mathematics

MDPI

Article

Enhancing Robustness of Viewpoint Changes in 3D Skeleton-Based Human Action Recognition

Jinyoon Park [1,2], Chulwoong Kim [2] and Seung-Chan Kim [1,*]

[1] Machine Learning Systems Lab., Department of Sport Interaction Science, Sungkyunkwan University, Suwon 16419, Republic of Korea; jin.park@taiipa.kr
[2] TAIIPA—Taean AI Industry Promotion Agency, Taean 32154, Republic of Korea; paul@taiipa.kr
* Correspondence: seungchan@ieee.org; Tel.: +82-31-299-6918

Abstract: Previous research on 3D skeleton-based human action recognition has frequently relied on a sequence-wise viewpoint normalization process, which adjusts the view directions of all segmented action sequences. This type of approach typically demonstrates robustness against variations in viewpoint found in short-term videos, a characteristic commonly encountered in public datasets. However, our preliminary investigation of complex action sequences, such as discussions or smoking, reveals its limitations in capturing the intricacies of such actions. To address these view-dependency issues, we propose a straightforward, yet effective, sequence-wise augmentation technique. This strategy enhances the robustness of action recognition models, particularly against changes in viewing direction that mainly occur within the horizontal plane (azimuth) by rotating human key points around either the z-axis or the spine vector, effectively creating variations in viewing directions. We scrutinize the robustness of this approach against real-world viewpoint variations through extensive empirical studies on multiple public datasets, including an additional set of custom action sequences. Despite the simplicity of our approach, our experimental results consistently yield improved action recognition accuracies. Compared to the sequence-wise viewpoint normalization method used with advanced deep learning models like Conv1D, LSTM, and Transformer, our approach showed a relative increase in accuracy of 34.42% for the z-axis and 10.86% for the spine vector.

Keywords: action recognition; machine learning; feature learning; skeletal data; data augmentation

MSC: 68T07

Citation: Park, J.; Kim, C.; Kim, S.-C. Enhancing Robustness of Viewpoint Changes in 3D Skeleton-Based Human Action Recognition. *Mathematics* **2023**, *11*, 3280. https://doi.org/10.3390/math11153280

Academic Editors: Xujuan Zhou, Lemai Nguyen and Guohun Zhu

Received: 1 June 2023
Revised: 3 July 2023
Accepted: 5 July 2023
Published: 26 July 2023

1. Introduction

During the last decade, 3D skeleton-based action recognition has gained much interest due to the increased affordability of obtaining 3D skeleton data, thanks to devices like Kinect and even methods for extracting data from 2D images [1–3]. However, real-world perturbations, such as viewpoint variations, make action recognition challenging. For instance, when action sequences are simple and short (e.g., Microsoft Research Action 3D (MSRA) [4], Florence 3D Action (FLO) [5], and UTKinect Action 3D (UTK) [6]), the sequence-wise view normalization approach can mitigate the issues arising from view variations. However, the sequence-wise view normalization approach may not be applicable to longer action sequences (e.g., Human 3.6M Pose (H36M) [7]) due to the complex and redundant nature of everyday motions. Despite efforts to improve robustness against the real-world perturbations in human action recognition, many current approaches still struggle with handling complex and diverse human motions, particularly when dealing with long sequences of everyday activities, such as having discussions, eating, talking on the phone, etc. These issues can be alleviated by collecting sufficient data from diverse viewpoints, enabling the development of models that can better adapt to variations in viewing angles [8]. However, as evidenced by the estimated kernel densities in Figure 1, most

existing datasets used for training and evaluation are often collected under controlled settings, which may not adequately represent the diverse range of viewing angles encountered in real-world situations [4–6]. This may lead to overfitting of the dataset during the training phase, resulting in the model's inability to extract motion features when it is arbitrarily rotated. To address these challenges, researchers have explored various techniques, such as view-invariant feature learning, data augmentation, and action synthesis. These techniques aim to enhance the model's ability to recognize actions from real-world perturbations, such as different viewpoints and diverse scenarios [9,10]. Building on the advancements made by researchers, we propose a simple yet robust sequence-wise augmentation technique that tackles the view-dependency issues in 3D skeleton-based action recognition, particularly focusing on long and complex motion sequences. In our proposed approach, we use an augmentation technique that applies random rotations about either the z-axis or the spine vector. These rotations create viewpoint perturbations mainly within the horizontal plane (azimuth) during the model's training phase.

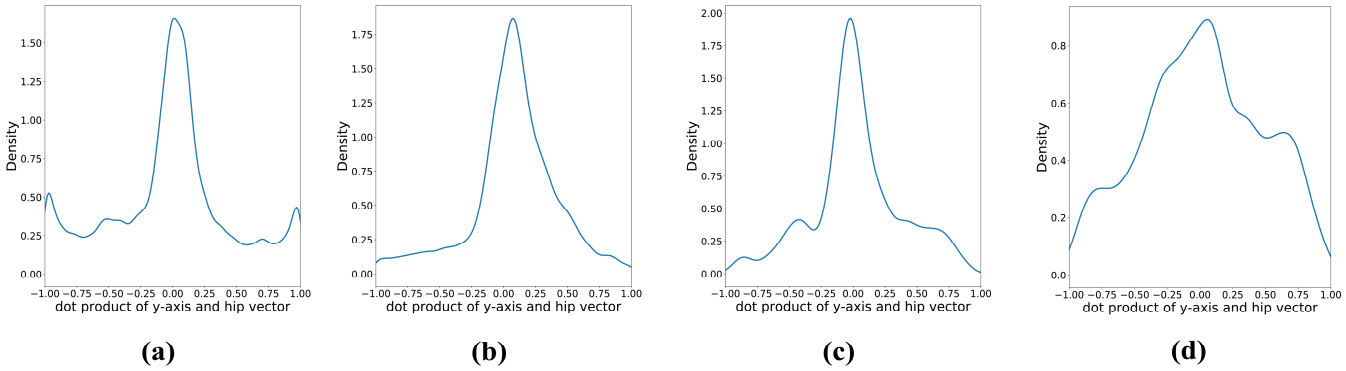

Figure 1. Kernel density estimation graph of the projection of the vector connecting the left/right hip joint onto the y-axis of the global coordinate system (i.e., <0, 1, 0>) for (**a**) H36M, (**b**) MSRA, (**c**) FLO, and (**d**) UTK datasets. Note that a peak-shaped distribution implies that the data collection procedure took place under controlled conditions (e.g., monotonous body-facing directions toward a fixed camera).

To validate our approach, we constructed deep learning-based models, such as a 1-dimensional convolutional neural network (Conv1D) [11,12], a long short-term memory network (LSTM) [13], and a Transformer [14] for the purpose of feature extraction, and investigated whether sequence-wise augmentation improves classification performance. Our experimental results demonstrate the robustness of our approach with existing deep learning-based action recognition models against real-world viewpoint variations, highlighting its potential for real-world applications. The main contributions of this paper are organized as follows:

1. We propose a novel augmentation method designed to bolster the robustness and generalization performance of 3D skeleton-based action recognition models by effectively mitigating the impacts of variations in viewing direction, primarily within the horizontal plane, for tasks involving natural human motions;
2. We extensively validate the robustness and generalizability of our proposed approach using four public datasets, as well as a custom dataset, thereby ensuring the broad applicability and robustness of our approach;
3. Through additional experiments, we determine the optimal ratio of original to augmented training data, providing a comprehensive guide for the practical implementation of our proposed methods in future applications.

In the next section, we will provide an overview of related works, discussing view-invariance in action recognition and data augmentation techniques for skeleton-based action recognition.

