Local Spherical Harmonics Improve Skeleton-Based Hand Action Recognition

Katharina Prasse[1] [0009−0003−9502−1313],
Steffen Jung[1,2] [0000−0001−8021−791X], Yuxuan Zhou[3] [0000−0002−7688−803X], and
Margret Keuper[1,2] [0000−0002−8437−7993]

[1] University of Siegen, 57076 Siegen, Germany katharina.prasse@uni-siegen.com
[2] Max Planck Institute for Informatics, Saarland Informatics Campus, Germany
[3] University of Mannheim, 68131 Mannheim, Germany

Abstract. Hand action recognition is essential. Communication, human-robot interactions, and gesture control are dependent on it. Skeleton-based action recognition traditionally includes hands, which belong to the classes which remain challenging to correctly recognize to date. We propose a method specifically designed for hand action recognition which uses relative angular embeddings and local Spherical Harmonics to create novel hand representations. The use of Spherical Harmonics creates rotation-invariant representations which make hand action recognition even more robust against inter-subject differences and viewpoint changes. We conduct extensive experiments on the hand joints in the First-Person Hand Action Benchmark with RGB-D Videos and 3D Hand Pose Annotations, and on the NTU RGB+D 120 dataset, demonstrating the benefit of using Local Spherical Harmonics Representations. Our code is available at https://github.com/KathPra/LSHR_LSHT.

Keywords: Hand Action Recognition · Spherical Harmonics · Rotation Invariance · Relative Angular Embeddings.

1 Introduction

Hand actions are everywhere. They can be observed during conversations, and provide valuable information about the atmosphere, hierarchy, backgrounds, and emotions. Furthermore, their fine-grained differences make hand actions a challenging recognition task. Skeleton-based action recognition naturally contains hand actions, which remain challenging to distinguish between. Especially human-computer interactions require the accurate recognition of hand actions in order to understand e.g. air quotes or the thumbs-up gesture. When hand action recognition is further optimized, a vast field of applications stands to benefit from it; hand motion understanding in the medical field, gesture control, and robotics are mere examples. Moreover, the increasing number of online interactions creates many scenarios when only a subset of the body joints is available for action recognition. Out of all body joints, hand joints are most likely to be included in digital recordings and have an expressive nature. We thus focus on

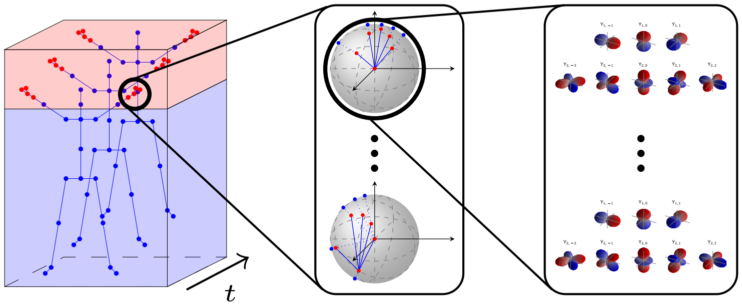


Fig. 1. Hands should receive particular emphasis as they contain the highest joint density and their mutual interaction is key to recognising hand actions. Hand joints (left) are first depicted as local spherical coordinates (middle) before being represented in terms of their Local Spherical Harmonics (right). The hand joint representation is then fed into the model as additional input.

hand actions explicitly and our proposed method allows to better distinguish between them. Since hands are very flexible, their embedding can benefit majorly from a local, relative description.

Action recognition is an active field of research since the late 1990s [26]. First advances into the field employed Recurrent Neural Networks [RNNs], among others Long Short-Term Memory Networks [LSTM], to capture the action over time [6,51,47]. Later, CNNs were fed action as a 2D map including semantic information as features e.g., joint type, time stamp, and position [19,48,42,24]. In recent years, Graph Convolutional Networks [GCN] are frequently employed and show a strong performance [32,44,31,3]. In contrast to CNNs, GCNs can use the inherent structure of the skeleton data. The joints are represented as nodes in the graph and the edges indicate their physical connections [31] or correlations [3]. Action recognition uses various forms of input data, ranging from RGB+Depth video data to skeleton joint Cartesian coordinates extracted from videos. Using skeleton data increases model efficiency and enhances robustness against variations in viewpoint and appearance [3,31,49]. In general, action recognition greatly benefits from understanding hand motion [36,11], which motivates us to focus particularly on hand joints. We describe each hand joint relative to the remaining hand joints. This way, the inter-joint relationships are explicitly included in the data and do not have to be inferred by the model.

