2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

SPDAGG-TransNet: Integrating Symmetric Positive Definite Networks with Transformers for UAV-Human Action Recognition*

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

***Abstract*— The advent of unmanned aerial vehicles (UAVs) has initiated a revolutionary era in human action recognition, profoundly influencing various domains. This transition underscores the critical necessity for comprehensive benchmarks crucial for formulating and evaluating UAV-centric models tailored to human behavior analysis.**

This paper presents an novel approach called SPDAGG-TransNet network for UAV-human action recognition, leveraging the resilience of skeletal-based features amidst these obstacles. Our approach revolves around a deep neural network adept at capturing the intricate spatial and temporal dimensions of human actions, leading to the development of Semi-Positive Definite (SPD) matrix representations. These representations are then transformed using a transformer encoder before being classified using a Multilayer Perceptron (MLP). To assess the effectiveness of our approach, we conduct thorough evaluations using publicly available datasets such as the UAV-Human Action Recognition and UAV-Gesture datasets. Our findings underscore the state-of-the-art performance achieved by our method, highlighting its potential to significantly advance UAV-based human action recognition.

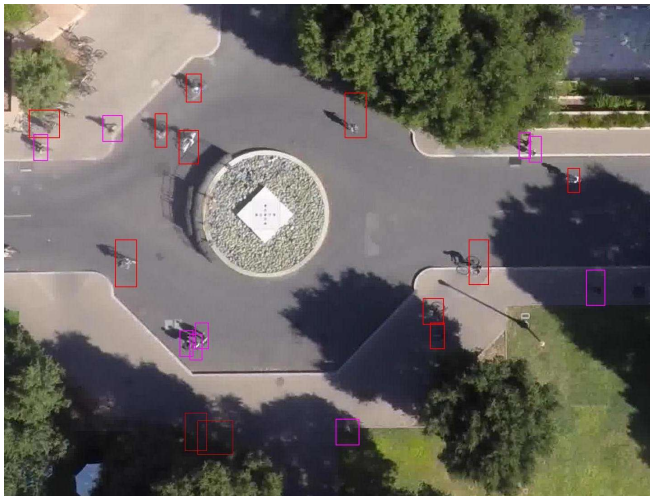

Fig. 1. Example of UAV-Human activities detection.

I. INTRODUCTION

Human action recognition from the viewpoint of unmanned aerial vehicles (UAVs) poses significant challenges, largely due to the dynamic nature of UAV flight, encompassing factors such as motion, attitude adjustments, and varying viewpoints. Moreover, complexity is heightened by challenges including occlusions, pose variations, and background clutter prevalent in UAV-captured images.

UAV-human action recognition holds promising applications across diverse domains. Primarily, it plays a pivotal role in surveillance, enabling the detection of suspicious behaviors to enhance security measures in various environments [2]. Additionally, in healthcare, it proves invaluable for remotely monitoring human vital signs, especially in scenarios requiring continuous health tracking such as remote or disaster-stricken areas [8]. Furthermore, it aids in disaster response and assessment by providing real-time insights into individual activities in affected regions, thereby facilitating damage assessment, survivor location, and optimizing rescue operations [12].

Beyond these critical domains, UAVs offer a unique advantage in monitoring human activities across expansive areas like stadiums, parks, and sprawling urban environments, ensuring comprehensive surveillance from diverse perspectives. Moreover, UAVs seamlessly engage in natural human interaction, participating in recreational activities, executing tasks, or providing assistance as needed. This multifaceted utility underscores the significant potential of UAV-human action recognition to enhance and revolutionize a wide array of applications, impacting domains essential to safety, well-being, and efficiency.

However, UAV-human action recognition presents new challenges not encountered in conventional action recognition methods. Notably, images captured by UAVs often exhibit attributes such as low resolution, noise, and distortion, primarily due to UAV motion and orientation. Additionally, these images frequently display significant variations in perspective and occlusions, stemming from the UAV's flight trajectory and altitude. Furthermore, backgrounds within UAV-captured imagery often feature intricate and ever-changing elements, potentially hindering the recognition of human actions. Lastly, human actions themselves encompass a wide spectrum of diversity and subtlety, necessitating precision and detail in the recognition process. (see Figure 1)

To address these challenges, various methods have been proposed for UAV-human action recognition, broadly categorized into appearance-based and skeletal-based approaches. We opt for skeletal-based methodologies due to their resilience to complex backgrounds and the flexibility to apply normalization techniques, enhancing robustness [18], [14].