2. Related Works

The growing interest in computer vision-based action recognition has enabled the analysis of human motions in videos captured from arbitrary camera viewpoints, resulting in a diverse range of observable viewing directions for human movement. However, datasets are often created under controlled and limited experimental settings, leading to discrepancies in the viewing angles of 3D human skeletal data. As illustrated in Figure 1, these viewing angles frequently appear too monotone, despite some public datasets being intentionally designed that way. While many studies have addressed the robustness of classification models under realistic scenarios, the challenge of managing complex and diverse human motions still remains, particularly in action recognition for long sequences of everyday activities. To address this issue, this section provides a brief summary of data augmentation techniques for skeleton-based activity recognition and robustness against viewpoint-related real-world perturbations.

2.1. Data Augmentation for Skeleton-Based Action Recognition

Data augmentation techniques for 3D skeleton-based action recognition aim to increase the diversity and quantity of training data by applying various transformations to the original samples. Typical approaches for augmenting 3D skeletons include geometric transformations, temporal transformations, and action synthesis. Geometric transformations involve rotation, scaling, shear, and translation of 3D skeleton data to simulate variations in camera viewpoints and human body sizes [15–17]. They also include joint noise injection, such as adding Gaussian noise to the 3D joint coordinates, which simulates inaccuracies in the data acquisition process and consequently helps the model become more robust to real-world noise [18]. However, these techniques can distort the human body's natural configurations, potentially harming the model's generalization performance. Unlike these manipulations, our sequence-wise augmentation technique is grounded on physically viable human movements, ensuring more realistic data augmentation. Temporal transformations involve altering the time-related aspects of the data, such as speeding up or slowing down the action sequences, time warping, or randomly dropping frames to simulate variations in the speed of actions and temporal misalignments [16]. This concept can be extended to graph-based representations of human poses (e.g., graph sparsification [19]) without loss of generality. Action synthesis focuses on generating new action sequences by merging or modifying existing ones. A crucial requirement of this process is that the resulting sequences should contain continuous, meaningful human actions while maintaining coherence among body parts [20]. A previous study presented a novel action synthesis approach by utilizing a proposed graphical representation, which effectively captures the high-dimensional nature of human pose sequences [17]. These synthesized actions, however, might not reflect real-world scenarios effectively. Our sequence-wise augmentation technique generates variations by rotating existing sequences, thereby keeping the underlying human actions intact while enhancing the diversity of training data.

2.2. Robustness against Viewpoint Variations

To enhance the robustness of classification models under realistic scenarios, studies have proposed effective representation techniques for human motion that are robust against real-world viewpoint variations. For instance, a previous study introduced the self-similarity matrix (SSM) of action sequences, which represents an action sequence as an image corresponding to a matrix computed using joint distances between each frame [21]. This method produces similar SSM patterns even when the viewing angles of the camera are different, making the action recognition process robust to various viewing angles in the real scene. However, the SSM-based approach is often limited by its inability to differentiate between different actions that have similar patterns in joint distances, an issue our method effectively addresses by considering the whole 3D skeleton sequence for augmentation.

In another study, a compact posture representation method was introduced using histograms of 3D joint locations [6]. The clustered posture visual words were then modeled using discrete hidden Markov models (HMMs) to capture the temporal evolutions of actions. This approach demonstrated significant view invariance and achieved impressive results in various action recognition tasks. Although this method exhibits significant view invariance, it often fails to capture complex motion dynamics due to its discretized nature. In contrast, our sequence-wise augmentation approach considers continuous motion sequences, preserving complex motion patterns and temporal dynamics more effectively.

Additionally, a sequence-wise view-invariant transform was proposed to eliminate the effects of view variations on the spatiotemporal locations of skeleton joints [17]. As this approach simultaneously transforms all human poses within a sequence using a single transformation matrix, it exhibits robustness to measure noises in the joint positions. However, it fails to consider the possible variations in the viewing angles.

3. Methods

In this section, we outline our proposed approach, as shown in Figure 2, which utilizes a simple, geometry-based augmentation technique designed to address variations in the camera's rotation predominantly occurring within the horizontal plane (azimuth), a situation commonly encountered when an individual captures a video while standing or moving on a flat surface such as the ground. To that end, we begin by detailing the data preprocessing steps, which include scale normalization and translation removal. Then, following the preprocessing steps, we introduce a sequence-wise augmentation technique. Finally, we provide an overview of the experimental procedures conducted using public and custom datasets for the purpose of validating our proposed approach.

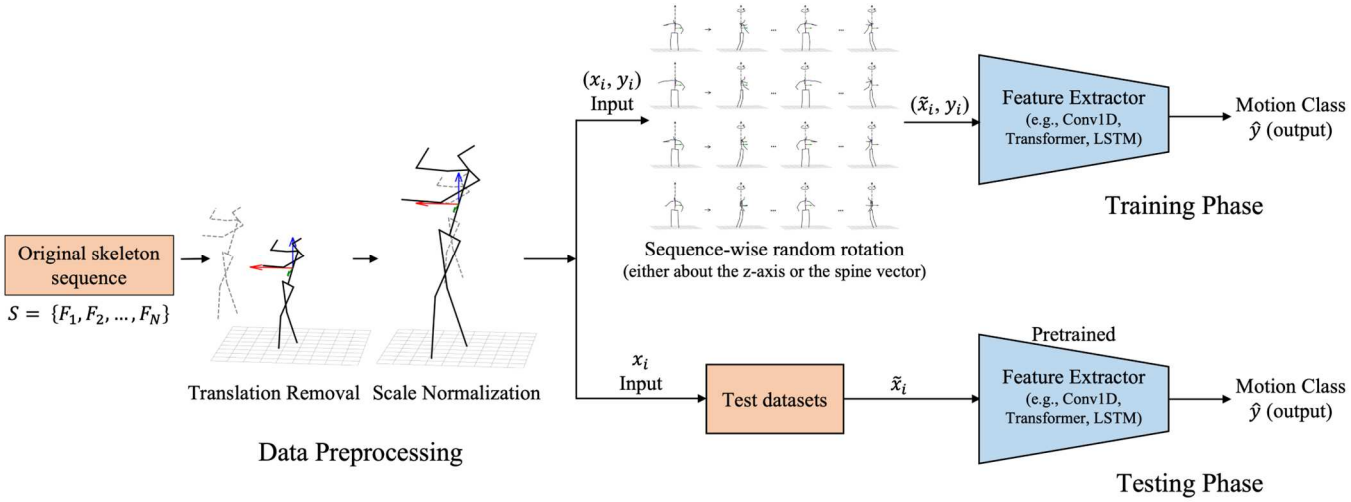

Figure 2. Overall framework of the proposed approach.

3.1. Data Preprocessing

To remove the influences of the physical stature of the subjects (e.g., height), arbitrary camera positions, and viewpoints, we normalized the skeleton data by eliminating the translation and scale of the human skeleton.

3.1.1. Translation Removal

To address the issue of the variability in the initial positions of motion sequences, we first removed the translation of action sequences by positioning the spine joint of each frame at the origin <0, 0, 0>. By doing so, we ensured a consistent position for all action sequences, allowing for more accurate comparisons and analysis of the actions, irrespective of the relative camera positions.