We propose to represent fine-grained motion through angular embeddings and local Spherical Harmonics, as visualized in Figure 1. We hypothesize that local spherical representations are better suited than Cartesian coordinates for several reasons: (1) Hand joints are very close in absolute position. When depicting them in terms of their relative position, small changes become more apparent. (2) The mutual interaction between fingers within each hand is one of the most important cues for action recognition. (3) Hands are very flexible. Considering their relative joint positions in terms of spherical projections facilitates robustness to slight motion variations. In previous works, cylindrical [41] coor-

dinates have been shown to improve detection accuracy. We propose an angular embedding based on local Spherical Harmonics representations [12] as we find it suitable for describing the relative positions of the hand joints, i.e. modelling inter-joint relationships. Further, the data transformation makes the representation more robust against inter-subject differences and viewpoint changes, as it is partially rotation-invariant. We further propose hand joints' Local Spherical Harmonic Transforms [12], a fully rotation-invariant representation. To our knowledge, we are the first to employ Spherical Harmonics in the field of hand action recognition.
This study's main contributions can be summarized as follows:

1. We depict each hand joint in terms of its local neighbourhood to enable better hand-action recognition.
2. We combine angular embeddings with the standard input as an explicit motion description to better represent the hand's inter-joint relationships.
3. We combine Spherical Harmonic Transforms with the standard input as a rotation-invariant motion description to increase the robustness of hand action recognition against viewpoint or orientation changes.
4. We show that angular representations, explicitly representing inter-joint relations, can improve the egocentric hand action recognition on FPHA [11] by a significant margin.
5. We further show that angular spherical embeddings can also be leveraged by classical human skeleton-based action recognition models such as CTR-GCN [3], improving hand-based action recognition accuracy.

2 Related Work

2.1 Skeleton-Based Action Recognition

Skeleton-based action recognition initially employed Recurrent Neural Networks (RNN) to learn features over time [6,51,47]. Long Short-Term Memory Networks (LSTMs) were popular as they were able to regulate learning over time. Furthermore, Temporal State-Space Models (TF) [10], Gram Matrices [50], and Temporal Recurrent Models (HBRNN) [6] were popular choices. Prior to using learned features, hand-crafted features were employed and achieved competitive levels of accuracy compared to modern methods [37,11]. Lastly, Key-pose models such as Moving Pose made use of a modified kNN clustering to detect actions [46], whereas Hu et al. proposed a joint heterogeneous feature learning framework [16]. Graph Convolution Networks (GCNs) for skeleton-based action recognition have been proposed e.g., in [25,30,34,44,45] and are defined directly on the graph, in contrast to Convolutional Neural Networks (CNNs). GCNs can be divided into two categories, i.e. spectral [1,5,21,14] and spatial [44,3] methods. While the research is mainly pushed in the direction of GCNs, CNNs and transformers remain part of the scientific discourse on action recognition [7,15,43].

While different methods have been proposed over time, one main tendency can be distilled from previous research: It is beneficial in terms of model accuracy to incorporate additional information [46,31,49,3,16,28]. Qin et al. [28] have included angles between several body joints into their model. Recently, multimodal ensembles have been shown to increase the overall accuracy [3,4,28]. While skeleton-based action recognition mainly focuses on the whole body, e.g., the popular NTU RGB-D 120 benchmark dataset [NTU 120] [23], several researchers have seen the merit in focusing on hand actions [36,11], and among them, Garcia et al. published the First-Person Hand Action Benchmark [FPHA] [11]. This dataset consists of egocentric recordings of daily hand actions in three categories, i.e. *social*, *kitchen* and *office actions*. Both datasets are included in this research. It is our goal to advance skeleton-based hand action recognition. We thus limit the review of skeleton-based action literature to the most relevant works.

2.2 Frequency Domain Representations for Action Recognition

Representing the skeleton information in the frequency instead of the spatial domain has two main advantages. Firstly, noise can be removed more easily and secondly, rotation invariance can be introduced. When transferring data from the spatial to the frequency domain, the data is represented as a sum of frequencies. When the input's data format is Cartesian coordinates, a Fourier Transformation is commonly used; when the input is spherical coordinates, Spherical Harmonics can be employed. Besides the different input formats, these methods are equivalent. They take a function and map it from the spatial domain to the frequency domain by approximating it as sums of sinusoids [2]. Many previous works on action recognition have employed Fourier Transformation [29,41,39].

Removing noise from human recording is always beneficial, e.g. when a waving hand is shaking, excluding the shaking, a high-frequency motion, makes it easier to correctly identify the action. Since the frequency domain data is complex, it has to be transformed in order to be fed into standard neural networks. Using the magnitude, a representation invariant to rotations, renders data normalization superfluous [35] and can improve recognition accuracies in use cases with varying viewpoints.

2.3 Spherical Harmonics in 3D Point Clouds

We are the first to employ Spherical Harmonics in skeleton-based action recognition to our knowledge. In 3D Point Clouds however, Spherical Harmonics are explored to make data representations robust or equivariant to rotations [8,27,33]. Esteves et al. [8] map 3D point clouds to spherical functions and propose spherical convolutions, which are robust against random rotations. Similarly, Spezialetti et al. propose a self-supervised learning framework for using spherical CNNs to learn objects' canonical surface orientation in order to detect them independent of SO(3) rotations [33]. Poulard et al. [27] propose spherical kernels for convolution which directly operate in the point clouds. Li et al. [22] compare several aforementioned methods [8,27,33] in settings where the data points are either rotated around one or all three axes during training. They highlight the strong performance of SPH-Net [27] whose accuracy remains on the same level independent of the number of axis rotated around during training. Fang et al. [9] points out the superiority of common CNNs over Spherical CNNs. Hence, we employ Spherical Harmonics as embeddings to which standard CNNs or GCNs can be applied.

