Spatio-Temporal Hyperbolic Aggregation Neural Network for Human Action Recognition

Mohamed Sanim Akremi, Najett Neji, Hedi Tabia

Spatio-Temporal Hyperbolic Aggregation Neural Network for Human Action Recognition*

Mohamed Sanim Akremi[1] Najett Neji [2] Hedi Tabia[3]

***Abstract*— Human action recognition (HAR) is a critical task in the field of robotics. Traditionally, HAR methods rely on either perceptual features from RGB images or skeletal features. While RGB-based features are typically represented in 2D Euclidean space, few approaches differentiate between methods developed for RGB data and those for skeletal features, often treating both as Euclidean representations. This conventional approach, which typically leverages standard deep learning techniques, limits the descriptive power of skeletal data, which naturally exhibits a tree-like structure. In this paper, we introduce a novel framework that, for the first time, utilizes skeletal data while preserving its inherent structure to fully capture its descriptive potential. Our proposed deep neural network embeds skeletal joints into hyperbolic space, followed by a spatio-temporal processing framework that incorporates established transformations to optimize performance while maintaining the advantages of hyperbolic analysis. Extensive experiments on publicly available datasets, including UAV-Human, UAV-Gesture, and DHG 14/28, demonstrate that our approach achieves state-of-the-art results, underscoring its ability to enhance robotic systems' performance in dynamic environments.**

***Index Terms*— Hyperbolic Geometry, Skeletal-Based Features, Human Action Recognition, Spatio-Temporal Processing, Hyperbolic Neural Network.**

I. Introduction

Human Action Recognition (HAR) is essential in robotics as it allows robots to understand, interpret, and interact with human behavior in real-time. This capability is important for robots operating in dynamic settings where human collaboration is key, such as in healthcare [1], service [2], manufacturing, and autonomous systems [3].

Significant advancements in HAR have been achieved through various approaches. Recent methods typically rely on perceptual features from RGB images or skeletal features obtained through conventional deep learning techniques. Convolutional Neural Networks (CNNs) [4] and transformers [5] are particularly effective for learning perceptual features, making them well-suited for video-based HAR [6].

RGB-based features are usually represented in 2D Euclidean space, which aligns well with traditional machine learning methods suited for Euclidean structures. However, skeletal data, which inherently has a tree-like structure, requires specialized processing to fully leverage its descriptive power since it has been proved by recent studies that Euclidean space struggles to represent complex, hierarchical data. The use of skeletal features has been shown to be more effective for recognition tasks, particularly in capturing body dynamics and structural dependencies between joints. Many methods have been developed to address this, including geometric approaches [7]–[9] such as Symmetric Positive Definite (SPD) matrices, Lie group-based methods, and the Grassmann manifold. SPD matrices capture the covariance structure of features, providing robust action representations. Lie group-based methods model human actions as curves in a Lie group to capture 3D rigid body motions, while the Grassmann manifold represents subspaces in higher-dimensional space, which is beneficial for video-based HAR.

Despite their advancements, these methods face limitations, including the need for extensive training datasets, computational challenges with high-dimensional data, and insufficient capture of hierarchical structures in tree-like data such as skeletal data. To address these issues, recent research indicates that hyperbolic embeddings offer a promising solution by improving accuracy and extending model applicability to hierarchical structures [10], as they naturally preserve the correlations in tree-like data.

In this paper, we propose a novel neural network architecture named the Spatio-Temporal Hyperbolic Aggregation Neural Network (ST-HAgg-Net) for processing sequences of 3D joint coordinates. Our proposed network consists of three main modules. The first module, called the Hyp-Map component, performs hyperbolic mapping through temporal and spatial segmentation, embedding these segments into the Poincaré unit ball. The second module, known as the Spatial-Temporal Hyperbolic (ST-H) component, handles both spatial and temporal segmentation at local and global levels, aggregating these segments using the Möbius gyromidpoint to represent the entire sequence as a single vector. The final module, the Classification component, maps this vector into Euclidean space and applies fully-connected layers to carry out the classification task.