In this paper, we propose a novel approach for UAV-human

[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 - 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

action recognition, leveraging skeletal data and SPD matrices properties. By utilizing the 3D coordinates of skeletal joints, our approach employs an SPD deep neural network to extract Temporal-spatial features that encapsulate both the spatial and temporal intricacies of human actions, transforming them into a sequence of SPD matrices. These SPD matrices are then mapped into Euclidean space and vectorized. Subsequently, a transformer encoder is employed to process the sequential vectors. It uses self-attention mechanisms to capture relationships between elements in the sequence efficiently.

We evaluate our approach rigorously using UAV-human dataset, and UAV-Gesture dataset. Remarkably, our methodology achieves state-of-the-art results on these benchmark datasets, demonstrating its prowess in action recognition within UAV-captured imagery.

This paper makes the following contributions:

- Introducing an innovative model that integrates SPD net and transformer architectures for skeletal-based UAV-human action recognition, filling a gap in an emerging field with limited prior research.
- Providing a comprehensive approach for addressing both temporal and spatial aspects in local segments and global sequences, enhancing the model's ability to capture nuanced human actions.
- Validating the efficiency of the proposed algorithm through rigorous testing using skeletal registration datasets, affirming its effectiveness in real-world applications.

The remainder of this paper is organized as follows: Section 2 reviews related work on UAV-human action recognition and deep learning manifold-based approaches. In Section 3, we describe our proposed approach, followed by experimental evaluations in Section 4. Finally, Section 5 concludes the paper.

II. Related works

A. Manifold-based approaches for Human action recognition

To explore the realm of action recognition, numerous research endeavors focus on methodologies that leverage skeletal data acquired through depth-sensing cameras. These studies often delve into the intricacies of non-Euclidean spaces, such as the utilization of elastic functional coding for the analysis of Riemannian trajectories [1]. Certain methodologies center their attention on Lie group techniques, exemplified by the deep learning neural network developed by Huang et al. in their work on skeleton-based action recognition [5]. This innovative approach hinges on the succession of a novel RotMap layer and RotPooling block. The RotMap layer adeptly transforms the initial rotation matrices, while the ensuing RotPooling layer meticulously temporally and spatially aggregates the resultant rotation matrices. The entire network culminates in a LogMap Layer, thereby completing this comprehensive framework. In the same field, Vemulapalli et al. proposed in [21] a neural network. It starts with a Skeletal representation using 3D rotations between the body joints. Then comes the warping layer to compute the nominal curves and warp all the curves to it. Along these nominal curves, a RollingMap layer are applied on the Lie group over its lie algebra and the actions are unwrapped onto this lie algebra. These unwrapped actions are finally transformed into feature vectors and are classified using the linear Support Vector Machine (SVM) classifier.

Different investigations have directed their attention toward Grassmann manifolds, exemplified by the introduction of a novel Grassmann network architecture as proposed in [6]. This innovative network structure comprises three significant components: the Projection block, the Pooling block, and the output block. The Projection block serves the purpose of transforming the orthonormal input matrices, while the Pooling block is meticulously designed to map these orthonormal matrices and subsequently apply mean pooling operations. The resultant outputs are then vectorized and subjected to classification within the output block.

In a related research domain, Vemulapalli et al. introduced a neural network framework in their study documented in [21]. Their methodology starts with the utilization of a skeletal representation that effectively captures 3D rotations occurring among the body joints. Following this initial step, they employ a warping layer to calculate nominal curves and align all these curves with a reference curve. Along the trajectory of these nominal curves, a specialized RollingMap layer operates on the Lie group, working within the confines of its Lie algebra to unwrap the actions onto this mathematical framework. Subsequently, these unwrapped actions undergo transformation into feature vectors, ultimately culminating in their classification using a linear SVM classifier. Nguyen et al. [10] proposes a network architecture to learn an SPD matrix representing a hand gesture. It consists of three main stages. The first stage is based on a convolutional layer. The second stage relies on different architectures for spatial and temporal Gaussian aggregation of joint features. The third stage learns a final SPD matrix from skeletal data Rui Wang et al.[22] 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 of each RAE an approximate identity mapping, which helps to prevent the degradation of statistical information.