3.1.2. Scale Normalization

In the context of 2D image-based computer vision, scale ambiguity refers to the difficulty in determining the true size or distance of objects within the image [22]. Projecting a 3D world onto a 2D plane results in a 2D image, which loses depth information and introduces uncertainties in the scale of the objects present in the image. This issue becomes particularly problematic when extracting skeletons using models like VideoPose3D [3], as the scale information of the resulting skeletons is also affected by these ambiguities. To address this issue, we implemented a scale normalization process for the 3D skeleton data across all action sequences, as depicted in Figure 3. This process involved normalizing each skeleton's z-coordinate values by the maximum value corresponding to the head joint in the first frame as a normalization factor, and then dividing all values in the 3D pose sequence by this factor. This operation ensures a consistent scale for all the sequences, improving the reliability of subsequent action recognition tasks.

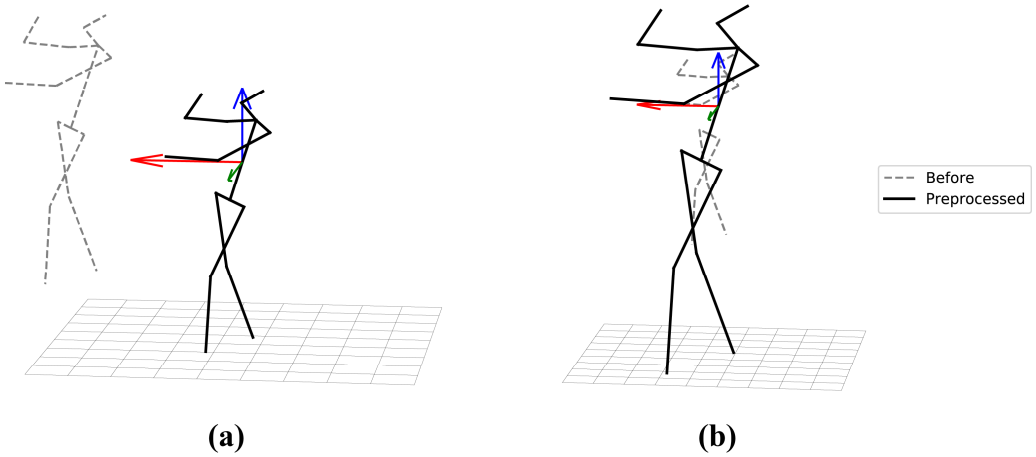

Figure 3. Examples of the proposed sequence-wise preprocessing methods. Application of (**a**) translation removal and (**b**) scale normalization process.

3.2. Data Augmentation

The proposed geometry-based sequence-wise skeleton augmentation method randomly rotates the 3D poses about the z-axis as in [23], and the spine vector. Unlike manipulations of joint positions that do not align with human capabilities (e.g., shearing [15] and jittering the 3D joints [18]), our approach is based on movements that are physically viable for humans. More specifically, the rotation matrix, R, that rotates an arbitrary 3D point by θ about either z-axis <0, 0, 1> or the spine vector that connects the neck joint and the middle hip joint of the first frame in each sequence, synchronously transforms all skeletons in the sequence, thereby preserving relative joint motions within the action sequence. During the training phase, we randomly select the degree θ for each action sequence to simulate diverse and arbitrary view transformations, imitating real-world viewpoint variations often seen when an individual captures a video while standing or moving on a flat surface, such as the ground. As an example, we select a random rotation degree from a set of 20 unique values, ranging between 1 and 19 in units of π. This introduces randomness and variability into the augmentation, thereby enhancing the technique's robustness and adaptability to view variations. As a result, this approach increases the total dataset size by a factor of 21, consisting of the original sequences and their 20 randomly rotated counterparts.

Let $\mathbf{p}_j^n$ be the position of the j-th joint in the n-th frame. The transformation can be described as

$$\begin{bmatrix} \mathbf{p}_j'^n \\ 1 \end{bmatrix} = \begin{bmatrix} R & \mathbf{0} \\ \mathbf{0}^T & 1 \end{bmatrix} \begin{bmatrix} \mathbf{p}_j^n \\ 1 \end{bmatrix} \tag{1}$$

where $\mathbf{p}_j'^n$ is the transformed position of the j-th joint in the n-th frame, R is the 3D rotation matrix calculated from the first frame of each sequence, and $\mathbf{0}$ is the zero-translation vector. The two sequence-wise random rotation techniques adopted in this study, rotation about the z-axis and along the spine vector, are demonstrated in Figure 4. Throughout the experiment, we assess which method yields more accurate and robust classification performance across different scenarios.

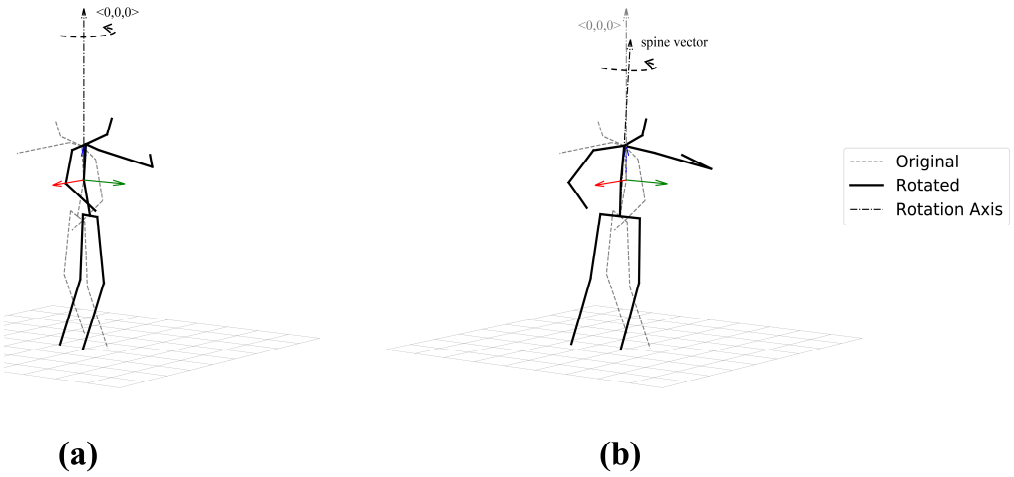

Figure 4. Illustration of 3D skeleton rotation about (**a**) z-axis and (**b**) spine vector (i.e., the vector that connects the neck joint and the mean position of left and right hip joints).

3.3. Datasets

To examine the validity of the proposed approach, we conducted extensive empirical studies using a diverse set of public datasets, namely MSRA, FLO, UTK, and H36M, as well as a custom dataset. Our choice of these particular datasets is twofold:

Firstly, MSRA, FLO, and UTK datasets are representative of typical collections in the field of action recognition, primarily featuring short, simple action sequences captured from limited viewpoints. This characteristic allows us to evaluate the performance of our approach in more controlled, experimental settings that align with common practices in the field. Secondly, the H36M dataset, though not commonly used for action recognition tasks, presents a different dimension of evaluation. It contains longer, more complex action sequences that are not strictly defined and better mirror real-world scenarios. The inclusion of this dataset hence provides a valuable benchmark for the performance of our approach when applied to more natural and real-world action classes. Additionally, we introduced a custom dataset for gait analysis of older people, which was specifically collected from real-world scenarios. This dataset serves a crucial role in affirming the robustness and applicability of our approach in authentic contexts, bridging the gap between controlled experimental conditions and the varying complexities of real-world data. Together, these datasets present a broad range of action sequence types, from simple and controlled to complex and realistic, enabling a comprehensive evaluation of our method's practical utility in handling diverse scenarios in skeletal action recognition tasks.

Microsoft Research Action 3D Dataset (MSRA) [4] is a public dataset of 3D skeleton-based action sequences captured by a depth camera at approximately 15 frames per second. The 3D skeleton data consists of 20 joints, and the dataset contains 20 action classes. Ten subjects were requested to perform predefined actions 2 to 3 times, resulting in a total of 567 action sequences. However, 20 sequences in this dataset are considered invalid due to missing data, leaving a total of 547 action sequences available for training and validation in this study. The proposed evaluation settings divide the dataset into three action sets, each containing a subset of 8 gestures.