To gauge the performance and robustness of our approach, we subject it to rigorous evaluation using three benchmarks: UAV-Human dataset, UAV-Gesture dataset, and DHG 14/28 dataset. Remarkably, our methodology achieves state-of-the-art results on this benchmark datasets, signifying its prowess in action recognition within ground-camera and UAV-camera captured imagery.

This paper makes three major contributions: (1) We propose an effective model that operates on skeletal sequences

[1]Mohamed Sanim Akremi is a PhD student at IBISC laboratory University of Paris-Saclay mohamed.akremi@universite-paris-saclay.fr

[2]Najett Neji is Associate Professor of Embedded Systems for drones at Université d'Evry Val d'Essonne, Université Paris Saclay najett.neji@univ-evry.fr

[3]Hedi Tabia is a professor of Computer Science, Université d'Evry Val d'Essonne, Université Paris Saclay hedi.tabia@univ-evry.fr

for human action recognition. (2)We introduce a new hyperbolic neural network tailored for the spatio-temporal analysis of skeletal action sequences. The representational power of hyperbolic space, particularly its capacity to capture hierarchical and temporal dependencies, proves especially beneficial when adapted to spatio-temporal processes, leading to improved modeling of the structured dynamics of human actions and enhanced performance. (3) We demonstrate the efficiency of the proposed algorithm through rigorous evaluation on skeletal registration datasets.

II. Related works

A. Manifold-based approaches for Human action recognition

Few works in action recognition utilize skeletal data from depth-sensing cameras and explores non-Euclidean spaces. For example, elastic functional coding is used for analyzing Riemannian trajectories [11]. Lie group techniques are also prominent, such as Huang et al.'s approach, which integrates a RotMap layer to adaptively transform rotation matrices, a RotPooling layer for aggregating these matrices, and a LogMap Layer to complete the framework [8]. Additionally, Vemulapalli et al. [12] introduced a neural network that represents skeletons using 3D rotations, applies a warping layer to nominal curves, and uses a RollingMap layer on the Lie group. The actions are then transformed into feature vectors and classified using a linear Support Vector Machine (SVM). Investigations have explored Grassmann manifolds, such as the novel Grassmann network architecture introduced in [13]. This architecture includes three key components: the Projection block, which transforms orthonormal input matrices; the Pooling block, which maps these matrices and applies mean pooling; and the Output block, where the processed data is vectorized and classified. Rui Wang et al. [7] propose DreamNet, which uses SPDNet as the backbone and builds a stacked Riemannian autoencoder (SRAE) on the tail. The associated reconstruction error term can make the embedding functions of both SRAE and each RAE approximate identity mappings, which helps prevent the degradation of statistical information.

B. Hyperbolic neural network

Recent research has highlighted spiral approaches, notably in the prominent work by Ganea et al. [10]. The authors combine Möbius gyrovector spaces with the Riemannian geometry of the Poincaré model, resulting in hyperbolic adaptations of essential deep learning tools. These adaptations include multinomial logistic regression, feed-forward networks, and recurrent neural networks such as gated recurrent units. This framework supports the embedding of sequential data and enhances classification within hyperbolic space. Shimizu et al. [14] introduced a novel methodology that integrates multinomial logistic regression, fully-connected layers, convolutional layers, and attention mechanisms into a unified mathematical framework without increasing the number of parameters. This approach demonstrates superior parameter efficiency, stability, and performance compared to traditional hyperbolic and Euclidean components. Dai et al. [15] introduced a hyperbolic-to-hyperbolic graph convolutional network (H2H-GCN) that operates directly on hyperbolic manifolds. They developed a manifold-preserving graph convolution featuring a hyperbolic feature transformation and hyperbolic neighborhood aggregation. The feature transformation performs a linear transformation on hyperbolic manifolds, maintaining node representations on the manifold by enforcing orthogonal constraints. The neighborhood aggregation updates node representations using the Einstein midpoint. The HMANet proposed by Meng et al. [16] introduces a unified mathematical interpretation for fundamental components of neural networks in a single hyperbolic geometry model, the Poincaré ball model. This includes a multinomial logistic regression, fully-connected layers, convolutional layers, and attention mechanisms. Chen et al. [17] proposed a fully hyperbolic framework for constructing hyperbolic networks using the Lorentz model. They adapt Lorentz transformations, including boosts and rotations, to formalize key neural network operations. They also demonstrate that the linear transformations in tangent spaces used by existing hyperbolic networks are a simplified version of Lorentz rotations and do not account for boosts, thereby limiting the capabilities of these networks.

