Transformers with 3D heatmaps for skeletal action recognition.

Ben Ahmed, Mohamed Amine

MT-RO 300371, mohamed.benahmed@epfl.ch

1 Introduction

The accurate recognition of human actions from video data is a pivotal challenge in computer vision, bearing significant implications for advancements in many fields such as autonomous systems. The inherent problem is the complexity of classifying human movements within diverse and often unpredictable environmental contexts.

Skeleton-based action recognition presents a promising alternative by focusing on the movement of human joints, offering a more robust approach less prone to environmental variabilities. However, the primary challenge in this approach is to accurately capture and process the dynamic nature of human motion. While existing methods have made strides in this direction, there is a critical need for more advanced techniques that can effectively interpret complex human actions with higher accuracy and computational efficiency.

This project proposes a novel solution to address these challenges. We aim to explore the integration of Transformer models, known for their exceptional performance in processing sequential data, with 3D heatmaps that effectively represent the spatial-temporal characteristics of skeletal movements. This approach is expected to harness the strengths of Transformer architectures in understanding the temporal dynamics and spatial configurations of human actions. By applying this advanced model to 3D heatmaps, this project seeks to push the boundaries of accuracy and robustness in skeletal action recognition.

2 Related work

PoseConv3D [3] was unveiled as a pioneering framework for skeleton-based action recognition, which represents a notable shift from the conventional Graph Convolutional Networks (GCNs). The innovation of PoseConv3D lies in its use of 3D heatmaps to depict human skeletons, derived from 2D pose data. This methodology offers a richer and more resilient representation of human movement, addressing several limitations of GCN-based methods.

PoseConv3D employs a 3D-CNN backbone to process these heatmaps, effectively capturing the spatiotemporal dynamics of skeleton sequences. This approach demonstrates significant improvements in handling pose estimation errors and excels in multi-person scenarios, areas where GCN methods have traditionally struggled. The paper highlights the framework's strong performance across diverse datasets, showcasing its robustness to input perturbations like dropped keypoints and its maintained accuracy with noisy or lower-quality pose data. This robustness and generalization ability mark a substantial advancement over previous models, particularly in practical, real-world scenarios.

However, a notable limitation of PoseConv3D is its reliance on a 3D-CNN architecture, rather than a Transformer-based model. While 3D-CNNs are proficient in handling spatial-temporal data, they may not capture the complex, long-range dependencies in human actions as effectively as Transformer models, which are renowned for their ability to process sequential data and uncover intricate patterns over extended temporal scales. This distinction is crucial, especially given the recent successes of Transformer models in various domains of artificial intelligence. Consequently, while PoseConv3D offers substantial improvements in robustness and generalization, the potential of a Transformer-based approach in this context remains an open and compelling area for further exploration and development.

Later MotionBERT [4] was introduced as a novel Transformer-based framework, which significantly advances the field of skeleton-based action recognition. MotionBERT employs a dual-stream spatio-temporal Transformer architecture, innovatively designed to process and analyze human motion. This architecture is composed of two distinct streams: the spatial (S) stream, which interprets the spatial relationships between skeletal joints, and the temporal (T) stream, which captures the temporal dynamics of movements. This dual-stream approach allows MotionBERT to concurrently process and integrate both spatial and temporal information, ensuring a more holistic understanding of human actions.

The model is particularly adept at capturing long-range dependencies within motion sequences, a capability that traditional methods often lack. MotionBERT's design facilitates an adaptive fusion mechanism, enabling the model to adjust the emphasis between spatial and temporal streams depending on the specific characteristics of the input data. This feature underscores MotionBERT's flexibility and its ability to handle a wide range of motion patterns and complexities.

One of the most significant advantages of MotionBERT is its exceptional ability to generalize across various human-centric video tasks, demonstrating its versatility and applicability. However, despite these strengths, MotionBERT's reliance on skeletal data presents a notable limitation. While skeletal data provides a robust and efficient way to represent human actions, it may lack certain nuances and details present in more comprehensive data representations, such as 3D heatmaps. 3D heatmaps offer a richer depiction of human movements by incorporating additional spatial information, potentially enhancing the model's ability to capture subtle and complex motion patterns.

ViViT [2] was also introduced as an advanced transformer-based model specifically designed for video classification, introducing a series of innovations that significantly enhance the analysis of video data. The ViViT model utilizes spatio-temporal tokens, a novel concept in video processing, which are encoded by transformer layers to efficiently handle the complex, multidimensional nature of video sequences.