Florence 3D Action Dataset (FLO) [5] is a public dataset of 3D action captured by a depth camera at approximately 30 frames per second, with the 3D skeleton data featuring

15 joints. The dataset contains 9 action classes, and 10 subjects were asked to perform the predefined actions 2 to 3 times. Altogether, the dataset contains 205 action sequences.

UTKinect Action 3D Dataset (UTK) [6] is captured using a Kinect camera for Window SDK Beta Version at about 30 frames per second, with the 3D skeleton data consisting of 20 joints. The dataset contains 10 action classes, and 10 subjects were asked to perform each action twice. In total, there are 200 action sequences in this dataset.

Human 3.6M Pose Dataset (H36M) [7] is a public dataset of human motion captured by a MoCap System. Utilizing four calibrated cameras, the system captures at approximately 50 frames per second, with the 3D skeleton data consisting of 17 joints. The dataset contains 15 action classes, performed by 7 subjects across two separate trials. In total, the action sequences sum up to 209 sequences. H36M dataset represents the long and complex action sequences of everyday activities, as shown in Table 1. The average length of the motions is 45.30 s (SD 11.07 s).

Table 1. Overall exploratory action sequence analysis for H36M.

Action Class	# of Total Trials	Max # of Frames	Min # of Frames	Average # of Frames (Std.)	Average Length (Secs)
Directions	13	4973	1381	2449 (884)	48.99
Discussion	14	6090	1931	3983 (1574)	79.65
Eating	14	3763	2010	2656 (467)	53.12
Greeting	14	3199	1149	1841 (551)	36.81
Phoning	14	4354	2085	3070 (665)	61.40
Photo	14	3325	1036	1882 (599)	37.63
Posing	14	2320	992	1728 (526)	34.55
Purchases	14	3124	1026	1471 (577)	29.42
Sitting	14	4533	1817	2799 (832)	55.98
SittingDown	14	6343	1501	2903 (1494)	58.06
Smoking	14	4870	2410	3372 (774)	67.43
Waiting	14	4856	1440	2735 (1193)	54.70
WalkDog	14	2732	1187	1924 (422)	38.45
Walking	14	3737	1612	2893 (782)	57.86
WalkTogether	14	3016	1231	2027 (617)	40.53
Average length in secs (std.)					45.30 (11.07)

Elderly Walking Dataset (EW) is a custom data collection derived from multiple sources. These sources include YouTube, one of the most popular social media platforms, and our own dataset of elderly individuals walking, which is partially sourced from our previous work [24]. The videos featuring elderly individuals include a variety of gait patterns. These encompass cane-assisted gait, walker-assisted gait, gait with disturbance, and gait without disturbance, aiming to provide a comprehensive representation of various walking patterns, particularly those observed among the elderly population.

4. Experiment

We conducted extensive empirical experiments to validate the robustness of our proposed approach against real-world variations in viewing direction, thereby providing a comprehensive analysis of its performance.

4.1. Experimental Setup

4.1.1. Data Segmentation

Each dataset was segmented into fixed-length sequences, T, yielding the input signal $\mathbf{x} \in \mathbb{R}^{T \times D}$. Here, D represents the dimension of each individual time step. The segmented timesteps differed depending on the dataset used, with H36M having 100 timesteps (equivalent to 2 s at 50 fps), MSRA 70 timesteps (equivalent to 4.6 s at 15 fps), FLO 35 timesteps (equivalent to 1.1 s at 30 fps), UTK 70 timesteps (equivalent to 4.6 s at 15 fps), and the EW dataset 60 timesteps (equivalent to 2 s at 30 fps). The number of timesteps was chosen in consideration of the dataset's sampling rate and motion characteristics, aligning with the methodology utilized in [25–27]. For instance, the FLO dataset, which contains relatively shorter actions, has a maximum of 35 timesteps, equivalent to 1.1 s. The datasets were segmented using a sliding window approach, with no overlaps between adjacent segments, ensuring that each data point was only included in one specific segment.

4.1.2. Data Augmentation

During the training phase, all datasets underwent random augmentation as detailed in Section 3.2, resulting in one of three overarching categories: (1) the original (T0), untouched segments; (2) the segments subject to sequence-wise view normalization (T1), serving as our baseline; and (3) the segments subjected to sequence-wise rotation, either about the z-axis (T2) or along the spine vector (T3), as per our proposed methodology. Importantly, the rotations were implemented in a way that maintained the intrinsic structure and motion within the skeletal action sequences, preserving the natural characteristics of the data.

4.1.3. Network Architecture and Training Process

For training the datasets, we leveraged a set of recent deep neural networks, namely Conv1D [11,12], LSTM [13], and Transformer (i.e., self-attention-based classification model) [14]. Each of these models is adept at learning critical features from multivariate time-series data, significantly enhancing the accuracy of classifications and predictions [28–30]. The Conv1D and LSTM models are particularly proficient in processing sequence data, while the Transformer model, with its self-attention mechanism, has shown remarkable performance in various sequence understanding tasks. Each model was selected due to its established use and recognized efficacy in handling sequence data, such as time-series. The models were independently trained and evaluated, allowing us to provide a comprehensive comparison of performance across various deep learning architectures. Consequently, they serve as a holistic platform for assessing the effectiveness of our proposed method [16,31–35].

Delving deeper into the characteristics of each model, Conv1D proves to be an effective model for processing sequence data as it employs 1-dimensional convolution operations combined with max-pooling layers to extract features from the input 3D skeleton sequences. In our experiments, our architecture comprised three blocks, each consisting of a convolutional and pooling layer. For the convolutional operations, the number of kernels in each layer was set to 32, 64, and 128, respectively, with a kernel size of 3. This relatively smaller kernel size was chosen based on prior research indicating its effectiveness in capturing local and finer-grained details [36]. To consolidate the feature maps into a single vector, we implemented a global average pooling operation at the end of the network.

LSTM, a variant of recurrent neural networks, has shown impressive prowess in managing sequence data and learning temporal dependencies. The selection of the number of LSTM layers and their respective sequence lengths T (ranging from 35 to 100) was guided by cross-validation to strike a balance between performance and computational efficiency. We stacked the recurrent cells twice (i.e., we used a stacked two-layer LSTM) with the number of recurrent units set to T, mirroring the length of the input signal $\mathbf{x} \in \mathbb{R}^{T \times D}$, following the approach adopted in previous studies [24,28]. The final recurrent hidden state then connects to a dense layer consisting of T units, functioning as a hidden layer within our network.

Lastly, we utilized a partial Transformer variant model for our experiments, incorporating only the encoder and classifier elements, thereby excluding the decoder structure. Even with this adaptation, our model upholds the fundamental feature of the Transformer architecture: the self-attention mechanism. This mechanism permits efficient comprehension of intricate relationships within the input data. The partial Transformer variant we used is particularly suitable for our classification task and presents a robust mechanism for managing multivariate time series data. For the Transformer model, the choice of specific hyperparameters, such as an embedding dimension of 64 for input data, a number of heads set to 4, and a stack number of 4, was motivated by empirical observations from previous research [37] and our own iterative testing process, which showed these settings yielded the best performance for our dataset.

Our chosen optimizer, the adaptive moment estimation (ADAM) optimizer with a learning rate of 0.001, has been widely adopted due to its efficiency and robustness in various deep learning tasks. The mini-batch size of 16 and training for up to 100 epochs were chosen to balance computational feasibility and learning performance. These parameters were iteratively refined during our model development process.