B. Vision transformer for action recognition

Mazzia et al.[9] propose the Action Transformer (AcT), a novel architecture for human action recognition. AcT is a fully self-attentional model that consistently outperforms more complex networks combining convolutional, recurrent, and attention layers. The approach leverages 2D pose representations over small temporal windows, providing a low-latency solution for accurate real-time performance. The model was evaluated on the newly introduced MPOSE2021 dataset, demonstrating its effectiveness and laying the groundwork for future research in human action

recognition. Shi et al.[15] introduces a model for skeleton-based human action recognition, employing sparse attention along the spatial dimension and segmented linear attention across the temporal aspect of data. Additionally, our model is adept at handling video clips of varying lengths aggregated into a single batch. Qiu et al.[13] proposes a novel method called STTFormer (Spatio-Temporal Tuples Transformer) to enhance the modeling of joint correlations. The problem lies in existing Transformer-based methods, which struggle to effectively capture the correlation of different joints across frames. To address this, the STTFormer divides the skeleton sequence into parts, each containing consecutive frames. It introduces a spatio-temporal tuples self-attention module to capture relationships between joints in consecutive frames. Additionally, a feature aggregation component improves the ability to distinguish similar actions by considering non-adjacent frames. Yang et al.[24] propose a method called RViT. It leverages spatial-temporal representation learning and incorporates an attention gate to connect the current frame input with the previous hidden state, allowing it to aggregate global inter-frame features over time. RViT processes videos recurrently, considering both spatial and temporal features due to its attention gate and recurrent execution.

III. The Proposed Approach

In this section, we introduce our network model, known as Temporal-spatial SPDAGG-TransNet. Initially, we provide an outline of our methodology. Then, we elucidate our proposed network designed to build an SPD representation of the skeletal sequence. Lastly, we detail the classification procedure.

A. Overview

Our model, depicted in Figure 2, aims to improve the performance of an SPD network designed for body action recognition using skeletal data through integration with a transformer. An essential preprocessing step involves standardizing the number of frames for all actions to a uniform value, denoted as N frames, achieved via interpolation. Subsequently, the resulting arrays undergo normalization to streamline computational processes, priming the data for processing.

The proposed architecture comprises two primary modules: a Temporal-spatial SPDAGG network, responsible for representing segments extracted from the entire sequence, each encapsulating an executed action, as a sequence of SPD matrices. After transitioning these matrices into Euclidean space and vectorizing them, we employ a transformer to enable the model to discern varying levels of significance among different segments, leveraging the self-attention mechanism.

B. Temporal-spatial SPDAGG network

Our proposed network is structured around three core elements. The initial component, dubbed the CONV component, integrates convolution layers aimed at rectifying capturing

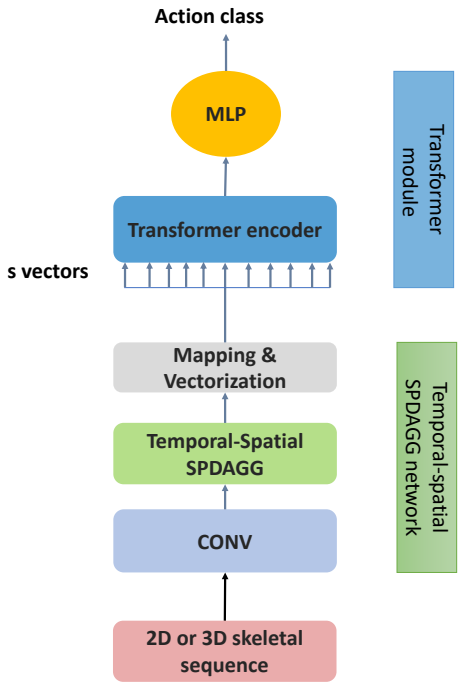

Fig. 2. Overview of the proposed SPDAGG-TransNet network. Given the joints coordinates, the data is processing by two consecutive subnetworks: a temporal-spatial SPDAGG network and a transformer model.

inaccuracies and strengthening correlations among various body parts. Following this, the Temporal-spatial SPDAGG Aggregation component partitions the sequence into a series of SPD matrices and consolidates them based on specified criteria, ultimately synthesizing the entire sequence into a unified SPD matrix representation. Lastly, the Projection component transforms this SPD matrix into Euclidean space and employs fully-connected layers to learn the best vectors representation.