III. Preliminaries

Hyperbolic space $\mathbb{H}^n$ is a Riemannian manifold of constant negative curvature [18]. Distances grow exponentially with separation, and parallel lines diverge, unlike in Euclidean geometry.

Unit Poincaré Ball Model $\mathbb{B}^n \subset \mathbb{R}^n$ is the most widely used representation of hyperbolic space, favored for its intuitive geometric interpretation. It maps hyperbolic space onto the interior of the unit ball.

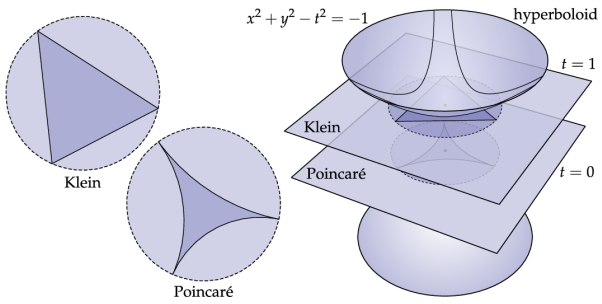

Fig. 1. Various types of hyperbolic spaces and their structural characteristics

In Riemannian geometry, Möbius addition and scalar multiplication are key operations commonly applied in the Poincaré disk model, defined by:

$$x \oplus y = \frac{(1+2\langle x,y\rangle + \|y\|^2)x + (1-\|x\|^2)y}{1+2\langle x,y\rangle + \|x\|^2\|y\|^2} \tag{1}$$

$$r \otimes x = tanh(r \times tanh^{-1}(\|x\|))\frac{x}{\|x\|} \tag{2}$$

where x and $y \in \mathbb{B}^n$, $r \in \mathbb{R}$ and $\langle .,. \rangle$ and $\|.\|$ denote respectively the Euclidean inner product and the Euclidean norm.

Additionally, the exponential map ($\exp_x(v)$) and logarithmic map ($\log_x(y)$) are fundamental tools for navigating curved spaces. They are computed as:

$$\mathrm{Exp}_x(v) = x \oplus (tanh(\frac{\lambda_x \|v\|}{2}) \frac{v}{\|v\|} \tag{3}$$

$$\mathrm{Log}_x(y) = \frac{2}{\lambda_x} tanh^{-1}(\| -x \oplus y\|) \frac{-x \oplus y}{\| -x \oplus y\|} \tag{4}$$

where $\lambda_x = \frac{2}{1-\|x\|^2}$ represnts the conformal factor that adjusts for the curvature of hyperbolic space.

IV. Spatio-Temporal Hyperbolic Aggregation Neural Network

In this section, we introduce our network model, called Spatio-Temporal Hyperbolic Aggregation Neural Network (ST-HAgg-Net). Initially, we provide an outline of our methodology. Then, we elucidate our proposed network designed to build an hyperbolic representation of the skeletal sequence. Lastly, we detail the classification procedure.

A. Model overview

The model depicted in Figure 2 is designed to enhance the efficacy of our hyperbolic network in recognizing bodily movements using skeletal data. Prior to the commencement of model execution, a crucial preprocessing step is undertaken to standardize the number of frames across all actions to a uniform value, represented as N frames, which is accomplished through interpolation. Subsequently, the resulting arrays are normalized to streamline computational operations, thereby priming the data for execution.

Our proposed network consists of three main modules. The first, Hyp-Map, performs hyperbolic mapping by segmenting skeletal sequences in time and space, embedding them into the Poincaré ball. The second, Spatial-Temporal Hyperbolic (ST-H), conducts both local and global segmentation, aggregating segments via the Möbius gyromidpoint to produce a compact sequence representation. Finally, the Classification module projects this representation into Euclidean space and applies fully connected layers for action classification.

