Modeling Spatial Relations of Human Body Parts for Indexing and Retrieving Close Character Interactions

Edmond S. L. Ho*[1,2], Jacky C. P. Chan[1], Yiu-ming Cheung[1], and Pong C. Yuen[1]

[1]Department of Computer Science, Hong Kong Baptist University
[2]Science Faculty, HKBU Institute of Research and Continuing Education

Abstract

Retrieving pre-captured human motion for analyzing and synthesizing virtual character movement have been widely used in Virtual Reality (VR) and interactive computer graphics applications. In this paper, we propose a new human pose representation, called *Spatial Relations of Human Body Parts (SRBP)*, to represent spatial relations between body parts of the subject(s), which intuitively describes how much the body parts are interacting with each other. Since *SRBP* is computed from the local structure (i.e. multiple body parts in proximity) of the pose instead of the information from individual or pairwise joints as in previous approaches, the new representation is robust to minor variations of individual joint location. Experimental results show that *SRBP* outperforms the existing skeleton-based motion retrieval and classification approaches on benchmark databases.

CR Categories: I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Virtual reality, Animation I.3.8 [Computer Graphics]: Applications—;

Keywords: Close Interaction, Motion Retrieval, Motion Classification, Human Motion, Spatial Relations

1 Introduction

Human motion data have been used in a wide range of VR applications, such as virtual training [Pronost et al. 2008; Kyan et al. 2015] and virtual rehabilitation [Celiktutan et al. 2013], for analyzing the performance of the human subject as well as animating virtual avatars to interact with the subject to enhance the realism of the system. In particular, pre-captured human motions can be used as examples to guide the movement of virtual characters in response to the performance of the human subject [Ho et al. 2013; Pronost et al. 2008]. However, when retrieving relevant examples from the motion database, maintaining the temporal coherency of the poses over successive frames is a crucial factor in producing realistic movement of the virtual characters. One of the fundamental problems in existing methods in retrieving character motion is the use of low-level representations such as 3D joint positions to represent human poses. Subtle movements of the subject or variations in the pose may result in significant changes in the representation and easily results in interpenetrations of body parts when handling close character interactions. While detecting and resolving the collisions between characters can be a solution, additional computation cost is required which may not be available in real-time applications.

*edmond@comp.hkbu.edu.hk

In this paper, we propose a new spatial relation-based approach to tackle the aforementioned difficulties by taking the advantages of modeling the relationships of body parts to represent human motions. In particular, human motions are represented by *Spatial Relations of Human Body Parts (SRBP)* which encodes the local relationship of body parts in proximity to represent motions in different classes. *SRBP* is intuitive and easy to interpret - the larger the *SRBP* value, the more the body parts interact. In addition, poses captured from subjects with different body sizes can be represented and compared using the same *SRBP* representation. Experimental results show that our proposed method outperforms the state-of-the-art skeleton-based approaches on classifying and retrieving motions on benchmark databases.

2 Related work

An early work proposed by Kovar et al. [Kovar et al. 2002] compares the similarity of human skeletal poses by calculating the Euclidean distance between the point clouds sampled from 3D motion data. However, logically similar postures are not necessarily having similar 3D joint configurations which limit the discriminative power of the proposed method. Yun et al. [Yun et al. 2012] compared the performance of two-character interaction classification using commonly used spatial and spatio-temporal features such as geometric relational, velocities and logical features. Müller et al. [Müller et al. 2005] proposed a semantic approach based on the correlation of four joint positions for content-based human pose retrieval. Different combinations of joints are used to index and retrieve 3D human postures effectively. However, this requires the user to manually specify the combination that is suitable for each action class. Ho and Komura [Ho and Komura 2009] proposed to use *tangle* in knot theory for indexing and retrieving two-character close interactions (such as dancing and wrestling). However, the method becomes less effective when the body parts of the characters are not entangled. Tang et al. [Tang et al. 2012] proposed Spacetime Proximity Graphs to extract spatio-temporal relations between the body parts in two-character interactions such as dancing and fighting. While the representation proposed in this paper share a similar idea in extracting spatial relations of body parts from 3D point cloud using Delaunay Tetrahedralization, the spatial relation-based features are computed differently and we will show the advantages of using our method over [Tang et al. 2012] in Section 5.2.