3 Method

We propose multi-modal hand joint representations from which our model learns feature representations. Skeleton-based action recognition generally uses the five-dimensional Cartesian coordinates $X \in \mathbb{R}^{N \times M \times C \times T \times V}$, where N describes the batch size, M the number of persons in the action recording, C the channel dimension, T the number of frames, and V the number of joints. We combine the Cartesian coordinates with a local embedding, created using Spherical Harmonics basis functions.

3.1 Local Spherical Coordinates

We first incorporate the hand joints' local neighbourhood by using each joint as the center of the coordinate system once while depicting the other joints' position relative to the center joint, as shown in Figure 2. We combine the local representations of all hand joints to capture fine-grained motion e.g., the differences between "making ok sign" and "making peace sign", even though the absolute positions of the thumb and index finger are very similar between both actions.

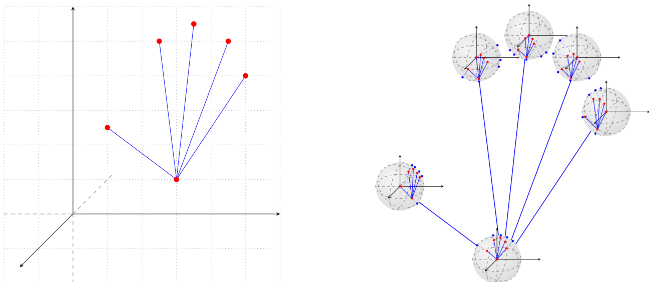


Fig. 2. Conversion between Cartesian global coordinates (left) and local spherical coordinates (right), where all coordinates are computed relative to each other joint. The global coordinates are Cartesian, while the local coordinates are spherical.

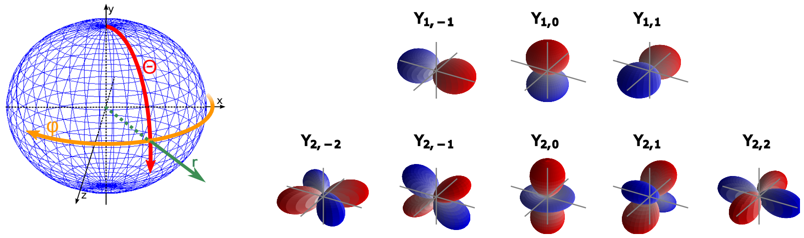
This transforms $X \in \mathbb{R}^{N \times M \times C \times T \times V}$ into $X_{loc} \in \mathbb{R}^{N \times M \times C \times T \times V \times V}$ where the last dimension contains the joint's local neighbourhood relative to its center joint's position. We compute the local neighbourhood for all hand joints, as we expect to gain the biggest advantage from investigating their local neighbourhood.

Moreover, we convert the local coordinates from Cartesian to Spherical coordinates as their angles are more suitable for describing hand motion than positions. Spherical coordinates consist of three values (r, θ, ϕ) and are used to describe point positions in 3D space, as visualized in Figure 3 (a). They can be computed using Cartesian coordinates as inputs. The radius $r \in [0, \infty)$ is defined as the length of the vector from the origin to the coordinate point, i.e. $r = \sqrt{x^2 + y^2 + z^2}$. The polar angle $\theta = \arctan \frac{\sqrt{x^2+y^2}}{z}$ describes the point's latitude and is defined for $\theta \in [0, \pi]$. The azimuthal angle $\phi = \arctan \frac{y}{x}$ describes the point's longitude and is defined for $\phi \in [0, 2\pi]$. The azimuth describes the point's location in the xy-plane relative to the positive x-axis.

3.2 Spherical Harmonics based Hand Joint Representations

We propose angular embeddings and Spherical Harmonics as novel representations for hand action recognition, which we realize through local spherical coordinates. All Spherical Harmonics-based representations naturally offer a coarse to fine representation while being interpretable in terms of frequency bands developed on a sphere. This includes representations using the Spherical Harmonics basis functions and the full Spherical Harmonic Transform. Moreover, the magnitude of Spherical Harmonic Transforms is rotation-invariant which is a helpful property for hand action recognition [12].

Figure 3 visualizes how we first transform the hand coordinates from Cartesian to spherical coordinates (a) before representing them in terms of their Spherical Harmonic basis functions (b).



(a) (b)

Fig. 3. (a) Conversion between Cartesian coordinates (x, y, z) and spherical coordinates (r, θ, ϕ). Spherical coordinates consist of the radius r, the polar angle θ and the azimuthal angle ϕ; (b) Visualization of the real part of Spherical Harmonics, where red indicates positive values while blue indicates negative values. The distance from the origin visualizes the magnitude of the Spherical Harmonics in the respective angular direction.