1) Convolution component: Utilizing 3D joints extracted from RGB images captured by UAVs and identified via real-time Pose estimation algorithms such as OpenPose [3], we employ manifold techniques in data engineering to reduce dimensionality. However, these data may be prone to significant capturing inaccuracies due to factors like camera precision, tracking algorithm performance, and UAV navigation conditions (e.g., angle deviation, weather). To tackle this challenge, we propose the integration of a convolution component to compensate for these errors and amplify correlations among body joints.

The input data, represented as $X \in \mathbb{R}^{N\times P\times J_P\times coor}$, where N denotes the sequence count, P indicates the number of body parts, J_P signifies the joints per body part, and *coor* denotes the 2D or 3D joint coordinates (depending on the output of the tracking algorithm). We partition the body skeleton into P parts, as depicted, and apply a (1×1) convolution layer devoid of bias to rectify errors. Subsequently, we employ a 2D convolution layer with a (3×3) kernel size and f filters to amplify correlations among the body joints. (refer to the first block of the SPDAGG module in Fig 2)

The resulting output Y is computed as

$$Y = Conv_{2D,3\times3}(conv_{1D,1\times1}(X)); \quad (1)$$

Where $Y \in \mathbb{R}^{N\times P\times J_P\times f}$

2) Temporal-spatial SPDAGG Aggregation component: The primary aim of this component is to acquire spatial and temporal insights into the positions of body joints within each segment of the sequence, while considering joint partitioning. This process yields an SPD matrix that encapsulates both first-order and second-order information regarding the mean and covariance of spatial distribution across frames. To achieve this, we propose employing the segmentation method introduced by [10]. We segment the sequence X, derived from the Convolution component, into six distinct sub-sequences $(X_s)_{s=1..6}$. Specifically, X_0 encompasses all N frames of Y. Each of X_1 and X_2 corresponds to one of the halves of all sequences within Y. Similarly, X_3, X_4, and X_5 each account for one-third of all sequences present in Y. Subsequently, we further divide each subsequence into NS segments, where each segment within the same subsequence contains an equal number of frames denoted as $nb[s]$. We obtain: $X_i \in \mathbb{R}^{NS\times nb[s]\times P\times J_p\times f}$

As illustrated in Fig 3, our proposal incorporates a modified Gauss aggregation layer that integrates both first-order and second-order information across frames. This approach aims to locally extract temporal features and reduce the frame count, which is particularly beneficial considering the resource limitations of UAVs.

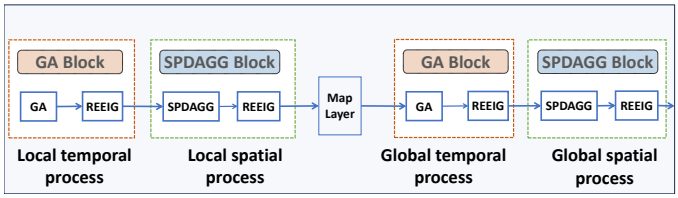

Fig. 3. Temporal-spatial SPDAGG Aggregation Module architecture. We build with Gauss aggregation block the first set of SPD matrices with respect to local segments. Then we apply a series of SPDAGG block alternating between spatial analysis and temporal analysis to end up with a final single SPD representation

First, we compute the mean μ and covariance Σ based on the number of frames along each subsequence axis, as follows:

$$\mu_{s,t,p,j} = \frac{1}{nb[s]} \sum_{i=1}^{nb[s]} X_{s,t,i,p,j} \quad (2)$$

$$\Sigma_{s,t,p,j} = \frac{1}{nb[s]} \sum_{i=1}^{nb[s]} (X_{s,t,i,p,j} - \mu_{s,t,p,j})(X_{s,t,i,p,j} - \mu_{s,t,p,j})^T \quad (3)$$

In situations where high-dimensional data is involved, inaccuracies in estimation can greatly diminish classification performance by making the logarithm operator unstable and ill-conditioned. As a result, distance calculations become exceptionally sensitive to noise. To tackle this challenge, shrinkage emerges as a commonly employed technique to alleviate bias in estimation. Covariance matrices are often substituted with a modified version, denoted as:

$$\hat{\Sigma}_{s,t,p,j} = (1-\lambda)\Sigma_{s,t,p,j} + \lambda \times \frac{trace(\Sigma_{s,t,p,j})}{nb[s]} I \quad (4)$$

Where λ represents the shrinkage parameter and I is the identity matrix.