3 Overview

This section gives an overview of the proposed approach for representing human motion with close interactions. Given a motion in the form of a sequence of frames which contain 3D human skeletal data, two types of spatial relations-based features are extracted (Section 4). Firstly, Relative Position features *(RP)* are extracted (Section 4.1) to represent the underlying skeletal structure of the subject. Next, the proposed method for extracting *Spatial Relations of Human Body Parts (SRBP)* will be explained in Section 4.2. Specifically, feature points are extracted from the human skeleton

in each frame for analyzing the spatial relations. We then compute *SRBP* based on the interaction between every pair of body parts. Once the frames in each motion are represented by *RP* and *SRBP*, temporal alignment is applied to the motions to remove the temporal variations to facilitate motion comparison. Without loss of generality, we will first explain our proposed method using an example of a single pose of a subject in Sections 4.1-4.2 and present how to extend our method to represent two-character interactions in Section 4.3.

4 Spatial Relation-based Representation

4.1 Relative Position Features (RP)

The Relative Position features *(RP)* represent the spatial relations between joints in the underlying skeletal structure of the subject. Using Relative Position features to model the spatial relations between human body parts and object has been proposed in skeleton-based human-object interaction recognition (e.g. [Delaitre et al. 2011]). Given the skeletal pose in each frame, the translation and rotation around the vertical axis of the root joint (i.e. pelvis) are removed as normalization to facilitate pose comparison in the later stages. Next, the *RP* feature of every pair of connected joints can be calculated by:

$$RP_i = \frac{v_i - v_{i,parent}}{|v_i - v_{i,parent}|} \tag{1}$$

where v_i and $v_{i,parent}$ are the 3D positions of the i-th joint and the parent of the i-th joint. The relative positions of the connected joints are normalized (as in Eq. 1) to facilitate the comparison of motions performed by subjects with different body sizes (i.e. bone lengths).

4.2 Spatial Relations of Human Body Parts (SRBP)

4.2.1 Feature Points Sampling

The first step is to compute feature points from the joint positions in Cartesian coordinates in each pose. While directly using joint positions as feature points have been widely used in existing spatial relation-based representations [Ho et al. 2010; Tang et al. 2012], the sparsely and unevenly distributed joint positions may results in significant change in spatial relations with subtle body movements. For this reason, our proposed method samples feature points at a high resolution by uniformly upsampling points from each body part formed using the joint positions on the two ends. Sampling feature points in high resolution can represent the topology of the body parts, which is an abstraction of the shape of the body parts and such information will be used for analyzing how the two body parts interact.

4.2.2 Extraction of Spatial Relations

When analyzing the spatial relations between the body parts of human(s), one of the criteria is to evaluate the spatial distances between them. In most of the human motions, the spatial relations between body parts in proximity represent the characteristics of the motions (e.g. the *Right Hand-Left Hand* pair in Hand Clap motion) since body parts in close distance tend to be more influential to each other. Therefore, a reasonable way to extract *important* spatial relations is to analyze the structure of the feature points in proximity. In this work, we propose to apply Delaunay Tetrahedralization to the sampled feature points to construct a volumetric mesh to represent the proximity information:

$$M = DelaunayTetrahedralization(\mathbf{V}) \tag{2}$$

where $\mathbf{V}$ contains the 3D positions of all sampled feature points and M is the volumetric mesh computed from $\mathbf{V}$. M contains the connectivity information of the feature points. Since the sampled points in proximity tend to be connected by edges in Delaunay Tetrahedralization, connectivity of the mesh can be used to analyze which entities are interacting with each other. The effectiveness of using the connectivity of the sampled points to extract the spatial relations of body parts for motion analysis and synthesis has been demonstrated in [Ho et al. 2010; Tang et al. 2012].