B. Hyperbolic mapping

From 3D joints, captured either by a skeletal tracker or extracted from RGB images using real-time Pose estimation algorithms like OpenPose [19], we employ manifold-based approaches in data preprocessing to reduce dimensionality. However, these data may suffer from remarkable capturing errors due to factors such as camera precision, tracking algorithm performance, and camera position (e.g., angle deviation, weather conditions). To mitigate these issues, we propose applying a convolutional module to compensate for errors and enhance correlations between body joints.

Initially, we partition the body skeleton (or hand skeleton) into P parts. A common partitioning method divides the skeleton into segments like two arms, two legs, and a torso. The input data is denoted as $X \in \mathbf{R}^{K1}, K1 = N \times P \times J_P \times coor$, where N is the number of sequences, J_P is the number of joints in each body part, and $coor$ represents

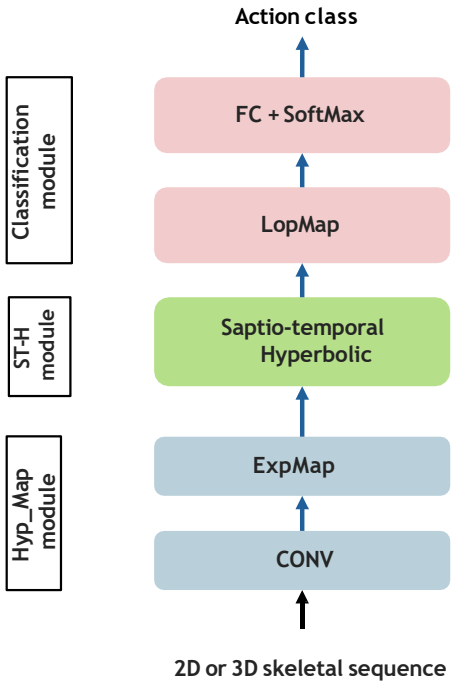

Fig. 2. Spatio-Temporal Hyperbolic Aggregation Neural Network (ST-HAgg-Net).

the 2D or 3D joint coordinates depending on the tracking algorithm output. Next, we apply a (1×1) convolutional layer without bias to mitigate errors. Subsequently, we map these features into hyperbolic space using the exponential map at $x = 0$. Substituting x with 0 in Equation 3, we obtain:

$$\mathrm{Exp}_{\mathbf{0}}(\mathbf{X}) = \tanh(\|\mathbf{X}\|) \frac{\mathbf{X}}{\|\mathbf{X}\|} \tag{5}$$

C. Spatial-Temporal hyperbolic Module

The primary goal is to accurately capture the spatial and temporal positions of body joints within each segment of a sequence, taking into consideration how joints are partitioned. In Euclidean space, the mean and covariance are typically utilized to derive significant insights along specified axes. In hyperbolic space, however, the concept of mean is replaced by methods like the midpoint due to the curvature properties of the space. Common approaches to generalize the mean in hyperbolic settings include tangent space aggregation, the Einstein midpoint method applied in the Beltrami-Klein model [20], and the Fréchet mean method. Our chosen method is the Möbius gyroaverage midpoint introduced by [14]. For a set of points $\{b_i\}_{1 \le i \le N} \in \mathbb{B}^n$, where $\mathbb{B}^n$ denotes the hyperbolic ball, the gyroaverage midpoint $\bar{b}$ is defined as:

$$\bar{b} = f_{mid}(\{b_i\}_{1 \le i \le N}) = \frac{1}{2} \otimes \left(\frac{\sum_{i=1}^{N} v_i \lambda_{b_i} b_i}{\sum_{i=1}^{N} |v_i| (\lambda_{b_i} - 1)} \right) \tag{6}$$

where $\{v_i\}_{1 \le i \le N}$ are real scalars. This approach enables the calculation of a representative midpoint in hyperbolic space, crucial for effectively capturing geometric relationships and

intrinsic structures among body joints across sequences. It leverages the unique properties of hyperbolic geometry to enhance the interpretation and analysis of spatial and temporal data in various applications.

In our context, we adopt the segmentation method proposed by [21]. We partition the sequence X, derived from the Convolution component, into six distinct sub-sequences $(X_s)_{s=1..6}$. The sub-sequence X_0 encompasses all N frames of Y. Sub-sequences X_1 and X_2 each correspond to one half of all sequences in Y, while X_3, X_4, and X_5 each account for one-third of all sequences in Y. Subsequently, each sub-sequence is further divided into NS segments, where each segment within the same sub-sequence contains an equal number of frames $nb[s]$.