Lastly, we performed a 10-fold cross-validation to further ensure the robustness and generalizability of our models. This process aids in minimizing overfitting and provides a more accurate estimate of model performance on unseen data.

4.2. Test Sets

To validate the effectiveness of our proposed approach, we carefully designed two distinct test sets that emulate real-world viewpoint variations: (1) Sequence-wise View Normalized (SW VN) and (2) Sequence-wise Randomly Rotated about the z-axis (SW RR). The SW VN test set represents normalized viewing directions of action sequences, in which the 3D poses are transformed to face the camera direction. This test set is designed to evaluate the robustness of our approach against viewpoint variations, thus providing a benchmark for comparing performance deviations. On the other hand, the SW RR test set embodies scenarios with arbitrary viewing directions of humans in the scene, while maintaining the relative orientation of joint motions in each frame. This test set effectively simulates real-world conditions where actions are performed under diverse viewing directions.

The selection of these two particular test sets is motivated by their potential to emulate different scenarios. The SW VN test set represents a controlled environment, while the SW RR test set closely mimics the unpredictable viewing directions experienced in real-world scenarios. We compared the test accuracies of models trained under the original condition to those trained under either the baseline or our proposed conditions, which provides a comprehensive evaluation of our approach's effectiveness and robustness.

5. Results and Discussions

This study evaluated classifier performance using accuracy as the primary metric. Table 2 shows the experimental results of the classifiers trained with the original dataset, and Table 3 summarizes the classification performances with manipulated datasets, comparing test results across all test sets. A detailed overview of the empirical results is explained in this section.

Table 2. Classification accuracies of each dataset when trained with only original data.

Dataset	Model	Accuracies on Test Set (%)	
		Original	SW RR (z-Axis)
H36M	Conv1D	64.12	14.07
	LSTM	54.26	14.07
	Transformer	60.09	13.57
UTK	Conv1D	83.34	29.08
	LSTM	63.82	31.21
	Transformer	86.92	28.61
FLO	Conv1D	86.93	34.88
	LSTM	78.28	37.12
	Transformer	90.60	42.88
MSRA	Conv1D	95.40	36.49
	LSTM	69.69	15.70
	Transformer	95.85	40.04
EW	Conv1D	94.87	57.28
	LSTM	93.85	54.87
	Transformer	94.72	56.81

SW RR: Sequence-wise Randomly Rotated about the z-axis test set.

Table 3. Classification accuracies of each dataset when trained under either baseline or proposed approach conditions.

Dataset	Model	Training Condition	Accuracies on Test Set (%)		
			Original	SW VN	SW RR (z-Axis)
H36M	Conv1D	T1	16.75	66.52	15.49
		T2	90.94	**93.01**	89.80
		T3	82.74	74.46	64.94
	LSTM	T1	14.68	**57.92**	14.43
		T2	45.81	46.89	41.05
		T3	36.07	30.66	29.94
	Transformer	T1	15.61	61.71	14.92
		T2	96.65	**96.82**	96.00
		T3	89.49	81.02	70.75
UTK	Conv1D	T1	11.00	81.34	15.50
		T2	95.00	94.50	**95.50**
		T3	94.00	67.32	80.82
	LSTM	T1	12.05	55.26	13.11
		T2	88.97	**91.97**	89.95
		T3	90.00	61.84	64.34
	Transformer	T1	16.00	87.45	21.05
		T2	95.50	91.00	**92.50**
		T3	91.37	65.79	77.32
FLO	Conv1D	T1	83.23	**89.31**	39.55
		T2	85.13	85.61	86.08
		T3	89.31	86.54	74.87
	LSTM	T1	67.90	71.17	36.36
		T2	86.00	**86.00**	82.73
		T3	85.41	83.59	67.84
	Transformer	T1	86.49	**89.29**	41.93
		T2	83.77	82.79	79.03
		T3	82.27	78.98	66.95

Table 3. *Cont.*

Dataset	Model	Training Condition	Accuracies on Test Set (%)		
			Original	SW VN	SW RR (z-Axis)
MSRA	Conv1D	T1	16.72	69.16	32.05
		T2	94.93	74.16	**82.82**
		T3	95.98	29.12	56.19
	LSTM	T1	12.98	48.26	24.41
		T2	88.70	85.10	**88.27**
		T3	92.25	28.21	54.62
	Transformer	T1	14.26	69.46	31.01
		T2	90.25	88.44	**89.20**
		T3	92.17	28.81	55.59
EW	Conv1D	T1	33.06	95.11	46.04
		T2	95.11	**95.39**	94.87
		T3	95.35	63.93	73.49
	LSTM	T1	33.38	92.27	45.13
		T2	94.87	**95.23**	94.64
		T3	96.10	65.45	72.03
	Transformer	T1	33.57	94.04	48.25
		T2	95.58	**95.58**	95.11
		T3	95.13	64.34	73.57

SW VN: Sequence-wise View Normalized test set; SW RR: Sequence-wise Randomly Rotated about the z-axis test set; T1: Baseline training condition (Sequence-wise View Normalization applied); T2: Proposed approach of sequence-wise rotation about the z-axis; T3: Proposed approach of sequence-wise rotation about the spine-vector.

5.1. Classification Results

5.1.1. Original Dataset (T0) vs. View Normalization (T1)

When trained solely with the original data (T0), the classification performance of data exhibiting arbitrary viewpoints—those unseen during the training phase—lacks consistency and satisfaction, as evidenced in Table 2. However, as illustrated in Table 3, employing sequence-wise normalized viewing directions for training (T1, our baseline condition), significantly enhances classification performance for both sequence-wise view-normalized (SW VN) and sequence-wise randomly rotated (SW RR) datasets. This results in an average performance increase of approximately 45.94% compared to the original data training scenario (T0).

5.1.2. View Normalization (T1) vs. Proposed Augmentations (T2 and T3)

Notably, classification performance presents significant enhancements for the sequence-wise randomly rotated ones (SW RR), especially on the long and complex action sequences (i.e., H36M and EW). When compared to the baseline T1 condition, we observed an average performance improvement of 9.98% under the T2 condition. As for the classification improvement of the original test set, there is a clear and consistent uplift in accuracy when comparing T1 to T2. However, it is noteworthy that under the T3 condition, despite some increase in accuracy compared to T1 in certain instances (notably the H36M dataset), the performance was neither as impressive nor as consistent as in the T2 condition. These results underscore the efficacy of our proposed rotation augmentation method.

Contrastingly, models trained under the T3 condition failed to demonstrate improved classification performance. Specifically, when comparing the evaluation results of the SW VN test set to the T1 baseline condition, we observed a performance decline by 40.65%, 21.66%, 10.31%, and 29.70% for the MSRA, FLO, UTK, and EW datasets, respectively, with the Transformer model. The H36M dataset was the sole exception, exhibiting a performance improvement of 19.31% under the T3 condition. Nevertheless, this performance was still 15.80% lower compared to the improvement observed under the T2 condition.

5.1.3. Rotation about z-Axis (T2) vs. Spine Vector (T3)

For recognizing the short action sequences (e.g., MSRA, FLO, and UTK), the two methods exhibited negligible differences in performance improvements, likely due to the limited variation in upper body movements within these sequences. In scenarios where upper body movements have limited variation, the sequence-wise random rotation about the z-axis effectively becomes a similar method to the rotation about the spine vector, as the axis of rotation tends to be the same. Hence, this similarity in approach under these specific conditions results in negligible performance differences between the two methods. On the other hand, for the sequences of everyday activities characterized by their complex and redundant nature, T2 demonstrated more improvements in classification performances over T3. Our comprehensive experiments on these two rotation techniques indicate that augmenting via rotation around the z-axis is optimal for motion classification robust to viewpoint changes, particularly when dealing with intricate and lengthy sequences common in daily activities.