4.2.3 The new representation - SRBP

In this subsection, the calculation of *SRBP* is explained. The *SRBP* indicates *how much the body parts are interacting with each other*. We compute *SRBP* based on the connectivity of the volumetric mesh constructed using Eq. 2. Since the volumetric mesh is constructed by performing Delaunay Tetrahedralization on the whole point cloud (i.e. 3D positions of all feature points), the connectivity of the mesh is computed based on the global structure and distribution of the feature points. By this, *SRBP* will be mainly affected by the global structure instead of variations of individual feature points to robustly represent the human pose. For each pair of body parts, the *SRBP* which is a scalar value can be calculated by:

$$SRBPi,j = \frac{\sum_{a=1}^{n_i}\sum_{b=1}^{n_j} conn(M, v_{i,a}, v_{j,b})}{n_i + n_j} \tag{3}$$

where $SRBP_{i,j}$ is the *SRBP* of the body parts BP_i and BP_j, $v_{i,a}$ and $v_{j,b}$ are the a-th and b-th feature point sampled from BP_i and BP_j respectively, $conn$ (Eq. 4) is a function to compute the connectivity between the feature points $v_{i,a}$ and $v_{j,b}$, M is the volumetric mesh computed using Eq. 2, and n_i and n_j are the total number of feature points sampled from BP_i and BP_j, respectively. The larger the *SRBP*, the more the pair interacts. In Eq. 3, however, longer body parts usually have more connections with other body parts as more feature points were sampled. To tackle this problem, we divide the connectivity returned from the $conn$ function by the total number of feature points sampled from the two body parts as normalization.

While most of the edges are short and connecting nearby points, the extreme points in the point cloud will be connected to each other since Delaunay Tetrahedralization produces a convex volumetric mesh. The long edges connecting the extreme points cannot truly reflect the interactions between the body parts. Instead of determining a suitable threshold value to exclude long edges from *SRBP* calculation, we propose to compute the connectivity between two feature points by a value that is inversely proportional to the Euclidean distance between the two feature points if they are connected; otherwise, zero will be returned. By this, an insignificant small value will be returned from $conn$ if the edge is long:

$$conn(M, v_{i,a}, v_{j,b}) = \begin{cases} \frac{1}{||v_{i,a} - v_{j,b}||} & \text{Connected in } M \\ 0 & \text{Not connected in } M \end{cases} \tag{4}$$

4.2.4 Abstraction of SRBP at Limb-level

The main purpose of computing *SRBP* at Limb-level is to reduce the dimensionality of the representation. We propose to compute the *SRBP* at Limb-level (*SRBP-Limb*) as an abstract representation. More specifically, a human-like skeletal structure is divided into 6 limb groups: *1) Head and Neck, 2) Right Arm, 3) Left Arm, 4) Right Leg, 5) Left Leg and 6) Torso*. The calculation of *SRBP-Limb* is similar to computing *SRBP*:

$$SRBP\text{-}Limb_{i,j} = \frac{\sum_{a=1}^{m_i}\sum_{b=1}^{m_j} conn(M, v_{i,a}, v_{j,b})}{m_i + m_j} \tag{5}$$

(a) (b)

Figure 1: *(a)-(b) Example poses used in the experiment presented in Section 5.2: (a) Hug from behind and (b) Assist walk/stand in the two-character close interaction dataset [Ho and Komura 2009].*

where $SRBP_{i,j}$ is the *SRBP* of the limbs L_i and L_j, $v_{i,a}$ and $v_{j,b}$ are the a-th and b-th feature point sampled from L_i and L_j respectively, $conn$ is the function defined in Eq. 4, and m_i and m_j are the total number of feature points sampled from L_i and L_j, respectively. Finally, the pose in each frame is represented by the Relative Position features *RP* and a ${}_6C_2 = 15$-dimensional vector of pairwise *SRBP-Limb*.