Subsequently, we compute the midpoint μ and covariance Σ along the local spatial axis, corresponding to joints located in each part P:

$$\bar{X}_{s,t,P,n} = \sum_{i=1}^{J[P]} f_{mid}(\{X_{s,t,P,nb[s]}\}_{j\in P}) \tag{7}$$

where $J[P]$ denotes the joints in part P.

Given the risk of estimation errors in high-dimensional data—which can destabilize the logarithm operator and hinder classification—we apply orthogonal transformations after each midpoint aggregation to maintain stability. In the Poincaré ball model, points lie within the unit disk or ball, and geodesics are either circular arcs orthogonal to the boundary or diameters. Preserving orthogonality through such transformations is essential for maintaining geometric structure. Additionally, orthogonal transformations preserve angles between lines, a key property in hyperbolic geometry, where triangle angle sums are always less than 180 degrees. This preservation is critical to maintaining consistency in geometric relations. In summary, the use of orthogonal matrices ensures geometric integrity, simplifies transformations, and improves both computational efficiency and accuracy in modeling and calculations. To this end, we apply the following transformation:

$$Y = M \otimes \bar{X}_{s,t,P,nb[s]} = M \times \bar{X}_{s,t,p,nb[s]} \tag{8}$$

Where $M \in \mathbb{O}_n$ and $Y \in \mathbb{R}^{\mathrm{K}}, \mathrm{K} = 6 \times \mathrm{P} \times \mathrm{NS} \times \mathrm{nb[s]} \times 3$

We conduct a local temporal analysis, focusing on each subsequence divided into NS segments of elementary sequences. Within each segment, we apply operations involving f_{mid} followed by orthogonal transformations along this axis. Next, we transition to the global spatial axis, considering the entire body part. We then extend our analysis to the global temporal axis. Finally, we aggregate the six subsequences to derive the final hyperbolic vector representation. This sequential approach ensures a comprehensive analysis of both local temporal dynamics and global spatial characteristics, culminating in a unified representation suitable for further processing and analysis. (see Fig 3).

D. Classification module

After completing the necessary spatial and temporal analyses, we transform our final output into Euclidean space

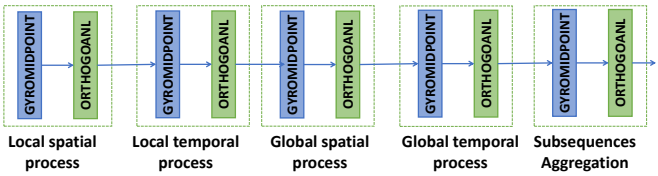

Fig. 3. ST-H module architecture. We denote by MID the midpoint operation, by ORTH the orthogonal transformation and by Trans Transformation.

using the logarithmic mapping equation 4 at $x = 0$. This transformation is defined as:

$$\mathrm{Log}_0(y) = \tanh^{-1}(\|y\|)\frac{y}{\|y\|} \tag{9}$$

Subsequently, we employ the CrossEntropyLoss function and the SGD (Stochastic Gradient Descent) algorithm for optimization and classification tasks. These steps ensure that our hyperbolic vector representations are appropriately mapped to Euclidean space for final processing and model training.

V. Experiments and Results

In this section, we evaluate our model's performance using three datasets: UAV-Human, UAV-Gesture, and DHG 14/28. The UAV-Human and DHG 14/28 datasets provide skeletal data, while for UAV-Gesture, we used OpenPose [19] for joint tracking. We detail our experimental setup and results for each dataset and compare our approach with state-of-the-art methods. The model is implemented in Python 3.9.7, running on an octa-core CPU at 3.2 GHz with 32GB of RAM. Computation time is about 140 ms per sequence. We performed numerous experiments to determine the optimal settings for our study.

A. Datasets

UAV-Human Dataset [22] is a large-scale dataset for UAV-based human behavior analysis, containing 67,428 RGB and skeleton video sequences from 119 subjects. For our study, we use 23,031 skeletal sequences (72% for training) across 155 action classes—122 solo and 33 involving objects.