To explore the balance between the effects of the mean and the covariance, we introduce a parameter $\beta \in [0,1]$. The modified Gauss Aggregation layer is formulated as follows:

$$Y(s,t,f) = \begin{bmatrix} \beta\hat{\Sigma} + (1-\beta)\mu\mu^T & \sqrt{1-\beta}\mu \\ \sqrt{1-\beta}\mu^T & 1 \end{bmatrix} \quad (5)$$

Noting that $Y \in \mathbb{R}^{6\times P\times NS\times J\times(f+1)\times(f+1)}$

Following this, we introduce the ReEig layer similar to the role of ReLU layers in traditional Convolutional Neural Networks. It rectifies the resulting SPD matrices with a non-linear function as below:

$$Y = U\ max(\varepsilon I, S)\ U^T \quad (6)$$

where $X = USV$ is the eigen-decomposition, I is the identity tensor (the same size as S) and ε is a rectification threshold.

We proceed with a localized spatial analysis at the joint level within each body part employing the SPD aggregation layer. For this purpose, we establish a set of weights $W \in \mathbb{R}^{6\times P\times NS\times J\times d_{out1}\times(f+1)}$, ensuring their residence on a compact Stiefel manifold. The aggregation process unfolds as follows:

$$Y_{s,t,p} = \sum_{j\in[1..J]} W_{s,t,p,j} X_{s,t,p,j} W_{s,t,p,j}^T \quad (7)$$

After the SPD aggregation layer, we integrate a ReEig layer. Then, we apply the block SPD Aggregation layer ($d_{out2} \times d_{out1}$) followed by another ReEig layer on the NS subsequences to perform a comprehensive global temporal analysis of the sequences. Subsequently, we proceed to utilize the block SPD Aggregation layer ($d_{out3} \times d_{out2}$) followed by a ReEig layer on the P body parts to execute a holistic global spatial analysis of the sequences.

3) Projection component: Upon the conclusion of the necessary spatial and temporal analyses, we utilize the resulting s SPD matrices to investigate their vector representation. To achieve this, we map them into Euclidean space by using the LogEig layer.

$$Y = U\ log(S)\ U^T \quad (8)$$

where $X = USV$ is the eigen-decomposition of X. Then we vectorize it and apply series of Fully-Connected (FC) layers and a SoftMax layer. Let $X \in \mathbb{R}^{s\times d_{model}}$ be the set of the final vectors.

C. Transformer module

The Transformer model [20] is a pivotal architecture in natural language processing and sequence modeling. It is is made of N_t layers with alternating H multi-head self-attention and feed-forward layers. each layer is accompanied by a normalization layer and residual connections to facilitate stable training(see Fig. 4).

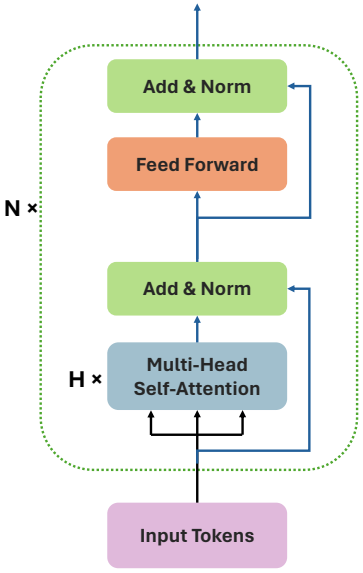

Fig. 4. Transformer encoder layer architecture. Input tokens go through N encoder layers and H self-attention heads.