5.1.4. Overall Assessment

The T1 condition's performance, focusing on view normalization, substantially dropped in accuracy under the SW RR test set across all models and datasets, suggesting limitations in handling varied viewing angles found in real-world scenarios. In contrast, T2, applying z-axis rotation, consistently improved accuracy across all tests, demonstrating its efficacy and robustness against real-world perturbations such as actions of varied viewing directions. It suggests T2 as a superior approach for real-world scenarios with arbitrary viewpoints. This approach facilitates the training process to accommodate a variety of viewing angles, thereby enhancing its generalizability and flexibility. The T3 condition, involving rotation about the spine vector, showed inconsistent performance. While it enhanced results for the H36M dataset, it generally underperformed compared to T1, particularly for the MSRA dataset, indicating the spine vector rotation might not be universally applicable across different datasets and scenarios.

On the whole, our findings suggest that employing data augmentation via random rotations around the z-axis during training (T2), in conjunction with view normalization (SW VN) during testing, is an optimal approach, particularly for interpreting complex human movements such as those found in the H36M dataset.

5.1.5. Model

The Transformer model, across all datasets, consistently outperformed or matched the performance of the Conv1D and LSTM models under the T2 condition for both SW VN and SW RR tests. This trend suggests that the Transformer model might be inherently better suited to handle the z-axis rotation augmentation, which enhances its robustness against variations in viewing direction encountered in a real-world environment. Whether this performance advantage arises from the Transformer model's architecture, which allows better context understanding across time steps, or its compatibility with the rotation augmentation technique is a point for further investigation. This outcome, nonetheless, underscores the potential of combining appropriate model architectures with effective augmentation techniques for enhancing robustness in skeletal action recognition performance.

5.2. Validation of Unseen Data

To test the robustness and generalizability of our proposed methods, we carried out a blind test using an unseen subset of the EW dataset. In this evaluation phase, we selected the Transformer models that were previously trained under T0 and T2 conditions. This choice was influenced by findings from our preceding experiments, which consistently demonstrated the superior performance of the Transformer model over the other models. As in Section 4.2., we devised two blind test datasets: (1) the original, and (2) SW RR. The results from this experiment showed that the T2 training condition yielded higher classification accuracies, specifically 77.86% for the original and 78.36% for the SW RR

dataset. This performance was superior to that of the T0 condition, which recorded classification accuracies of 74.63% for the original and 50.99% for the SW RR dataset. The superior performance of the T2 condition, as compared to the T0 condition, not only on the original dataset but also on the SW RR dataset, signifies the robustness and scalability of our proposed rotation augmentation method. This suggests its capability to maintain a high level of accuracy even when dealing with unseen data that has undergone random rotations, thus highlighting its applicability in real-world scenarios with varied viewing directions.

We also examined the high-dimensional internal features ($D = 64$ in our case) learned by the Transformer model, using t-distributed stochastic neighbor embedding (t-SNE) [38]. The two-dimensional embeddings projected from the last fully connected layer are shown in Figure 5. It is worth noting that the projected feature points of the SW RR blind test dataset were not effectively clustered together when using the Transformer model trained under the T0 condition. However, when we employed the model trained under our proposed approach (T2), the projected feature points of both blind test datasets were robustly clustered according to their respective motion classes.

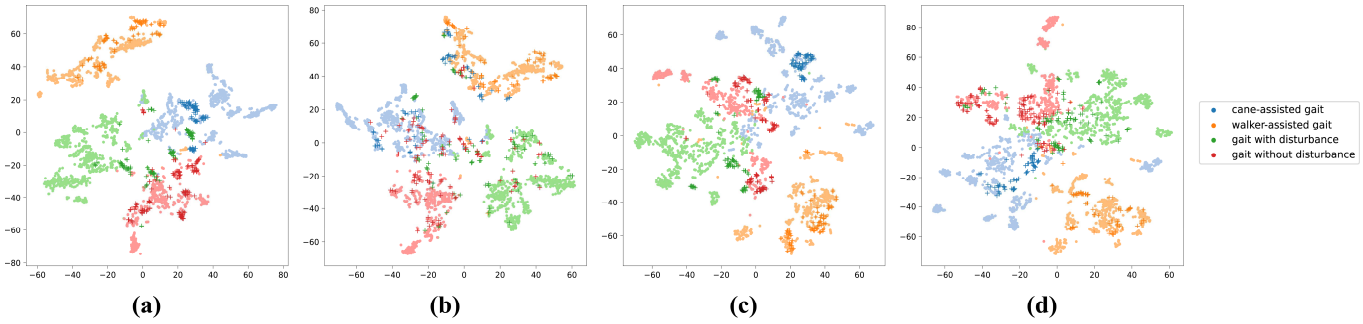

Figure 5. t-SNE visualization of high-dimensional ($D = 64$) internal features of the Transformer model, each of which is displaying (**a**,**c**) the original and (**b**,**d**) randomly rotated blind test dataset when the Transformer classification model trained with either (**a**,**b**) the original dataset or (**c**,**d**) the augmented dataset was employed. Each point is colored according to the predicted class, while the blind test datasets are marked with a cross according to the respective color of the predicted class.

5.3. Optimal Ratio of Original to Augmented Training Data

We performed a hyperparameter search to determine the optimal ratio of original to augmented training data. This allows insights into the applications of data augmentation for skeleton-based action recognitions in the future. Figure 6 illustrates the improvement in classification accuracies based on the augmentation ratio when using the Transformer model trained under the T2 condition. From these results, the optimal ratio for augmentation in the skeleton-based action recognition model was identified to be 1:7, representing a blend of original (N%) and augmented data (100%—N%). It's noteworthy that beyond this ratio, the relative improvement in experimental outcomes becomes insignificant, suggesting a point of diminishing returns for further augmentation.

5.4. Visualization of Attention Maps

A visual analysis of the Transformer model's self-attention weights provides additional insight into the effectiveness of our proposed rotation-based data augmentation technique. Figure 7 presents the attention maps generated by each head at the first self-attention layer in response to rotated test inputs. For illustrative purposes, the attention maps and visualizations of the rotated test inputs are displayed side-by-side for comparison.

In the attention maps, vertical line patterns indicate that for a particular output token, the model is paying attention to multiple specific input tokens. This suggests that the model aggregates information from multiple tokens in the input to produce a particular latent output. Note also that the attention maps appear to be quite sparse, with a lot of areas receiving no attention. This implies that only specific parts of the input sequence are

influential in generating the corresponding latent output, highlighting the model's ability to focus on pertinent information while disregarding irrelevant details.