4.3 Representing Two-character Interaction

In two-character interactions, the context of the interaction is concentrated in the inter-relations (i.e. pairing up the limbs from different subjects). Using two-character interaction as an example, $6 \times 6 = 36$ *SRBP-Limbs* (i.e. a 36-dimensional vector) are used as the representation. We expect that multi-subject interactions can also be represented by ${}_hC_2 \times 36$ *SRBP-Limbs* using the concepts described above, where h is the number of subjects in the interaction. Experimental results show that comparing poses using *SRBP-Limb* in two-character interactions improves the intra- and inter-class classification accuracy and the details are presented in Section 5.2.

5 Experimental Results

In this section, we present the experimental results on retrieving and classifying two-character interactions (Section 5.1 and 5.2). In particular, the benchmark SBU Kinect Interaction Dataset [Yun et al. 2012] and 3D human motions from [Ho and Komura 2009] were used. Finally, we analyze how the parameters affect the performance of our proposed method in Section 5.3.

Implementation Details In all experiments, we sample the feature points at every 9cm along the body segment (i.e. bone). The volumetric mesh structure (explained in Section 4.2.2) in each frame is computed using the Delaunay Tetrahedralization function in MATLAB. To avoid the temporal misalignment between motions, a classical DTW function [Rabiner and Juang 1993] is used. In the motion classification tasks , we trained linear binary SVM classifiers by LIBSVM [Chang and Lin 2011] to classify motions in the benchmark datasets in a one-versus-all manner. Finally, we pick the action class which returns the highest decision value as the class label of the testing motion.

5.1 SBU Kinect Interaction Dataset

In this experiment, we evaluate the accuracy of classifying motions captured from two subjects simultaneously. The 3D skeletal data in the SBU Kinect Interaction Dataset [Yun et al. 2012] are used in this experiment. The results (in Table 1) show that our proposed method outperforms [Yun et al. 2012] in the overall accuracy in both sets. Firstly, for Set 1, the results obtained using *SRBP* alone outperforms the Joint Distance (JD) descriptor [Yun et al. 2012] by 3.28%. By further using *SRBP* and JD to represent each frame, 93.90% classification accuracy is achieved, which improves the results presented in [Yun et al. 2012] significantly by 5.01%. For Set 2, using *SRBP* alone achieved 1.44% higher classification accuracy over JD. By using both *SRBP* and JD, we achieved 95.08% which significantly outperforms [Yun et al. 2012] by 5.33%. These results highlight the consistency and robustness of our method over [Yun et al. 2012].

Method	Accuracy	
	Set 1	Set 2
Joint distance (JD) [Yun et al. 2012]	88.89%	89.75%
Our method ($SRBP$ only)	92.17%	91.19%
Our method (JD and $SRBP$)	**93.90%**	**95.08%**

Table 1: *Comparison of recognition accuracy on the SBU Kinect Interaction Dataset.*

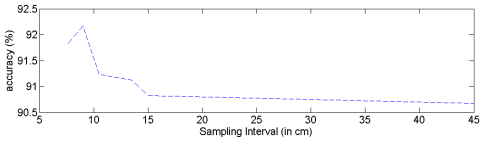

Figure 2: *Parameter evaluation of our proposed method on SBU Kinect Interaction Dataset - Accuracy VS Feature points sampling resolution.*

5.2 Wrestling and Dancing Interactions

In this experiment, we evaluate the effectiveness of using our representation in classifying two-character close interactions. The poses used in the experiment are obtained from the close interaction dataset in [Ho and Komura 2009]. The dataset contains 14 interactions groups. Differentiating poses between different action groups in the dataset is challenging because the poses in same interaction group have large variations in terms of low-level features such as joint angles and positions (see Figure 1 (a) and (b)). In summary, the poses in the same action class are having large variations while poses from different action groups are similar in low-level joint configurations.