In the realm of multi-head self-attention, the model begins by processing an input sequence of embeddings $X \in \mathbb{R}^{s \times d_{\text{model}}}$, outputs of the SPDAGG module. The attention scores between queries and keys are then determined using the scaled dot-product mechanism : Through learned projection matrices $W_Q, W_K, W_V \in \mathbb{R}^{d_{\text{model}} \times d_k}$, queries Q, keys K, and values V are computed as follows: $Q = XW_Q$, $K = XW_K$ and $V = XW_V$. The attention score (SA) between each query and key is computed as:

$$SA_X(Q,K,V) = \text{SoftMax}(\frac{QK^T}{\sqrt{d_k}})V \tag{9}$$

d_k is the dimension of the projected queries, keys, and values. We perform this SA operation for all the H heads of the i^{th} layer of the encoder. This process is executed across multiple attention heads, with the results concatenated and linearly transformed as follows:

$$Y = [SA_1(X), SA_2(X)..SA_H(X)]W_o \tag{10}$$

where $\text{SA}_i(X) = SA_X(Q_i, K_i, V_i)$ and $W_O \in \mathbb{R}^{hd_k \times d_{\text{model}}}$ is the output projection matrix.

Moving to the feed-forward neural network stage, the output from the multi-head self-attention undergoes further processing. This involves passing the activations through position-wise feed-forward layers, each comprising two linear transformations followed by a rectified linear unit (ReLU) activation function. This mechanism enables the model to capture intricate relationships and features within the input sequence.

Then, we apply a normalization layer and residual connections to ensure stable training and efficient information flow. These components are integrated after each sub-layer, where layer normalization normalizes the output of each sub-layer and residual connections facilitate the smooth propagation of gradients during training.

Finally, we employ a Multilayer Perceptron (MLP) to perform the classification task, discerning and categorizing the actions executed within these sequences based on the learned representations.

IV. Experiments and Results

We intend to showcase the model's performance using the UAV-Human dataset and UAV-Gesture dataset. The skeletal data information is provided on UAV-Human dataset. We apply OpenPose [3] to track joints of UAV-Gesture dataset. We extensively provide details on the experimental settings and results obtained for each dataset. We also compare the state-of-the-art methods to our approach. Our model is implemented in python 3.9.7 environment.

A. Datasets

1) UAV-Human dataset:: constitutes an impressive compilation of 67,428 multi-modal video sequences involving the participation of 119 individuals. Its expansive nature renders it suitable for a diverse array of tasks, spanning action recognition, pose estimation, person re-identification, and attribute recognition. Meticulously gathered over the course of three months using UAV deployment, the dataset encompasses various urban and rural environments, captured both day and night. This comprehensive collection encapsulates a wide range of diversities, including subjects, backgrounds, lighting conditions, weather variations, occlusions, camera movements, and UAV flying orientations. Consequently, it serves as a robust resource catering to the burgeoning field of UAV-based human behavior understanding, offering invaluable insights for numerous applications. In our specific scenario, we focus on utilizing skeletal coordinates from the dataset, comprising a substantial collection of 23,031 skeletal sequences, with 72% allocated for training purposes. This dataset encompasses a total of 155 distinct action classes, with 122 performed independently by individuals and the remaining 33 involving actions executed with the assistance of auxiliary elements. Figure 5 provides a general overview of the various types of data available within this dataset.

2) UAV-Gesture dataset: [11]: The dataset is tailored for UAV control and gesture recognition purposes, serving as a valuable resource for analyzing and understanding commands directed towards unmanned aerial vehicles (UAVs). It comprises outdoor recorded video clips capturing UAV commanding signals, encompassing a set of 13 gestures conducive to basic UAV navigation and control. With a focus on general aircraft handling and helicopter handling signals, the dataset contains 119 high-definition video clips, spanning a total of 37,151 frames.

Fig. 5. Examples of action videos and different data representations in UAV-Human dataset.

These gestures, carefully selected from the realm of general aircraft handling and helicopter handling signals, are annotated with body joint data and corresponding gesture classes. This comprehensive annotation enhances the dataset's versatility, extending its utility across a broad spectrum of research areas, including gesture recognition, action recognition, human pose recognition, and situation awareness. (Refer to Fig. 6 for an illustration.)

Fig. 6. Instances of gestures captured in both RGB and skeletal representations within the UAV-Gesture dataset..

