UPSR: a Unified Proxy Skeleton Retargeting Method for Heterogeneous Avatar Animation

Wenfeng Song
Computer School, Beijing Information Science and Technology University, Beijing, China

Xinyu Zhang
Computer School, Beijing Information Science and Technology University, Beijing, China

Yang Gao*
State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China

Yifan Luo
Computer School, Beijing Information Science and Technology University, Beijing, China

Haoxiang Wang
Computer School, Beijing Information Science and Technology University, Beijing, China

Xianfei Wang
Computer School, Beijing Information Science and Technology University, Beijing, China

Xia Hou
Computer School, Beijing Information Science and Technology University, Beijing, China

2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW) | 979-8-3503-4839-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/VRW58643.2023.00276

ABSTRACT

Pose-driven avatar animation is widely applied in VR fields. The flexible creation of animations is a significant yet challenging task due to the heterogeneous topologies and shapes of avatars. To alleviate this, we propose a unified proxy skeleton for retargeting (UPSR) to achieve consistent motions of the virtual avatar. Particularly, the heterogeneous topologies are converted into a unified skeleton topology of the avatars by using a learned nonlinear mapping function. Furthermore, we propose to retarget the skeletons with different body shapes into 3D virtual avatars. Our UPSR can produce avatar animations with a higher level of authenticity without the dependency on high-cost motion capture devices or the restrictions of its topology. Additionally, we could drive detailed motions based on multiple sources of motion datasets, including monocular videos, motion capture devices, and standard motion files, with the precise motion capture of eyes, lips, hands, and bodies.

Index Terms: Computing methodologies—Computer graphics—Animation—Motion capture;

1 INTRODUCTION

High-fidelity avatar animation generation has been widely used in VR-related applications, such as virtual anchors and 3D animation creation. Virtual character animations have lately gained popularity in both the industrial and research sectors due to their efficacy when compared to time- and money-consuming traditional manual animation creation [1]. In this study, to adapt to the various standards of motion representation, we propose a unified proxy skeleton that encodes various topologies of structures as an immediate one, including the primary control points of facial emotions, hand, and motion clues. Furthermore, to bridge the body shape gaps between arbitrary keypoints and the arbitrary virtual avatar, we propose to solve the homeomorphic topologies via an improved retargeting network. We propose a unified motion re-targeting paradigm to non-linearly map the various structural skeletons with arbitrary topologies, which gives rise to a flexible yet effective method for virtual avatar generation. We provide a simple proxy skeleton structure that allows for the proper avatar animation generation for non-homomorphic topological skeletons.

2 HETEROGENEOUS AVATAR ANIMATION GENERATION

The purpose of our study is to animate high-fidelity avatars with arbitrary topologies from a variety of motion capture media, such as monocular RGB films, kinect skeletons, and Mixamo [2]. Fig. 1 shows the pipeline of our framework. There are three main components to our work: (1) We first feed the multiple sources of monocular video motion capture. Considering the multiple media motion capture process has various control points to represent human motion, we propose a unified framework to track the motion status. Different from Mixamo or Kinect-based motions, we directly retarget the skeletons to the virtual skeletons through nonlinear network mappings. This preprocessing module has the advantages of high efficiency and robustness. (2) We propose unifying the un-homeomorphism topologies structures. For example, body structures may have two line segments representing the upper body, while others may only have a single line segment in the center of the upper body. In this component, we first design a single-line body proxy kinematic skeleton, which is close to most virtual avatars. In the meantime, we create mapping relationships to convert other standards into proxy skeletons. (3) The main part of our UPSR is a nonlinear mapping learned from neural networks that retargets the proxy skeletons into homeomorphic virtual avatars. Consequently, we could drive virtual avatars through arbitrary typologies.

2.1 Motion Capture from Multiple Sources

We utilize the various source movements that are currently accessible in order to obtain sufficient motion source datasets. There are two ways to acquire massive data resources: Firstly, we utilize the BlazePose [3] method to detect the human pose in video clips. Secondly, we collect abundant 3D animation skeletons from the open resource library of Mixamo company. BlazePose employs three different types of convolutional neural network (CNN) models: the lite model, the comprehensive model, and the heavy model. Mixamo [4] provides two kinds of skeleton online data resources: characters and animations. Characters are the appearance of the skeletons, and Mixamo provides plenty of characters for users to choose from.

2.2 Multiple Sources Unified Proxy Skeleton

In this section, we create a proxy skeleton in a mediate topology to adapt to the highly diverse topologies of the multiple source media. The topology of a proxy skeleton (aka. proxy topology) is abstracted based on the fact that most virtual human avatars are driven from general control points, such as face points to control expressions and lip motion, and torso body components to govern activities.

On the one hand, in terms of the RGB videos, we first obtain the pose and keypoints of the bodies [3]. It should be noted that there are two difficulties in driving virtual avatars. The first is that the keypoints lack relationships between distinct nodes. The parent and child node topologies are not specified, which is crucial when driving the avatars. As a result, we must empirically increase the clues in order to generate armatures from keypoints. The second point to mention is that structures are not always homeomorphic.