We compared our proposed method with the Joint Distance (JD) descriptor [Yun et al. 2012] and Spacetime Proximity Graphs [Tang et al. 2012] by computing similarity matrices on the selected dataset and the results are shown in Figure 3. We also computed *SRBP* at two levels - *Body part-* and *Limb-* level to show the effectiveness of the abstraction of the spatial relations in classifying the poses. The results show that using all pairs of joint distances (Figure 3 bottom, (a)) to represent the poses in this challenging dataset results in low similarity within each action group. Results obtained using Spacetime Proximity Graphs [Tang et al. 2012] (Figure 3 bottom, (b)) shows high inter-class similarity as many poses in different classes are considered as similar (e.g. class 5-6, 8-9, 11-14) as well as low intra-class similar (e.g. class 1, 14). On the other hand, the poses represented by *SRBP* (Figure 3 bottom, (c) and (d)) show strong intra-group similarity. By further abstracting the spatial relations by computing the *SRBP* at Limb-level, logically similar poses can be detected as brighter color is shown within each action class in the similarity matrix (Figure 3 bottom, (d)). This experiment shows our proposed representation can effectively represent semantically similar two-character interactions even though they are different in low-level features such as joint positions and angles.

5.3 Parameters Evaluation

In this section, we analyze how the parameters affect the performance (i.e. classification accuracy) of the proposed method. We

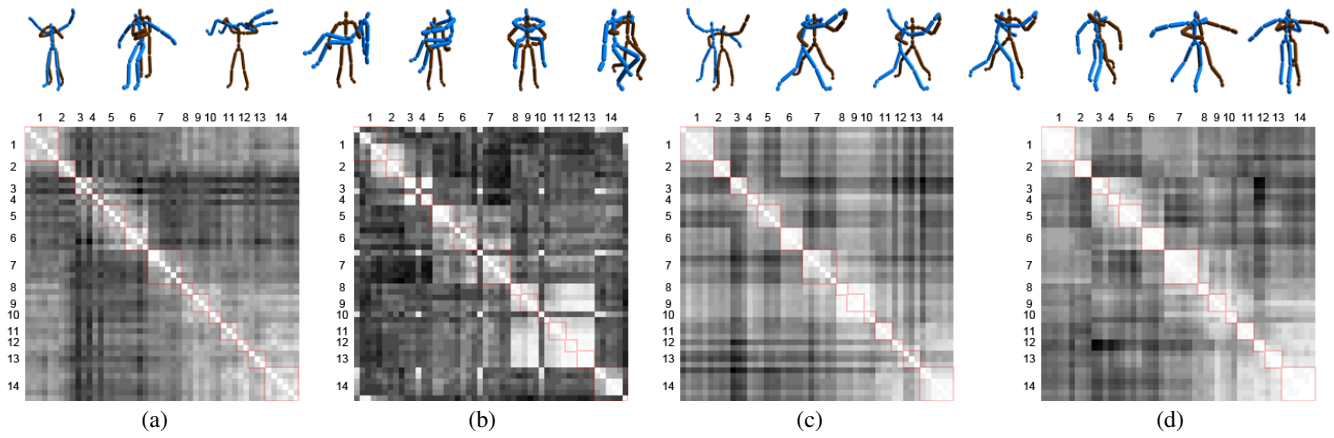

Figure 3: *Top row: Example poses in action groups 1-14 (from left to right). Bottom row: Similarity matrices computed using (a) Joint relative distances, (b) Spacetime Proximity Graphs [Tang et al. 2012], (c) Body segment-level SRBP, and (d) Limb-level SRBP. The red squares on the matrices indicate the 14 action groups.*

tested our method by sampling the feature points with different resolutions and use the computed *SRBP* features in classifying motions as in the cross subject experiment explained in Section 5.1. The results (in Figure 2) indicate that using high resolution feature points improves the classification accuracy and reached the best performance at 92.17% classification accuracy when sampling feature points from body part on every 9 cm. For timing information, computing *SRBP* for two-characters requires 0.73 seconds in our MATLAB implementation. We expect the computational cost can be reduced by using an optimized Delaunay Tetrahedralization implementation.