UAV-Gesture Dataset [23] is designed for UAV control via gesture recognition. It comprises 119 outdoor HD video clips (37,151 frames) featuring 13 gestures derived from standard aircraft and helicopter command signals.

DHG Dataset [24] contains depth images and hand skeletons recorded at 30 FPS using an Intel RealSense camera. It features 14 hand gestures, each performed with one finger and the whole hand, resulting in either 14 or 28 gesture classes. The dataset includes 1,960 training sequences (70%) and 840 test sequences (30%).

B. Ablation study

To evaluate the impact of hyperbolic geometry on model performance, we developed a Euclidean variant of our ST-HAgg-Net model, referred to as ST-Euc-Net. In this version, we replaced the expmap and logmap layers with Euclidean

counterparts and substituted the hyperbolic midpoint layer with a Euclidean mean calculation, while keeping all other layers unchanged. We then assessed the performance of both networks—ST-HAgg-Net in hyperbolic space and ST-Euc-Net in Euclidean space on the three datasets. The results of these experiments are presented in TableI.

Method	UAV-Human	DHG-14	UAV-Gesture
ST-Euc-Net	45.31%	90.12%	86.15%
ST-HAgg-Net	51.73%	95.71%	92.74%

TABLE I

EVALUATION OF ST-HAGG-NET (HYPERBOLIC) VS. ST-EUC-NET (EUCLIDEAN).

Table I illustrates the superior performance achieved with hyperbolic geometry. In the UAV-Human dataset, our hyperbolic model outperformed the Euclidean baseline by 6.42%. Similarly, in the DHG-14 dataset, across 14 gesture categories, the hyperbolic model exceeded the Euclidean version by 5.6%. These results highlight the substantial performance gains that can be achieved by adopting a hyperbolic manifold, reinforcing the benefits of leveraging non-Euclidean geometry for complex human action and gesture recognition tasks. The superiority of hyperbolic spaces stems from their ability to naturally model complex hierarchical relationships, which are often present in human action and gesture data. This underscores the effectiveness of defining the model within hyperbolic space, resulting in notable improvements in both accuracy and performance.

C. Comparison with the state-of-the-art methods

1) UAV-Human dataset: The effectiveness of our model on the UAV-Human skeleton database, along with its comparative performance against other models, is presented in Table II. Our results show that ST-HAgg-Net achieves higher accuracy compared to other networks due to several key factors.

Most of these approaches in HAR focus on improving Graph Convolutional Networks (GCNs) due to their ability to effectively capture local structures, which helps in extracting rich geometric information from graphs. However, GCNs typically have shallow architectures, often limited to three layers. Furthermore, as the complexity of the graph grows, scaling GCNs becomes increasingly difficult, which restricts their performance on large and intricate datasets, such as UAV-Human. Our model is specifically optimized to handle complex sequential data. By leveraging a deep embedding learning mechanism based on the Poincaré ball model, it generates rich geometric features. This approach emphasizes correlations between skeletal joints and introducing rotational transformations within the ball, leading to more accurate and nuanced interpretations of human actions.

2) UAV-Gesture dataset: The performance of our model on the UAV-Gesture skeleton database, compared to alternative models, is detailed in Table III.

Method	Accuracy(%)
2S-AGCN [25]	34.84
HARD-Net [26]	36.97
Shift-GCN [27]	37.98
TC-GCNs (joint+Bone) [28]	45.14
Dream-Net [7]	46.28
MITFAS [29]	50.8
ST-HAgg-Net(ours)	**51.73**

TABLE II

PERFORMANCE COMPARISON WITH THE STATE-OF-THE-ART METHODS ON UAV-HUMAN DATASET.

Method	Accuracy(%)
P-GNN [23]	91.9
PSD (Fitting, alignment) [30]	92.44
ST-HAgg-Net(ours)	**92.74**

TABLE III

PERFORMANCE COMPARISON WITH THE STATE-OF-THE-ART METHODS ON UAV-GESTURE DATASET.

Our ST-HAgg-Net neural network surpasses existing state-of-the-art methods, achieving high accuracy by capturing information through the combination of mean and covariance, followed by the application of a series of midpoint layers along different axes. This approach enhances the model's ability to generate superior vector representations for actions. As a result, our model demonstrates strong performance in both hand gesture recognition and body action recognition tasks.