A function on the sphere can be represented in Spherical Harmonics as

$$f(\theta, \phi) = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{m=\ell} a_m^\ell Y_m^\ell(\theta, \phi) \tag{1}$$

with

$$Y_{\ell,m}(\theta,\phi) = \sqrt{\frac{(\ell-m)!(2\ell+1)}{(\ell+m)!4\pi}}e^{im\phi}P_{\ell}^{m}cos\theta \tag{2}$$

where $Y_{m}^{\ell}(\theta,\phi)$ are the Spherical Harmonics basis functions and the *!* indicates a factorial operation. Spherical Harmonics take two spherical coordinates as input, the azimuth θ and the polar angle ϕ. They further include the associated Legendre polynomial P_{ℓ}^{m} and have the parameter degree ℓ, in accordance to which the order m is set, i.e. $m \in [-\ell,\ell]$; both ℓ and m are real numbers. The magnitude of the Local Spherical Harmonics basis function Representation [LSHR] is invariant against rotations around the y-axis, since the magnitude of the complex exponential function is 1. The magnitude of the Local Spherical Harmonic Transforms [LSHT] is SO(3) rotation invariant. We argue that such representations are beneficial when learning to recognize hand actions across different views. We further expect these representations to support recognition in spite of hand orientation differences between subjects.

The Spherical Harmonics for $\ell \in \{1,2\}$ are employed for skeleton-based hand action recognition, as visualized in Figure 3 (b). The inclusion of $\ell = 0$ does not contain action-specific information, and we assume that all relevant information is contained within the first two bands, i.e. $\ell \leq 2$. The parameter ℓ is defined as the number of nodal lines, or bands, thus with larger ℓ, higher frequencies are represented; the removal of high frequencies reduces the noise in the data.

The chosen Spherical Harmonics representation is stacked along the channel dimension and concatenated to the original input. Since Spherical Harmonics are complex numbers, i.e. $x = a+bi$, they cannot directly be fed into standard neural networks. They need to either be represented by their real and imaginary parts, or by their magnitude and phase. When using a single part, the representation is no longer complete and the input cannot be recovered entirely. The real and imaginary parts are a complete representation of the complex number. When only the real or only the imaginary part is used, the representation is no longer complete, mathematically speaking. The magnitude has the property of being rotation-invariant [12].

3.3 Models

We include angular embeddings, a local Spherical Harmonics representation (LSHR), and Spherical Harmonic Transforms (LSHT) both in a simple baseline model (GCN-BL) and an advanced model (CTR-GCN), both proposed by Chen et al. [3]. The Channel-wise Topology Refinement Graph Convolution Model (CTR-GCN), proposed by Chen et al., achieves state-of-the-art results with a clean model architecture and a small number of epochs [3]. Both models are described in the subsequent paragraph, closely following Chen et al.'s description. They each consist of 10 layers, where each layer contains both a spatial graph convolutional network (GCN) module and a temporal convolutional network (TCN) module. While the CTR-GCN model learns a non-shared topology for the channels and dynamically infers joint relations during inference, the GCN-BL model learns a static topology shared between all channels. Further implementation details can be found in the supplementary material.

3.4 Evaluation

The angular embeddings and Spherical Harmonics are evaluated in terms of their accuracy improvement both on all body joints [Imp.] and with a focus on hand joints [Hand Imp.]. During training, we randomly rotate the input data.

The local Spherical Harmonics representations are concatenated to the standard model input, the Cartesian coordinates, before the first layer. This causes a negligible increase in the number of parameters since only the input dimensionality is modified but not the layer outputs. When evaluating the effect of hand joints' angular embeddings or Spherical Harmonics, the model is compared to the original implementation. In order to maintain the data structure of $X \in \mathbb{R}^{N\times M\times C\times T\times V}$, zeros are inserted for all non-hand joints. Our ablation studies contain a model trained exclusively on angular embeddings evaluated on FPHA [11]. Furthermore, the NTU120 [23] ablations include a baseline model of identical dimensionality, where random numbers replace the angular embeddings.

4 Experiments

4.1 Datasets

The benefits of Local Spherical Harmonics representations (LSHR) and Local Spherical Harmonic Transforms (LSHT) of hand joints are shown on two datasets: Mainly, the First-Person Hand Action Benchmark is assessed. It is created by Garcia-Hernando et al. [11] and consists of shoulder-mounted camera recordings of six subjects each performing 45 action classes four to six times. The dataset contains 1175 skeleton recordings over time, including the wrist and four joints for each finger as shown in Figure 4. This dataset is split 1:1 into train and test sets. The viewpoints differ between recordings, which makes the recognition simultaneously Cross-Subject and Cross-Setup tasks.

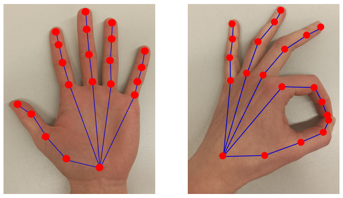


Fig. 4. Visualization of hand joints in First-Person Hand Action Benchmark [11] (own visualization).

Secondly, the benchmark dataset NTU RGB+D 120 [23] is employed. It contains 120 action classes, which can be split into daily, medical, and criminal actions [23]. The dataset contains 114,480 videos featuring 106 distinct subjects in 32 scenarios (camera height and distance to subject). The dataset can be evaluated in two settings, i.e. Cross-Subject and Cross-Setup. In the Cross-Subject evaluation, the 106 subjects are split into the train and test set. In the Cross-Setup setting, some recording angles and backgrounds are exclusively used for training, while others are used for testing. Both cases have the same train-test split, i.e. 1:1. The computation of local Spherical Harmonics was exclusively done for the eight hand joints, i.e. both left and right wrist, hand, thumb, and finger.