6 Conclusions

In this work, we proposed a new representation called *Spatial Relations of Human Body Parts (SRBP)* which is based on the spatial relations between the body parts of the subjects for retrieving and classifying character interactions. By representing human poses using the local spatial relations of body parts in proximity, the performance is improved as the proposed representation is robust to the variations on the individual joints locations. Experimental results show that our method outperforms other commonly used skeleton-based approaches in retrieving and classifying semantically similar motions, even though the motions are significantly different in low-level features such as joint positions and angles.

Acknowledgements

This work was supported in part by Hong Kong RGC (GRF210813), the National Natural Science Foundation of China (61302176) and the HKBU FRG (FRG2/13-14/092).

References

CELIKTUTAN, O., AKGUL, C. B., WOLF, C., AND SANKUR, B. 2013. Graph-based analysis of physical exercise actions. In *Proceedings of the 1st ACM International Workshop on Multimedia Indexing and Information Retrieval for Healthcare*, ACM, New York, NY, USA, MIIRH '13, 23–32.

CHANG, C.-C., AND LIN, C.-J. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology 2*, 27:1–27:27.

DELAITRE, V., SIVIC, J., AND LAPTEV, I. 2011. Learning person-object interactions for action recognition in still images. In *Advances in Neural Information Processing Systems*.

HO, E. S. L., AND KOMURA, T. 2009. Indexing and retrieving motions of characters in close contact. *Visualization and Computer Graphics, IEEE Transactions on 15*, 3 (May-June), 481–492.

HO, E. S. L., KOMURA, T., AND TAI, C.-L. 2010. Spatial relationship preserving character motion adaptation. *ACM Trans. Graph. 29*, 4 (July), 33:1–33:8.

HO, E. S. L., CHAN, J. C. P., KOMURA, T., AND LEUNG, H. 2013. Interactive partner control in close interactions for real-time applications. *ACM Trans. Multimedia Comput. Commun. Appl. 9*, 3 (July), 21:1–21:19.

KOVAR, L., GLEICHER, M., AND PIGHIN, F. 2002. Motion graphs. *ACM Trans. Graph. 21*, 3 (July), 473–482.

KYAN, M., SUN, G., LI, H., ZHONG, L., MUNEESAWANG, P., DONG, N., ELDER, B., AND GUAN, L. 2015. An approach to ballet dance training through ms kinect and visualization in a cave virtual reality environment. *ACM Trans. Intell. Syst. Technol. 6*, 2 (Mar.), 23:1–23:37.

MÜLLER, M., RÖDER, T., AND CLAUSEN, M. 2005. Efficient content-based retrieval of motion capture data. *ACM Trans. Graph. 24*, 3 (July), 677–685.

PRONOST, N., MULTON, F., LI, Q., GENG, W., KULPA, R., AND DUMONT, G. 2008. Interactive animation of virtual characters: Application to virtual kung-fu fighting. In *Cyberworlds, 2008 International Conference on*, 276–283.

RABINER, L., AND JUANG, B.-H. 1993. *Fundamentals of speech recognition*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

TANG, J. K., CHAN, J. C., LEUNG, H., AND KOMURA, T. 2012. Interaction retrieval by spacetime proximity graphs. *Computer Graphics Forum 31*, 2pt4, 745–754.

YUN, K., HONORIO, J., CHATTOPADHYAY, D., BERG, T., AND SAMARAS, D. 2012. Two-person interaction detection using body-pose features and multiple instance learning. In *CVPR Workshops*, 28–35.