B. Experimental settings

Many experiments were carried out in order to find the most convenient settings for this experiment. Due to space limitations, we will content ourselves with transcribing the settings values in the TableI. N.B: we applied 3 FC layers with size $(M \times dc1)$, $(dc1 \times dc2)$ and $(dc2 \times dc3)$ where $M = 20100$.

For the transformer module, we set the number of attention head $H = 3$ and the number of transformer blocks $N_t = 6$.

Parameter	Description	Value
N	Interpolation frames	500
f	output channels of CONV layer	9
s	Number of sub-sequences	6
NS	Number of segments in each subsequence	15
ε	Threshold of the ReEig layer	10^{-4}
d_{out1}	Output size of the 1^{st} SPDAGG layer	50
d_{out2}	Output size of the 2^{nd} SPDAGG layer	100
d_{out3}	Output size of the 3^{rd} SPDAGG layer	150
d_{out4}	Output size of the 4^{th} SPDAGG layer	200
d_{c1}	Output size of the 1^{st} FC layer	10000
d_{c2}	Output size of the 2^{nd} FC layer	5000
d_{c3}	Output size of the 3^{rd} FC layer	500
H	Number of attention heads	3
N_t	Number of Transformer layers	6

TABLE I
MODEL DEFAULT SETTINGS

To validate the efficacy of combining the Temporal-spatial SPDAGG network and transformer blocks, we present the performance obtained individually by each network and then by their combined architecture. This approach allows us to assess the effectiveness of each component separately as well as their synergistic performance when integrated together.

C. Comparison of the state-of-the-art methods on:

1) UAV-Human dataset: We conducted experiments to evaluate the effectiveness of our SPDAGG-TransNet model on the UAV-Human skeleton database and compared its performance with other existing models, as outlined in Table II. The results indicate that our Temporal-spatial SPDAGG network demonstrates superior efficiency across several key metrics compared to alternative networks. Notably, this network showcases remarkable proficiency in feature extraction, effectively capturing both first-order and second-order information. Additionally, its Gauss aggregation layers exhibit efficiency in spatial and temporal data aggregation. When considering UAV-human action recognition, these combined strengths render our Temporal-spatial SPDAGG neural network a highly effective choice in comparison to other models. The convolution component of the Temporal-spatial SPDAGG neural network is well-suited for handling noisy and distorted data to its intrinsic ability to model essential patterns and reduces random variations. Meanwhile, the transformer layers are leveraged for their capacity to manage variations in perspective and occlusions by capturing long-range dependencies and maintaining global contextual information across frames. While the transformer model alone did not surpass the state-of-the-art method, integrating it with the SPDAGG network led to an improvement in accuracy. Consequently, the combined model emerges as a superior neural network solution, yielding enhanced performance in terms of accuracy.

Method	Accuracy(%)
DGNN[16]	29.9
ST-GCN[23]	30.25
2S-AGCN[17]	34.84
HARD-Net[7]	36.97
Shift-GCN[4]	37.98
Dream-Net [22]	46.28
Transformer [20]	42.1
SPDAGG (ours)	51.24
SPDAGG-TransNet(ours)	**54,41**

TABLE II

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

2) UAV-Gesture dataset: The performance of our model on the UAV-Human skeleton database, compared to alternative models, is detailed in Table III. Our Temporal-spatial SPDAGG neural network demonstrates superior effectiveness across critical parameters. Notably, it excels in feature extraction by capturing both first-order and second-order information adeptly. Additionally, its Gauss aggregation layers proficiently handle spatial and temporal data aggregation. These combined strengths establish our Temporal-spatial SPDAGG neural network as a highly efficient choice for UAV-Gesture action recognition. Adding a transformer above leads to an enhanced performance. In addition, the results revealed that none of both Transfomer and SPDAGG network provides the best performance on this dataset. This motivates the need for their combination, resulting in our proposed SPDAGG-TransNet.

Method	Accuracy(%)
P-GNN[11]	91.9
PSD (Fitting, alignment)[19]	92.44
Transformer [20]	81.87
SPDAGG (ours)	91.24
SPDAGG-TransNet(ours)	**93,41**

TABLE III

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

V. CONCLUSION

In summary, our research endeavors in the field of UAV-human action recognition have yielded significant advancements and valuable insights. Understanding human behavior from the perspective of unmanned aerial vehicles (UAVs) poses a compelling challenge due to the dynamic and multifaceted nature of real-world scenarios. Through the development and application of our innovative SPDAGG-TransNet network, we have demonstrated effective strategies to address these challenges.

