Pacific Graphics 2025
M. Christie, N. Pietroni, and Y.-S. Wang
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 44 (2025), Number 7

Body-Scale-Invariant Motion Embedding for Motion Similarity

XIAN DU[1], CHUYAN QUAN[1], Ri Yu[†1,2]

[1]Dept. of Artificial Intelligence, Ajou University, South Korea
[2]Dept. of Software and Computer Engineering, Ajou University, South Korea

Abstract

Accurate measurement of motion similarity is crucial for applications in healthcare, rehabilitation, sports analysis, and human–computer interaction. However, existing Human Pose Estimation (HPE) approaches often conflate motion dynamics with anatomical variations, leading to body-scale-dependent similarity assessments. We propose a framework for learning body-scale-invariant motion embeddings directly from RGB videos. Leveraging diverse 3D character animations with varied skeletal proportions, we generate standardized motion data and train the SAME model to capture temporal dynamics independent of body size. Our approach enables robust cross-character motion similarity evaluation. Experimental results show that the method effectively decouples kinematic patterns from structural differences, outperforming scale-sensitive baselines. Key contributions include: (1) a scalable motion data processing pipeline; (2) a learning-based body-scale-invariant embedding method; and (3) validation of motion similarity assessment independent of anatomy.

CCS Concepts

• ***Computing methodologies*** → ***Computer graphics; Motion processing; Neural networks; Learning latent representations;***

1. Introduction

Measuring motion similarity precisely is an essential task for action recognition or movement analysis in several fields such as health care, medical rehabilitation, sports performance evaluation, and human-computer interaction. In these domains, accurately assessing how similar one movement is to another can help in diagnosing motor disorders, tracking patient recovery, providing real-time feedback in physical therapy, or comparing athletic performance. A common approach to extracting skeletal representations from video involves the use of Human Pose Estimation (HPE) techniques. Given two videos of individuals performing the same action, HPE can be applied to obtain 3D pose sequences for each subject. These pose sequences are then used to evaluate motion similarity between the two individuals. However, simply comparing joint positions or joint angles often fails to yield an accurate measure of similarity. This is because the 3D poses obtained from HPE are not purely representative of motion; rather, they are influenced by individual differences in body size and proportions. As a result, direct comparison of pose features may conflate motion similarity with anatomical variance, limiting the effectiveness of such methods in applications requiring precise motion analysis. Our goal is to find a method for extracting body-scale-independent motion embedding directly from RGB video for accurate motion similarity.

2. Framework Methodology

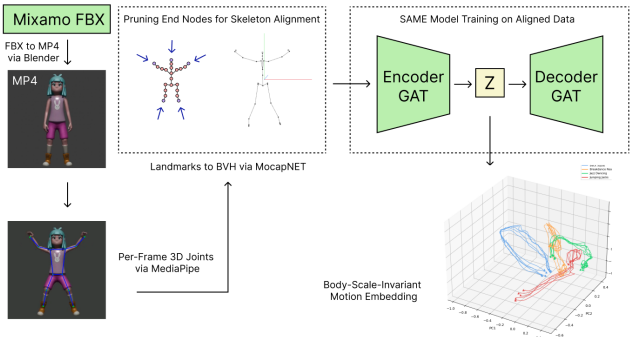

Figure 1: *Overview of Our Framework*

We propose a motion processing pipeline that converts heterogeneous character animations into a standardized format for training skeleton-agnostic motion embedding models. The workflow integrates data acquisition, motion extraction, format conversion, and skeletal alignment into a coherent process.

Data Acquisition and Preprocessing: As illustrated in Figure (Fig. 1), we obtain character and motion data from Mixamo [Ado], comprising 28 unique characters in FBX format. Each character includes a T-pose and four motion types. The FBX files are imported into Blender [Fou] and rendered into MP4 videos, preserving the full temporal and spatial motion details for subsequent analysis.

† Corresponding author

Motion Data Extraction and BVH Conversion: We employ MediaPipe [LTN*19] to process each MP4 video, extracting per-frame 3D pose coordinates by detecting and tracking body landmarks. These coordinates are then transformed into BVH format using MocapNET [QA21], which reconstructs the motion into a consistent skeletal hierarchy. This conversion ensures compatibility with motion processing frameworks and retains precise kinematic characteristics across different characters.

Data Alignment and Model Training: To meet the input requirements of the SAME model [LKP*23], we standardize the BVH data by retaining exactly 27 joints, unifying joint nomenclature, and pruning redundant nodes to match the target skeleton topology. The aligned dataset is then used to train the SAME model, enabling the learning of skeleton-agnostic motion embeddings that robustly capture motion features across heterogeneous skeletal structures.