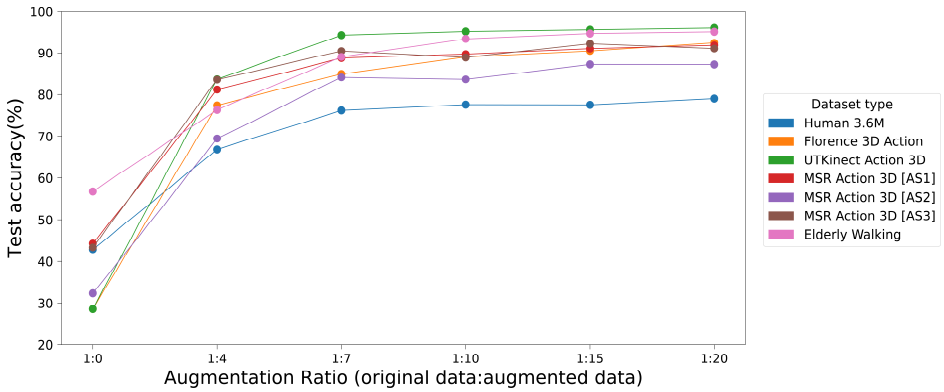

Figure 6. Test accuracy according to the ratio of original to augmented data for each dataset when the Transformer model is employed. The model is tested with the SW RR test set.

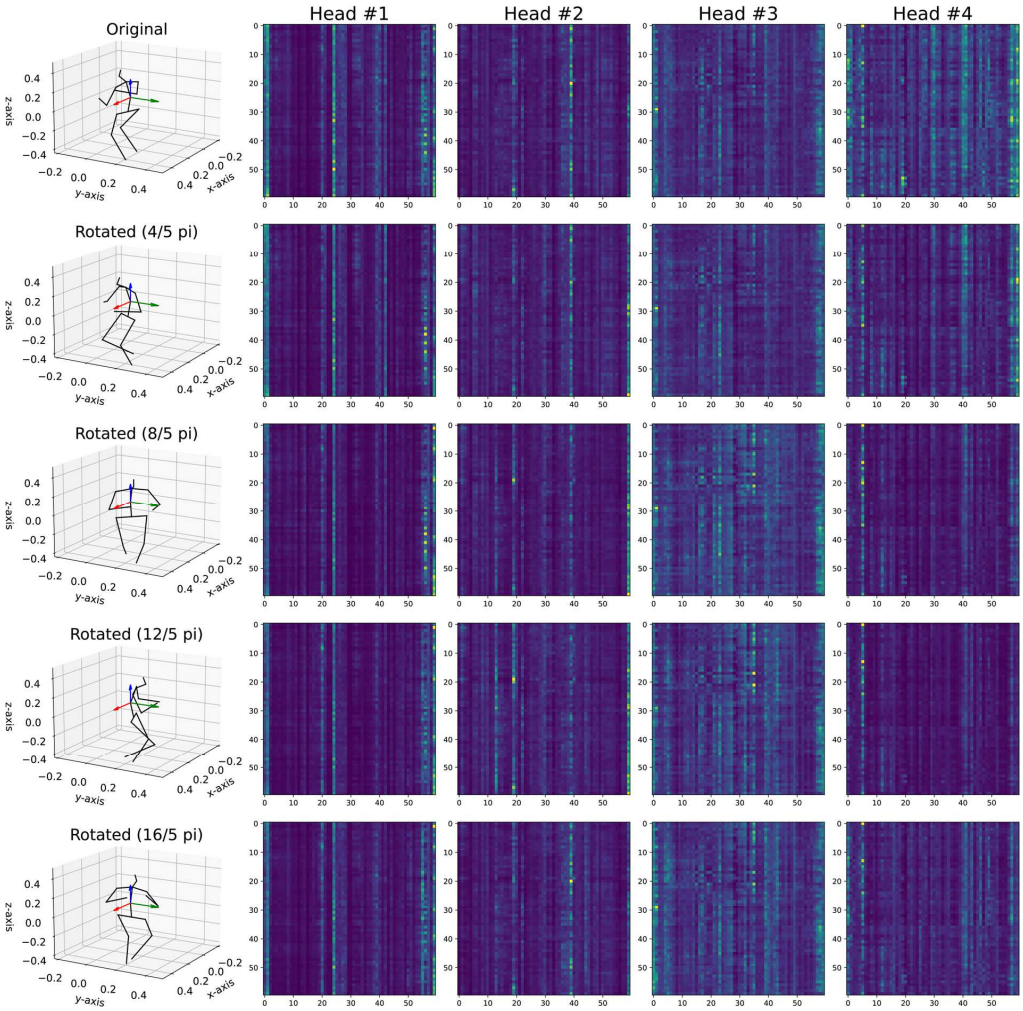

Figure 7. Illustration of the attention maps generated from each head at the first self-attention layer in response to the rotated test input. The attention maps and the visualization of the rotated test input are presented side-by-side for comparison. Given that the attention maps display similar patterns irrespective of the rotation of the input, it's noteworthy that the model focuses on similar input embeddings during the prediction process. This implies that the model has successfully learned robust motion features.

Overall, upon close examination, the attention maps exhibit strikingly similar patterns, regardless of the extent of input rotation. This observation suggests that the model consistently focuses on similar input embeddings during the prediction process. Such behavior indicates that the model has successfully learned motion features resilient to viewpoint changes from sequential motion inputs, which not only strengthens the generalization capabilities of action recognition models but also provides promising directions for future research.

5.5. Limitations and Future Work

While our study successfully validates the proposed approach, it's worth noting that our testing was confined to a limited number of publicly available datasets. Therefore, the generalizability and scalability of the proposed techniques might still need further validation across a wider array of datasets and scenarios. In future work, we intend to apply our proposed approach to other naturally occurring, non-repetitive action sequences, where learning intrinsic motion features is beneficial. We believe our rotation augmentation method can contribute significantly to these contexts, enhancing model performance across a variety of everyday scenarios. Additionally, future work includes a more in-depth analysis of our model's self-attention maps to elucidate how it learns rotation-invariant features. This could illuminate why these specific patterns emerge, refining our understanding of view-invariant feature learning, and ultimately aiding the advancement and effectiveness of our augmentation technique.

5.6. Applications

The proposed approach holds broad and practical applications, particularly in tasks necessitating view-invariant action recognition. Our study greatly contributes to learning to recognize actions from various viewpoints. With this view-invariant feature learning, human motions could be effectively recognized regardless of the user's position relative to the sensing device or changes in camera perspectives. This advantage opens up a plethora of potential applications. For instance, in video surveillance systems where cameras are positioned at varying angles and heights, our approach could enhance the identification and tracking of individuals based on their movements. Similarly, in fields such as human–computer interaction, healthcare, and sports analytics, the ability to recognize actions irrespective of the viewpoint could lead to significant improvements. More intuitive user interfaces, better patient monitoring, and comprehensive athlete performance analyses could be achieved, respectively.

6. Conclusions

In this paper, we propose a sequence-wise augmentation technique that boosts the robustness and generalization performance of 3D skeleton-based action recognition models. Our approach specifically addresses variations in viewing direction, predominantly within the horizontal plane, which commonly occur when an individual is capturing a video while standing or moving on a flat surface.

To that end, we have implemented an augmentation technique that applies random rotations to 3D key points about the z-axis or the spine vector to represent various viewing directions of humans in the scene. To validate the proposed approach, we utilized four public datasets, widely used as benchmarks in the 3D skeleton-based action recognition, to validate the proposed approach. Among the deep learning architecture employed in this study, the Transformer demonstrated the highest classification performance compared to other methods (i.e., Conv1D and LSTM). The experimental results suggest that our approach enables classifiers to learn features resilient to viewpoint changes from sequential motion inputs effectively, capturing the intricacies of everyday motions. Given that the learned attention maps exhibit consistent patterns regardless of input rotation, it suggests that the model concentrates on similar input embeddings during the prediction process, indicating the successful acquisition of robust motion features by the model. By combining

random z-axis rotations during training with view normalization during testing, we can enhance model performance significantly, particularly in the interpretation of complex human movements.

In summary, our findings have significant implications for the advancement of human action recognition models, such as those used in human–computer interaction, video surveillance, and health monitoring systems.