D. DHG 14/28 dataset

Method	DHG-14(%)	DHG-28(%)
STA-Res-TCN [31]	93.57	90.7
TCN-Summ [32]	93.57	91.43
ST-TS-HGR-NET [21]	94.29	89.4
DD-Net [33]	94.6	91.9
HMANET [16]	95.12	92.62
ST-HAgg-Net(ours)	**95.71**	**94.29**

TABLE IV

STATE-OF-THE-ART METHODS ON DHG DATASET.

Table IV summarizes various approaches using 3D skeletal hand sequences from the DHG-14/28 dataset and the performance achieved by each. The results confirm the effectiveness of our proposed network architecture for hand gesture recognition, as it achieves higher accuracy compared to other methods. Based on these findings, we conclude that our approach, ST-HAgg-Net, offers a pivotable advantage over previous works. Compared to HMANet [16], which also utilizes a hyperbolic-based approach, ST-HAgg-Net achieves superior results. This improvement can be attributed to our network's strategy of maintaining all transformations within the Poincaré unit ball, rather than hybridizing between Euclidean space and hyperbolic space.

VI. CONCLUSION

In conclusion, our research on integrating hyperbolic geometry into deep learning has led to notable advancements in human action recognition. Although hyperbolic neural networks have been underutilized in this domain, we have demonstrated their effectiveness in recognizing hand gestures and body movements using both UAV-based and ground-based cameras. The ST-HAgg-Net model showcases strong potential in addressing the complexities of understanding human behavior through skeletal-based features. By embedding skeletal joints into the Poincaré ball model and employing spatio-temporal data processing, we have achieved precise and efficient action classification, as demonstrated across various context of HAR-related datasets.

As we continue to push the boundaries of UAV-based human behavior analysis, we plan to expand our research into other hyperbolic architectures, such as the hyperboloid space and the Beltrami-Klein model.

REFERENCES

[1] V. A. Adewopo, N. Elsayed, Z. ElSayed, M. Ozer, A. Abdelgawad, and M. Bayoumi, "A review on action recognition for accident detection in smart city transportation systems," *Journal of Electrical Systems and Information Technology*, vol. 10, no. 1, p. 57, 2023.

[2] J. Llaurado-Fons, A. Martinez, F. A. Pujol-López, and H. Mora, "An architecture for human action recognition in smart cities video surveillance systems," in *Research and Innovation Forum 2020: Disruptive Technologies in Times of Change*, pp. 51–56, Springer, 2021.

[3] M. S. Akremi, R. Slama, and H. Tabia, "Spd siamese neural network for skeleton-based hand gesture recognition.," in *VISIGRAPP (4: VISAPP)*, pp. 394–402, 2022.

[4] C.-F. R. Chen, R. Panda, K. Ramakrishnan, R. Feris, J. Cohn, A. Oliva, and Q. Fan, "Deep analysis of cnn-based spatio-temporal representations for action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6165–6175, 2021.

[5] D. Zhuang, M. Jiang, H. Arioui, and H. Tabia, "Action text diffusion prior network for action segmentation," in *Proceedings of the 20th International Conference on Content-based Multimedia Indexing*, pp. 79–85, 2023.

[6] M. B. Shaikh, D. Chai, S. M. S. Islam, and N. Akhtar, "From cnns to transformers in multimodal human action recognition: A survey," *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024.

[7] R. Wang, X.-J. Wu, Z. Chen, T. Xu, and J. Kittler, "Dreamnet: A deep riemannian manifold network for spd matrix learning," in *Proceedings of the Asian Conference on Computer Vision*, pp. 3241–3257, 2022.

[8] Z. Huang, C. Wan, T. Probst, and L. Van Gool, "Deep learning on lie groups for skeleton-based action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6099–6108, 2017.

[9] R. Slama, H. Wannous, M. Daoudi, and A. Srivastava, "Accurate 3d action recognition using learning on the grassmann manifold," *Pattern Recognition*, vol. 48, no. 2, pp. 556–567, 2015.