This pipeline produces a uniform BVH representation optimized for the SAME model, enabling robust and skeleton-independent motion analysis across diverse character models. Our approach ensures that learned embeddings are independent of specific skeleton structures, facilitating robust motion comparison and analysis across diverse character models.

3. Experimental Results

We evaluate the proposed body-scale-invariant motion embedding framework on a dataset comprising four motion types — Back Squat, Breakdance Ready, Jazz Dancing, and Jumping Jacks — each performed by 28 character models with diverse skeletal proportions.

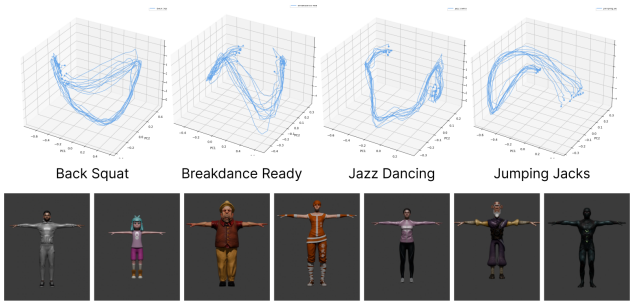

Figure 2: *Motion-Specific Embedding Clustering*

Motion-Specific Embedding Clustering: Figure 2 illustrates the 3D embedding trajectories for all sequences. Each motion type forms a distinct and compact cluster with minimal inter-class overlap, indicating that the learned embeddings capture discriminative motion dynamics independent of character identity or body scale. Cyclic actions such as Jumping Jacks yield highly regular, closed trajectories, while complex motions like Breakdance Ready produce more irregular yet clearly separable paths.

Invariance to Body Proportion Variability: To assess robustness against morphological differences, we compare synchronized Jumping Jacks sequences from two characters with markedly different body proportions (Figure 3). The DTW-aligned cosine similarity analysis shows a pooled similarity score of 0.9478, a mean per-frame resampled similarity of 0.9197, and a mean similarity along the DTW path of 0.9214, with an optimal path deviation of less than 5%. Minor discrepancies appear during shoulder abduction (Frames 15–20) and spinal flexion (Frames 40–45), attributable to anatomical differences, yet overall temporal alignment remains strong.

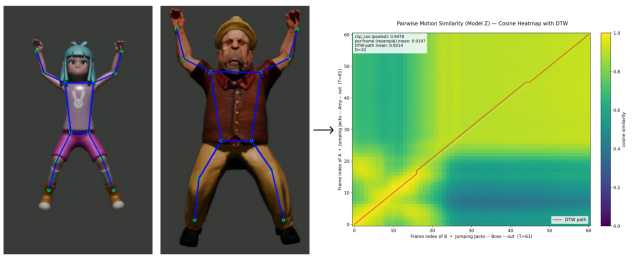

Figure 3: *Invariance to Body Proportion Variability*

These results confirm that the model effectively decouples kinematic semantics from anatomical variation, maintaining geometric separability in the embedding space and enabling style-preserving motion transfer across characters. The learned representation shows strong potential for applications in gait analysis, sports kinematics, and rehabilitation monitoring.

4. Conclusions

In this work, we introduced a body-scale-invariant motion embedding framework that decouples kinematic patterns from anatomical proportions, enabling objective motion similarity analysis across diverse body shapes. The learned embeddings have potential in rehabilitation monitoring, sports performance analysis, and related domains where proportional bias limits accuracy. While current results demonstrate robustness in controlled scenarios, the method remains sensitive to occlusions in unconstrained environments. Future research will focus on real-world video adaptation and multimodal sensor fusion to further enhance reliability and applicability.

References

[Ado] ADOBE: Mixamo animation services. URL: https://www.mixamo.com. 1

[Fou] FOUNDATION B.: Blender. URL: https://www.blender.org. 1

[LKP*23] LEE S., KANG T., PARK J., LEE J., WON J.: Same: Skeleton-agnostic motion embedding for character animation. In *SIGGRAPH Asia 2023 Conference Papers* (New York, NY, USA, 2023), SA '23, Association for Computing Machinery. URL: https://doi.org/10.1145/3610548.3618206, doi:10.1145/3610548.3618206. 2

[LTN*19] LUGARESI C., TANG J., NASH H., MCCLANAHAN C., UBOWEJA E., HAYS M., ZHANG F., CHANG C.-L., YONG M., LEE J., ET AL.: Mediapipe: A framework for perceiving and processing reality. In *Third workshop on computer vision for AR/VR at IEEE computer vision and pattern recognition (CVPR)* (2019), vol. 2019. 2

[QA21] QAMMAZ A., ARGYROS A. A.: Towards holistic real-time human 3d pose estimation using mocapnets. In *British Machine Vision Conference (BMVC 2021)* (November 2021), BMVA, p. 418. 2