Author Contributions: Conceptualization, J.P. and S.-C.K.; methodology, J.P. and S.-C.K.; validation, C.K. and S.-C.K. and J.P.; investigation, J.P.; resources, J.P.; data curation, J.P.; writing—review and editing, S.-C.K., C.K. and J.P.; visualization, J.P.; supervision, C.K. and S.-C.K.; project administration, S.-C.K.; funding acquisition, C.K. and S.-C.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Korea Evaluation Institute of Industrial Technology (KEIT), which was funded by the Korean government (MOTIE) (No. 20015188).

Data Availability Statement: We cited the details of each dataset in the document.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Shotton, J.; Sharp, T.; Kipman, A.; Fitzgibbon, A.; Finocchio, M.; Blake, A.; Cook, M.; Moore, R. Real-time human pose recognition in parts from single depth images. *Commun. ACM* **2013**, *56*, 116–124. [CrossRef]
2. Zhang, Z. Microsoft kinect sensor and its effect. *IEEE Multimed.* **2012**, *19*, 4–10. [CrossRef]
3. Pavllo, D.; Feichtenhofer, C.; Grangier, D.; Auli, M. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–19 June 2019.
4. Li, W.; Zhang, Z.; Liu, Z. Action recognition based on a bag of 3d points. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, San Francisco, CA, USA, 13–18 June 2010.
5. Seidenari, L.; Varano, V.; Berretti, S.; Bimbo, A.; Pala, P. Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Portland, OR, USA, 23–28 June 2013.
6. Xia, L.; Chen, C.-C.; Aggarwal, J.K. View invariant human action recognition using histograms of 3d joints. In Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, 16–21 June 2012.
7. Ionescu, C.; Li, F.; Sminchisescu, C. Latent structured models for human pose estimation. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011.
8. Shahroudy, A.; Liu, J.; Ng, T.-T.; Wang, G. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
9. Gao, L.; Ji, Y.; Gedamu, K.; Zhu, X.; Xu, X.; Shen, H.T. View-Invariant Human Action Recognition Via View Transformation Network (VTN). *IEEE Trans. Multimed.* **2021**, *24*, 4493–4503. [CrossRef]
10. Zhang, P.; Lan, C.; Xing, J.; Zeng, W.; Xue, J.; Zheng, N. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
11. LeCun, Y.; Bengio, Y. Convolutional networks for images, speech, and time series. In *The Handbook of Brain Theory and Neural Networks*; The MIT Press: Cambridge, MA, USA, 1995; Volume 3361, p. 1995.
12. Chen, Y. Convolutional Neural Network for Sentence Classification. Master's Thesis, University of Waterloo, Waterloo, ON, Canada, 2015.
13. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
14. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
15. Wang, H.; Wang, L. Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
16. Li, B.; Dai, Y.; Cheng, X.; Chen, H.; Lin, Y.; He, M. Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep CNN. In Proceedings of the 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Hong Kong, China, 10–14 July 2017.
17. Liu, M.; Liu, H.; Chen, C. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognit.* **2017**, *68*, 346–362. [CrossRef]
18. Thoker, F.M.; Doughty, H.; Snoek, C.G. Skeleton-contrastive 3D action representation learning. In Proceedings of the 29th ACM International Conference on Multimedia, Online, China, 20–24 October 2021.

19. Ahmad, T.; Jin, L.; Lin, L.; Tang, G. Skeleton-based action recognition using sparse spatio-temporal GCN with edge effective resistance. *Neurocomputing* **2021**, *423*, 389–398. [CrossRef]
20. Yan, S.; Li, Z.; Xiong, Y.; Yan, H.; Lin, D. Convolutional sequence generation for skeleton-based action synthesis. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019.
21. Junejo, I.N.; Dexter, E.; Laptev, I.; Pérez, P. Cross-view action recognition from temporal self-similarities. In Proceedings of the Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, 12–18 October 2008.
22. Kitt, B.M.; Rehder, J.; Chambers, A.D.; Schonbein, M.; Lategahn, H.; Singh, S. Monocular visual odometry using a planar road model to solve scale ambiguity. In Proceedings of the Proceedings of 5th European Conference on Mobile Robots (ECMR '11), Örebro, Sweden, 7–9 September 2011.
23. Yang, J.; Lu, H.; Li, C.; Hu, X.; Hu, B. Data Augmentation for Depression Detection Using Skeleton-Based Gait Information. *arXiv* **2022**, arXiv:2201.01115. [CrossRef] [PubMed]
24. Kim, H.-J.; Kim, H.; Park, J.; Oh, B.; Kim, S.-C. Recognition of Gait Patterns in Older Adults Using Wearable Smartwatch Devices: Observational Study. *J. Med. Internet Res.* **2022**, *24*, e39190. [CrossRef] [PubMed]
25. Rhif, M.; Wannous, H.; Farah, I.R. Action recognition from 3d skeleton sequences using deep networks on lie group features. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018.
26. Vemulapalli, R.; Arrate, F.; Chellappa, R. Human action recognition by representing 3d skeletons as points in a lie group. In Proceedings of the IEEE conference on COMPUTER Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014.
27. Lyu, H.; Huang, D.; Li, S.; Ng, W.W.; Ma, Q. Multiscale echo self-attention memory network for multivariate time series classification. *Neurocomputing* **2023**, *520*, 60–72. [CrossRef]
28. Kim, H.; Lee, H.; Park, J.; Paillat, L.; Kim, S.C. Vehicle Control on an Uninstrumented Surface with an Off-the-Shelf Smartwatch. *IEEE Trans. Intell. Veh.* **2023**, *8*, 3366–3374. [CrossRef]
29. Lee, K.-W.; Kim, S.-C.; Lim, S.-C. DeepTouch: Enabling Touch Interaction in Underwater Environments by Learning Touch-Induced Inertial Motions. *IEEE Sens. J.* **2022**, *22*, 8924–8932. [CrossRef]
30. Perol, T.; Gharbi, M.; Denolle, M. Convolutional neural network for earthquake detection and location. *Sci. Adv.* **2018**, *4*, e1700578. [CrossRef] [PubMed]
31. Meng, F.; Liu, H.; Liang, Y.; Tu, J.; Liu, M. Sample fusion network: An end-to-end data augmentation network for skeleton-based human action recognition. *IEEE Trans. Image Process.* **2019**, *28*, 5281–5295. [CrossRef] [PubMed]
32. Song, S.; Lan, C.; Xing, J.; Zeng, W.; Liu, J. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.
33. Li, C.; Wang, P.; Wang, S.; Hou, Y.; Li, W. Skeleton-based action recognition using LSTM and CNN. In Proceedings of the 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Hong Kong, China, 10–14 July 2017.
34. Liu, J.; Shahroudy, A.; Xu, D.; Kot, A.C.; Wang, G. Skeleton-based action recognition using spatio-temporal LSTM network with trust gates. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 3007–3021. [CrossRef] [PubMed]
35. Chen, D.; Zhang, T.; Zhou, P.; Yan, C.; Li, C. OFPI: Optical Flow Pose Image for Action Recognition. *Mathematics* **2023**, *11*, 1451. [CrossRef]
36. Supratak, A.; Dong, H.; Wu, C.; Guo, Y. DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2017**, *25*, 1998–2008. [CrossRef] [PubMed]
37. Mazzia, V.; Angarano, S.; Salvetti, F.; Angelini, F.; Chiaberge, M. Action Transformer: A self-attention model for short-time pose-based human action recognition. *Pattern Recognit.* **2022**, *124*, 108487. [CrossRef]
38. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.