[10] O. Ganea, G. Bécigneul, and T. Hofmann, "Hyperbolic neural networks," *Advances in neural information processing systems*, vol. 31, 2018.

[11] R. Anirudh, P. Turaga, J. Su, and A. Srivastava, "Elastic functional coding of riemannian trajectories," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 5, pp. 922–936, 2016.

[12] R. Vemulapalli and R. Chellapa, "Rolling rotations for recognizing human actions from 3d skeletal data," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4471–4479, 2016.

[13] Z. Huang, J. Wu, and L. Van Gool, "Building deep networks on grassmann manifolds," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.

[14] R. Shimizu, Y. Mukuta, and T. Harada, "Hyperbolic neural networks++," *arXiv preprint arXiv:2006.08210*, 2020.

[15] J. Dai, Y. Wu, Z. Gao, and Y. Jia, "A hyperbolic-to-hyperbolic graph convolutional network," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 154–163, 2021.

[16] J. Chen, C. Zhao, Q. Wang, and H. Meng, "Hmanet: Hyperbolic manifold aware network for skeleton-based action recognition," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 15, no. 2, pp. 602–614, 2022.

[17] W. Chen, X. Han, Y. Lin, H. Zhao, Z. Liu, P. Li, M. Sun, and J. Zhou, "Fully hyperbolic neural networks," *arXiv preprint arXiv:2105.14686*, 2021.

[18] T. Lin and H. Zha, *Riemannian Manifold*, pp. 1081–1086. Cham: Springer International Publishing, 2021.

[19] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7291–7299, 2017.

[20] I. Chami, Z. Ying, C. Ré, and J. Leskovec, "Hyperbolic graph convolutional neural networks," *Advances in neural information processing systems*, vol. 32, 2019.

[21] X. S. Nguyen, L. Brun, O. Lézoray, and S. Bougleux, "A neural network based on spd manifold learning for skeleton-based hand gesture recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12036–12045, 2019.

[22] T. Li, J. Liu, W. Zhang, Y. Ni, W. Wang, and Z. Li, "Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16266–16275, 2021.

[23] A. G. Perera, Y. Wei Law, and J. Chahl, "Uav-gesture: A dataset for uav control and gesture recognition," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 117–128, 2018.

[24] Q. De Smedt, H. Wannous, J.-P. Vandeborre, J. Guerry, B. Le Saux, and D. Filliat, "Shrec'17 track: 3d hand gesture recognition using a depth and skeletal dataset," in *3DOR-10th Eurographics Workshop on 3D Object Retrieval*, pp. 1–6, 2017.

[25] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12026–12035, 2019.

[26] T. Li, J. Liu, W. Zhang, and L. Duan, "Hard-net: Hardness-aware discrimination network for 3d early activity prediction," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pp. 420–436, Springer, 2020.

[27] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, "Skeleton-based action recognition with shift graph convolutional network," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 183–192, 2020.

[28] C. Li, S. Li, Y. Gao, L. Zhou, and W. Li, "Static graph convolution with learned temporal and channel-wise graph topology generation for skeleton-based action recognition," *Computer Vision and Image Understanding*, vol. 244, p. 104012, 2024.

[29] R. Xian, X. Wang, and D. Manocha, "Mitfas: Mutual information based temporal feature alignment and sampling for aerial video action recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6625–6634, 2024.

[30] B. Szczapa, M. Daoudi, S. Berretti, A. Del Bimbo, P. Pala, and E. Massart, "Fitting, comparison, and alignment of trajectories on positive semi-definite matrices with application to action recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 0–0, 2019.

[31] J. Hou, G. Wang, X. Chen, J.-H. Xue, R. Zhu, and H. Yang, "Spatial-temporal attention res-tcn for skeleton-based dynamic hand gesture recognition," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 0–0, 2018.

[32] A. Sabater, I. Alonso, L. Montesano, and A. C. Murillo, "Domain and view-point agnostic hand action recognition," *arXiv preprint arXiv:2103.02303*, 2021.

[33] F. Yang, Y. Wu, S. Sakti, and S. Nakamura, "Make skeleton-based action recognition model smaller, faster and better," in *Proceedings of the ACM multimedia asia*, pp. 1–6, ., 2019.