Our study underscores the critical role of skeletal-based representations in enhancing the robustness of UAV-human action recognition systems. By delving into the spatial and temporal dimensions of human actions, we have leveraged SPD matrices to establish a robust foundation for accurate and efficient action classification. The results from our evaluations, particularly on the UAV-Human and UAV-Gesture datasets, unequivocally establish the state-of-the-art performance of our method, highlighting its potential for practical applications in real-world settings.

As we continue to push the boundaries of UAV-based human behavior analysis, we intend to expand our research into enhancing the performance of tracking joints and 3D human body reconstruction from RGB sequences captured with the help of UAV, promising exciting prospects for future advancements in the field.

REFERENCES

[1] Anirudh, R., Turaga, P., Su, J., Srivastava, A.: Elastic functional coding of riemannian trajectories. IEEE transactions on pattern analysis and machine intelligence **39**(5), 922–936 (2016)

[2] Bousmina, A., Selmi, M., Ben Rhaiem, M.A., Farah, I.R.: A hybrid approach based on gan and cnn-lstm for aerial activity recognition. Remote Sensing **15**(14), 3626 (2023)

[3] Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: 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)

[4] Cheng, K., Zhang, Y., He, X., Chen, W., Cheng, J., Lu, H.: 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)

[5] Huang, Z., Wan, C., Probst, T., Van Gool, L.: 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)

[6] Huang, Z., Wu, J., Van Gool, L.: Building deep networks on grassmann manifolds. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)

[7] Li, T., Liu, J., Zhang, W., Duan, L.: 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)

[8] Li, T., Liu, J., Zhang, W., Ni, Y., Wang, W., Li, Z.: 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)

[9] 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 Recognition **124**, 108487 (2022)

[10] Nguyen, X.S., Brun, L., Lézoray, O., Bougleux, S.: 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)

[11] Perera, A.G., Wei Law, Y., Chahl, J.: Uav-gesture: A dataset for uav control and gesture recognition. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. pp. 0–0 (2018)

[12] Pham, H.H., Khoudour, L., Crouzil, A., Zegers, P., Velastin, S.A.: Video-based human action recognition using deep learning: a review. arXiv preprint arXiv:2208.03775 (2022)

[13] Qiu, H., Hou, B., Ren, B., Zhang, X.: Spatio-temporal tuples transformer for skeleton-based action recognition. arXiv preprint arXiv:2201.02849 (2022)

[14] Shaikh, M.B., Chai, D.: Rgb-d data-based action recognition: A review. Sensors **21**(12), 4246 (2021)

[15] Shi, F., Lee, C., Qiu, L., Zhao, Y., Shen, T., Muralidhar, S., Han, T., Zhu, S.C., Narayanan, V.: Star: Sparse transformer-based action recognition. arXiv preprint arXiv:2107.07089 (2021)

[16] Shi, L., Zhang, Y., Cheng, J., Lu, H.: Skeleton-based action recognition with directed graph neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7912–7921 (2019)

[17] Shi, L., Zhang, Y., Cheng, J., Lu, H.: 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)

[18] Silva, V., Soares, F., Leão, C.P., Esteves, J.S., Vercelli, G.: Skeleton driven action recognition using an image-based spatial-temporal representation and convolution neural network. Sensors **21**(13), 4342 (2021)

[19] Szczapa, B., Daoudi, M., Berretti, S., Del Bimbo, A., Pala, P., Massart, E.: 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)
[20] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
[21] Vemulapalli, R., Chellapa, R.: 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)
[22] Wang, R., Wu, X.J., Chen, Z., Xu, T., Kittler, J.: Dreamnet: A deep riemannian manifold network for spd matrix learning. In: Proceedings of the Asian Conference on Computer Vision. pp. 3241–3257 (2022)
[23] 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)
[24] Yang, J., Dong, X., Liu, L., Zhang, C., Shen, J., Yu, D.: Recurring the transformer for video action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14063–14073 (2022)