4.2 Implementation Details

Experiments are conducted on NVIDIA Tesla V100 GPUs with the PyTorch learning framework. The models proposed by Chen et al. [3], GCN-BL and CTR-GCN, have the same hyperparameter settings, i.e. SGD with a momentum of 0.9, weight decay of 0.0004, and training for 65 epochs including 5 warm-up epochs. The learning rate is initialized with 0.1 and decays by a factor of 0.1 at epochs 35 and 55. The batch size is 64 and all action recordings are resized to 64 frames. For the FPHA dataset, a batch size of 25 was chosen, due to the smaller size of this dataset. Experiments are run on AMD Ryzen 9 5900x.

4.3 Experimental Results

Experiments on the First-Person Hand Action Benchmark (FPHA) demonstrate the strength of angular embeddings, a local Spherical Harmonics representation [LSHR], as shown in Table 1. The model accuracy increases for both the GCN-BL [3] and the CTR-GCN [3] model. We ablate our method by selecting angular embeddings as the sole model input; they outperform the original model by a large margin. The original CTR-GCN model suffers from severe overfitting causing their test accuracy to remain below the GCN-BL model's test accuracy. The original GCN-BL model overfits less and thus has a higher accuracy.

Method	Param.	Rotation Invariance	Acc (%)
GCN BL	2.1 M	✗	80.52
+ LSHR (R&I)	2.1 M	✗	88.35
+ **LSHR (M)**	2.1 M	✓	**89.04**
+ LSHT (M)	2.1 M	✓	87.30
excl. LSHR (M)	2.1 M	✓	83.83*
CTR-GCN	1.4 M	✗	74.26
+ LSHR (R&I)	1.5 M	✗	90.26
+ **LSHR (M)**	1.5 M	✓	**92.52**
+ LSHT (M)	1.4 M	✓	89.04
excl. LSHR (M)	1.5 M	✓	85.57*

Table 1. Evaluation of the GCN-BL [3] and CTR-GCN [3] Model with Local Spherical Harmonics Representations [LSHR] and Local Spherical Harmonic Transforms [LSHT] evaluated on FPHA [11] using the real and imaginary parts (R& I) or the rotation-invariant magnitude (M), compared against the original model, and LSHR exclusively (*). Our Method significantly increases accuracy.

As Table 1 shows, the largest increase over the original model is achieved with the CTR-GCN model, using the rotation-invariant magnitude of the angular embedding, a local Spherical Harmonics representation [LSHR] (+18%). Local Spherical Harmonic Transforms [LSHT] also outperform the original model by a large margin (+15%). The results indicate that learning a local relative hand representation directly from the basis functions is favourable over employing their Spherical Harmonics. The performance of the GCN-BL model differs from the CTR-GCN model, possibly due to the design of the model. While GCN-BL has a shared topology, the CTR-GCN model uses a channel-wise topology. Further angular embedding formats are reported in the supplementary material.

When comparing our method to others evaluated on the first-person hand action benchmark, we clearly outperform previous models by a large margin, as shown in Table 2. Our accuracy is 7% higher than the best previously reported model using the Gram Matrix [50].

Method	Acc (%)
1-layer LSTM [11]	78.73
2-layer LSTM [11]	80.14
Moving Pose [46]	56.34
Lie Group [38]	82.69
HBRNN [6]	77.40
Gram Matrix [50]	85.39
TF [10]	80.69
JOULE-pose [16]	74.60
TCN [20]	78.57
LEML [18]	79.48
SPDML-AIM [13]	78.40
SPDNet [17]	83.79
SymNet-v1 [40]	81.04
SymNet-v2 [40]	82.96
GCN-BL [3]	80.52
CTR-GCN [3]	74.26
Ours	**92.52**

Table 2. Model Evaluation on First-Person Hand Action Benchmark [11]. Our model using the magnitude of the angular embeddings outperforms all other models.

Our method is clearly superior when evaluated on a hands-only dataset such as FHPH [11]. Furthermore, when evaluating it on a body dataset, it remains superior to the original model's performance. We evaluate angular embeddings, a local Spherical Harmonics representation [LSHR], on the NTU120 dataset. For both the Cross-Subject and Cross-Setup benchmarks, we assess various input formats. These benchmarks give further insights into the performance of rotation-invariant features. Within the Cross-Subject benchmark, we can evaluate the robustness of angular embeddings and Spherical Harmonics against inter-subject differences. The Cross-Setup benchmark allows us to evaluate the rotation-invariance of our method explicitly. The results using the GCN-BL model confirm the findings from the FPHA dataset, that angular embeddings, a local Spherical Harmonics representation [LSHR], increase the overall model accuracy (see Table 3). This improvement is even larger when exclusively assessing hand-related action classes [Hand Imp.]. The full list of hand vs. non-hand-related action classes can be found in the supplementary material. In the Cross-Subject setting, the full spectrum, i.e. real and imaginary parts of angular embeddings, induces the largest accuracy increase, as expected. In this case, the use of a rotation-invariant representation is not as advantageous for cross-subject action recognition. In the Cross-Setup setting, the largest accuracy increase was achieved when using the magnitude of the angular embedding, in line with our expectations, that the rotation invariant representation is well suited for this setting. The phase information reduces model accuracy, as it contains rotation information and thus hinders recognition when different viewpoints are compared. When investigating the differences between local angular embeddings, a local Spherical Harmonics representation [LSHR], and Spherical Harmonic Transforms [LSHT] on the NTU120 dataset, it becomes apparent, that the GCN-BL model benefits from the inclusion of Local angular Spherical Harmonics Representations [LSHR]. The CTR-GCN model, however, achieves higher levels of accuracy when the local Spherical Harmonic Transforms [LSHT] are employed.