*Email: gaoyangvr@buaa.edu.cn

979-8-3503-4839-2/23/$31.00 ©2023 IEEE
DOI 10.1109/VRW58643.2023.00276

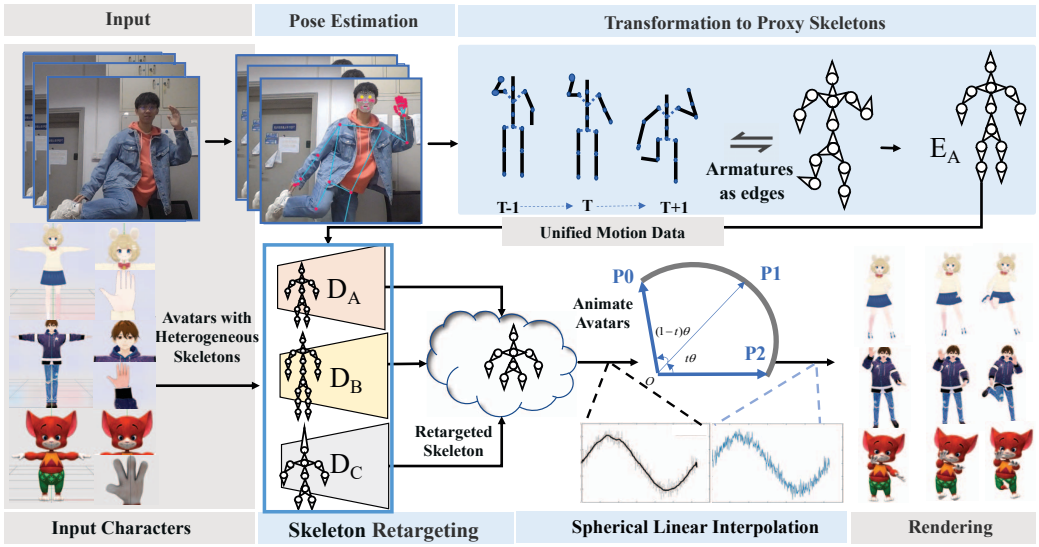

Figure 1: An overview of our unified proxy skeleton retargeting framework (1) Motion capture in monocular video. This module is realized through a unified face, hand, and body keypoint estimation model. (2) An adaptor that converts the raw motion skeletons into a topology proxy of the skeletons that match the rigid characters. (3) A kinematic key points map drives a virtual character. We provide a proxy skeleton map of keypoints, which is suitable for VRM, Mixamo, and other classical format bones' naming schemas and coordinate system standards.

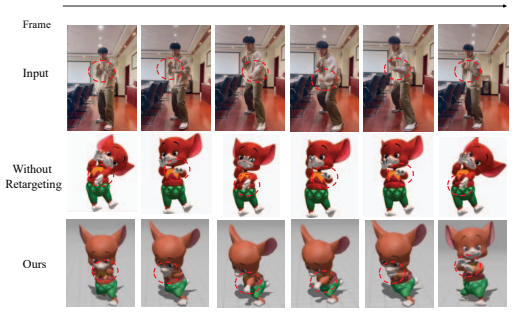

Figure 2: Illustration of our method. It is capable of adapting to avatars accurately.

Estimated keypoints, for example, merely provide a skeleton-like style without the edges, such as spines. And the edges of the spines are generally designed as virtual avatars. The raw motion from the estimated keypoints is difficult because the wrist has just four points. This could result in an inability to identify fine details, such as whether the back is straight when leaning over. Meanwhile, in terms of the motion data from the mixamo website, we can also feed it into our retargeting pipeline.

2.3 Motion Retargeting based on Proxy Skeleton

In the aforementioned section, we unify the different structures into a homeomorphic topology. Based on this, we can drive virtual avatars with high diversity. In practice, the captured skeletons may differ in their number of key points, bones, and length of arm or leg. To solve the potential consequences caused by the physical diversity of the virtual avatars, we introduce motion retargeting to unify the captured skeletons.

Extending from the existing high-performance homeomorphic methods, we propose a novel arbitrary homeomorphism retargeting pipeline. We also use two parallel branches in general: a dynamic branch and a static branch. The dynamic branch could encode the animations' time-dependent features, while the static branch encodes the spatial features of the keypoints positions. Accordingly, the two branches fully characterize the motion features. Therefore, it is critical that the convolution layer calculates with both features in consideration.

The dynamic branch and the static branch share the same map key points. This makes it possible to apply skeletal convolution to motion sequences and keep the branches' dimensions and meanings consistent. To represent the motion features for the skeleton structure, we utilize the following operations following [5]. The Deep Skeletal operators are defined to be applied to animated skeletons, which consider the skeleton structure. The operators allow the neural network to learn the low-level, local joint correlations in shallow layers and the high-level, global body part correlations in deeper layers. Hence, the operator is selected for our retargeting from the proxy skeleton. The results are shown in Fig. 2 (see supplement videos).

3 Conclusion

We propose a unified proxy skeleton for retargeting (UPSR) to achieve consistent motions of the virtual avatar. Furthermore, we propose a novel retargeting method for the skeletons with different body shapes into 3D virtual avatars. Our UPSR could produce avatar animations with a higher level of authenticity without the dependency on high-cost motion capture devices or the restrictions of its topology.

References

[1] Man To Tang, Victor Long Zhu, and Voicu Popescu. Alterecho: Loose avatar-streamer coupling for expressive vtubing. In *ISMAR*, pages 128–137, 2021.

[2] Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, et al. Blazepose: On-device real-time body pose tracking. *arXiv preprint arXiv:2006.10204*, 2020.

[3] Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, et al. Blazepose: On-device real-time body pose tracking. *CoRR*, abs/2006.10204, 2020.

[4] Sue Blackman. Rigging with mixamo. In *Unity for Absolute Beginners*, pages 565–573. Springer, 2014.

[5] Kfir Aberman, Peizhuo Li, Dani Lischinski, et al. Skeleton-aware networks for deep motion retargeting. *ACM Transactions on Graphics*, 39(4):62–1, 2020.