. 207–226, Springer, 1999.

[43] R. Osada, T. Funkhouser, B. Chazelle, and D. Dobkin, "Shape distributions," *ACM Transactions on Graphics (TOG)*, vol. 21, no. 4, pp. 807–832, 2002.