Dataset	Format	Rand. BL Acc. (%)	Ours Acc. (%)	Imp.	Hand Imp.
X-Sub	Real	83.89	84.38	↑ +0.5	↑ +0.9
	Imaginary	83.89	84.39	↑ +0.5	**↑ +1.0**
	Magnitude	83.89	84.20	↑ +0.3	↑ +0.3
	Real & Imag.	83.18	84.01	**↑ +0.8**	**↑ +1.0**
	Mag. & Phase	83.18	83.72	↑ +0.5	↑ +0.7
X-Set	Real	85.62	85.93	↑ +0.3	↑ +0.5
	Imaginary	85.62	85.98	↑ +0.4	↑ +0.7
	Magnitude	85.62	86.21	**↑ +0.6**	**↑ +1.0**
	Real & Imag.	85.49	85.91	↑ +0.4	↑ +0.7
	Mag. & Phase	85.49	85.39	↓ −0.1	↓ −0.5

Table 3. Format Comparison of the joint modality using GCN-BL with angular embeddings on NTU120. Our method increases overall accuracy (Imp.) and hand-related action accuracy (Hand Imp.) compared to a random baseline (Rand. BL).

Dataset	Model	Joint	Original Acc. (%)	LSHR (Ours) Acc. (%)	LSHT (Ours) Acc. (%)
X-Sub	GCN-BL	Loc.	83.75	84.20 **(↑ 0.5)**	84.03 (↑ 0.3
		Vel.	80.30	80.53 **(↑ 0.2)**	80.32 (±0)
	CTR-GCN	Loc.	85.08	85.31 **(↑ 0.2)**	85.27 **(↑ 0.2)**
		Vel.	81.12	81.45 (↑ 0.3)	81.53 **(↑ 0.4)**
X-Set	GCN-BL	Loc.	85.64	86.21 **(↑ 0.6)**	85.75 (↑ 0.1)
		Vel.	82.25	82.41 **(↑ 0.2)**	82.25 (±0)
	CTR-GCN	Loc.	86.76	86.63 (↓ 0.1)	87.01 **(↑ 0.3)**
		Vel.	83.12	83.32 (↑ 0.2)	83.42 **(↑ 0.3)**

Table 4. Single modality evaluation using local Spherical Harmonics representations [LSHR], and rotation-invariant Local Spherical Harmonic Transforms [LSHT] on both NTU120 benchmarks [23]. Rotation-invariant hand joint representations increase the model's accuracy.

The ensemble performance is increased by the inclusion of angular embeddings, local Spherical Harmonics representations [LSHR], or Spherical Harmonic Transforms [LSHT] as reported in Table 5. The model using the full Local Spherical Harmonic Transforms [LSHT] outperforms the other models by a small margin. As our method only includes joints, the bone modalities are always taken from the original model and do not contain any angular embeddings or Spherical Harmonic Transforms.

Modality	**Method**	**NTU RGB+D 120**	
		X-Sub (%)	**X-Set(%)**
Loc.	CTR-GCN	**88.8**	90.0
	LSHR (full spectrum)	88.6	90.1
	LSHR (rotation invariant)	**88.8**	90.0
	LSHT (rotation invariant)	**88.8**	**90.2**
Loc. & Vel.	CTR-GCN	89.1	90.6
	LSHR (full spectrum)	88.9	90.5
	LSHR (rotation invariant)	89.1	90.6
	LSHT (rotation invariant)	**89.2**	**90.7**

Table 5. Ensemble Evaluation of the CTR-GCN model in its original version compared with two local Spherical Harmonics representations [LSHR], full-spectrum (real and imaginary part) and rotation invariance (magnitude), and the full Spherical Harmonic Transforms [LSHT]. LSHT outperforms all other models by a small margin.

We conduct further experiments on the NTU120 benchmarks comparing rotation-invariant and complete angular embeddings, a local Spherical Harmonics representation [LSHR], for which we report both the hand accuracy and the overall accuracy in the supplementary material. As our method focuses on hand joints, the hand accuracy increases are larger than the overall accuracy increases, when compared to the original model.

5 Discussion & Conclusion

Our rotation-invariant joint representations on the basis of local Spherical Harmonics improve skeleton-based hand action recognition. Their inclusion allows us to better distinguish between fine-grained hand actions. Both the theoretical framework and the experimental results affirm that hand joints' local Spherical Harmonics improve skeleton-based hand action recognition. We clearly outperform other methods the hands-only First-Person-Hand-Action-Benchmark [11]. The evaluations on the NTU120 Cross-Subject and Cross-Setup benchmarks further confirm our findings. Our experiments have shown that rotation-invariant hand representations increase robustness against inter-subject orientation differences and viewpoint changes, thus resulting in higher accuracy levels. On the NTU120 dataset, the overall best model performance is achieved in the Cross-Setup setting using the joint location modality, as expected due to the rotation invariance of the features. The Cross-Subject setting benefited more from the rotation invariance than expected, hinting that inter-subject differences can be reduced using angular embeddings. Thinking of the action "making a peace sign" exemplifies this, as different subjects have slightly different orientations of their hands. We aim to open the door for future research in the adaptation of angular embedding to this and other data modalities. Furthermore, the number of hand joints included in the NTU120 is four per hand, which appears to make the sampling of additional data points a meaningful investigation.

Acknowledgements

This work is partially supported by the BMBF project 16DKWN027b Climate Visions and DFG research unit 5336 Learning2Sense. All experiments were run on the computational resources of the University of Siegen and the Max Planck Institute for Informatics. We would like to thank Julia Grabinski for her guidance.

References

1. Bruna, J., Zaremba, W., Szlam, A., LeCun, Y.: Spectral networks and locally connected networks on graphs. arXiv preprint arXiv:1312.6203 (2013)
2. Brunton, S.L., Kutz, J.N.: Fourier and Wavelet Transforms, p. 47–83. Cambridge University Press (2019). https://doi.org/10.1017/9781108380690.003
3. Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., Hu, W.: Channel-wise topology refinement graph convolution for skeleton-based action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13359–13368 (2021)
4. Chi, H.g., Ha, M.H., Chi, S., Lee, S.W., Huang, Q., Ramani, K.: Infogcn: Representation learning for human skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20186–20196 (2022)
5. Defferrard, M., Bresson, X., Vandergheynst, P.: Convolutional neural networks on graphs with fast localized spectral filtering. Advances in neural information processing systems **29** (2016)
6. Du, Y., Wang, W., Wang, L.: Hierarchical recurrent neural network for skeleton based action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1110–1118 (2015)
7. Duan, H., Zhao, Y., Chen, K., Lin, D., Dai, B.: Revisiting skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2969–2978 (2022)
8. Esteves, C., Allen-Blanchette, C., Makadia, A., Daniilidis, K.: Learning so (3) equivariant representations with spherical cnns. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 52–68 (2018)
9. Fang, J., Zhou, D., Song, X., Jin, S., Yang, R., Zhang, L.: Rotpredictor: Unsupervised canonical viewpoint learning for point cloud classification. In: 2020 International Conference on 3D Vision (3DV). pp. 987–996. IEEE (2020)
10. Garcia-Hernando, G., Kim, T.K.: Transition forests: Learning discriminative temporal transitions for action recognition and detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 432–440 (2017)
11. Garcia-Hernando, G., Yuan, S., Baek, S., Kim, T.K.: First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In: Proceedings of Computer Vision and Pattern Recognition (CVPR) (2018)
12. Green, R.: Spherical harmonic lighting: The gritty details. In: Archives of the game developers conference. vol. 56, p. 4 (2003)

13. Harandi, M., Salzmann, M., Hartley, R.: Dimensionality reduction on spd manifolds: The emergence of geometry-aware methods. IEEE transactions on pattern analysis and machine intelligence **40**(1), 48–62 (2017)
14. Henaff, M., Bruna, J., LeCun, Y.: Deep convolutional networks on graph-structured data. arXiv preprint arXiv:1506.05163 (2015)
15. Hu, H., Dong, S., Zhao, Y., Lian, D., Li, Z., Gao, S.: Transrac: Encoding multi-scale temporal correlation with transformers for repetitive action counting. arXiv preprint arXiv:2204.01018 (2022)
16. Hu, J.F., Zheng, W.S., Lai, J., Zhang, J.: Jointly learning heterogeneous features for rgb-d activity recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5344–5352 (2015)
17. Huang, Z., Van Gool, L.: A riemannian network for spd matrix learning. In: Proceedings of the AAAI conference on artificial intelligence. vol. 31 (2017)
18. Huang, Z., Wang, R., Shan, S., Li, X., Chen, X.: Log-euclidean metric learning on symmetric positive definite manifold with application to image set classification. In: International conference on machine learning. pp. 720–729. PMLR (2015)
19. Ke, Q., Bennamoun, M., An, S., Sohel, F., Boussaid, F.: A new representation of skeleton sequences for 3d action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3288–3297 (2017)
20. Kim, T.S., Reiter, A.: Interpretable 3d human action analysis with temporal convolutional networks. In: 2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW). pp. 1623–1631. IEEE (2017)
21. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
22. Li, F., Fujiwara, K., Okura, F., Matsushita, Y.: A closer look at rotation-invariant deep point cloud analysis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16218–16227 (2021)
23. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.Y., Kot, A.C.: Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. IEEE transactions on pattern analysis and machine intelligence **42**(10), 2684–2701 (2019)
24. Liu, M., Liu, H., Chen, C.: Enhanced skeleton visualization for view invariant human action recognition. Pattern Recognition **68**, 346–362 (2017)
25. Liu, Z., Zhang, H., Chen, Z., Wang, Z., Ouyang, W.: Disentangling and unifying graph convolutions for skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 143–152 (2020)
26. Minami, K.i., Nakajima, H., Toyoshima, T.: Real-time discrimination of ventricular tachyarrhythmia with fourier-transform neural network. IEEE transactions on Biomedical Engineering **46**(2), 179–185 (1999)
27. Poulenard, A., Rakotosaona, M.J., Ponty, Y., Ovsjanikov, M.: Effective rotation-invariant point cnn with spherical harmonics kernels. In: 2019 International Conference on 3D Vision (3DV). pp. 47–56. IEEE (2019)
28. Qin, Z., Liu, Y., Ji, P., Kim, D., Wang, L., McKay, R., Anwar, S., Gedeon, T.: Fusing higher-order features in graph neural networks for skeleton-based action recognition. IEEE Transactions on Neural Networks and Learning Systems (2022)
29. Rodriguez, M.D., Ahmed, J., Shah, M.: Action mach a spatio-temporal maximum average correlation height filter for action recognition. In: 2008 IEEE conference on computer vision and pattern recognition. pp. 1–8. IEEE (2008)
30. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: CVPR (2019)

31. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Adasgn: Adapting joint number and model size for efficient skeleton-based action recognition. In: ICCV (2021)
32. Si, C., Jing, Y., Wang, W., Wang, L., Tan, T.: Skeleton-based action recognition with spatial reasoning and temporal stack learning. In: Proceedings of the European conference on computer vision (ECCV). pp. 103–118 (2018)
33. Spezialetti, R., Stella, F., Marcon, M., Silva, L., Salti, S., Di Stefano, L.: Learning to orient surfaces by self-supervised spherical cnns. Advances in Neural information processing systems **33**, 5381–5392 (2020)
34. Tang, Y., Tian, Y., Lu, J., Li, P., Zhou, J.: Deep progressive reinforcement learning for skeleton-based action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5323–5332 (2018)
35. Temerinac, M., Reisert, M., Burkhardt, H.: Invariant features for searching in protein fold databases. International Journal of Computer Mathematics **84**(5), 635–651 (2007)
36. Trivedi, N., Thatipelli, A., Sarvadevabhatla, R.K.: Ntu-x: An enhanced large-scale dataset for improving pose-based recognition of subtle human actions. In: Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing. pp. 1–9 (2021)
37. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. arXiv preprint arXiv:1710.10903 (2017)
38. 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. pp. 588–595 (2014)
39. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Mining actionlet ensemble for action recognition with depth cameras. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1290–1297. IEEE (2012)
40. Wang, R., Wu, X.J., Kittler, J.: Symnet: A simple symmetric positive definite manifold deep learning method for image set classification. IEEE Transactions on Neural Networks and Learning Systems **33**(5), 2208–2222 (2021)
41. Weinland, D., Ronfard, R., Boyer, E.: Free viewpoint action recognition using motion history volumes. Computer vision and image understanding **104**(2-3), 249–257 (2006)
42. Weng, J., Liu, M., Jiang, X., Yuan, J.: Deformable pose traversal convolution for 3d action and gesture recognition. In: Proceedings of the European conference on computer vision (ECCV). pp. 136–152 (2018)
43. Xu, K., Ye, F., Zhong, Q., Xie, D.: Topology-aware convolutional neural network for efficient skeleton-based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 2866–2874 (2022)
44. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Thirty-second AAAI conference on artificial intelligence (2018)
45. Ye, F., Pu, S., Zhong, Q., Li, C., Xie, D., Tang, H.: Dynamic gcn: Context-enriched topology learning for skeleton-based action recognition. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 55–63 (2020)
46. Zanfir, M., Leordeanu, M., Sminchisescu, C.: The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2752–2759 (2013)
47. 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. pp. 2117–2126 (2017)

48. Zhang, P., Lan, C., Xing, J., Zeng, W., Xue, J., Zheng, N.: View adaptive neural networks for high performance skeleton-based human action recognition. IEEE transactions on pattern analysis and machine intelligence **41**(8), 1963–1978 (2019)
49. Zhang, P., Lan, C., Zeng, W., Xing, J., Xue, J., Zheng, N.: Semantics-guided neural networks for efficient skeleton-based human action recognition. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1112–1121 (2020)
50. Zhang, X., Wang, Y., Gou, M., Sznaier, M., Camps, O.: Efficient temporal sequence comparison and classification using gram matrix embeddings on a riemannian manifold. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4498–4507 (2016)
51. Zhu, W., Lan, C., Xing, J., Zeng, W., Li, Y., Shen, L., Xie, X.: Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In: Proceedings of the AAAI conference on artificial intelligence. vol. 30 (2016)