A central innovation in ViViT is the introduction of tubelet embeddings. A tubelet in this context refers to a small, localized segment of video that encompasses a sequence of frames, capturing both spatial and temporal information. These tubelets represent a compact yet comprehensive snapshot of the video data, allowing the model to process dynamic sequences more effectively. By extending the embedding concept from the Vision Transformer (ViT) to 3D tubelets, ViViT can more adeptly fuse spatial and temporal information during tokenization, providing a richer understanding of the video content.

The paper discusses four distinct variations of the ViViT model, each with unique features to optimize the processing of video data:

- Spatio-temporal Attention Model: This model treats all spatio-temporal tokens uniformly through a single transformer encoder. It allows for a complete analysis of interactions across the entire video sequence but requires considerable computational resources.
- Factorized Encoder: To improve computational efficiency, this variant splits the encoding process into two stages, using one transformer encoder for spatial information and another for temporal information. This separation allows for more specialized and efficient processing.
- Factorized Self-Attention: This model variation refines efficiency by splitting the multi-headed self-attention mechanism within each transformer block into separate spatial and temporal components. This factorization reduces the computational load while maintaining the model's capacity to analyze complex video data.
- Factorized Dot-Product Attention: Keeping the same parameter size as the non-factorized model, this version divides the attention mechanism into spatial and temporal dimensions using different heads. This approach streamlines the attention process, enhancing efficiency without compromising performance.

ViViT's tubelet embeddings are particularly innovative, as they enable the model to capture intricate spatial and temporal patterns within video sequences. This ability is crucial for understanding the nuanced dynamics of video content, from subtle human movements to complex interactions within scenes. The various architectural optimizations in ViViT, such as the factorized encoder and attention mechanisms, demonstrate the model's versatility and efficiency in handling diverse video classification tasks. These features position ViViT as a powerful tool in the field of video understanding, capable of advancing video analysis applications across a broad spectrum of domains.

3 Method

3.1 *Overview*

This section discusses the combination of 3D heatmaps with transformer models, specifically MotionBERT and ViViT, for skeletal action recognition. This integration aims to leverage the strengths of both the transformer architecture and the rich spatial-temporal representation provided by 3D heatmaps.

3.2 *MotionBERT + 3D Heatmaps*

This approach utilizes the MotionBERT framework, enhanced with 3D heatmap inputs to enrich the spatial-temporal representation of skeletal movements. The integration with 3D heatmaps allows the Dual-stream Spatio-temporal Transformer or DSTformer (figure 1), central to the MotionBERT architecture, to effectively capture the intricacies of human motion in three dimensions.

The DSTformer architecture is adapted to process the 3D heatmap representation of skeletons. Given a sequence of 3D heatmaps that encapsulate the skeletal structure over time, the DSTformer initially projects this data into a high-dimensional feature space, $F_0 \in \mathbb{R}^{T(frames) \times J(joints) \times C_f(channel\_features)}$, enriching the input with spatial depth and temporal dynamics. To this representation, spatial positional encoding and temporal positional encoding are added, providing the model with a keen sense of joint position and frame sequence.

The DSTformer is composed of Spatial and Temporal Blocks that utilize Multi-Head Self-Attention (MHSA) mechanisms:

- **Spatial Block (S-MHSA)**: This block is responsible for understanding the spatial arrangement of the joints within each frame of the 3D heatmap, capturing the static posture and inter-joint relationships.

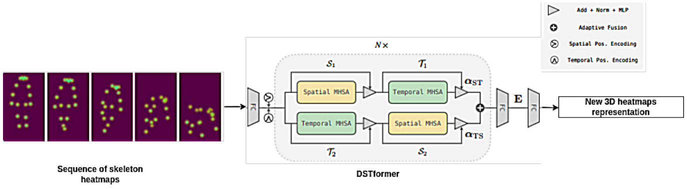

Fig. 1: Adapted Dual-stream Spatio-temporal Transformer (DST-former) [4]

- **Temporal Block (T-MHSA)**: Complementing the spatial aspect, this block analyzes the progression of the joints' positions across the sequence of heatmaps, capturing the dynamics of movement.

By processing the 3D heatmap data through these blocks, the DSTformer is able to maintain a comprehensive understanding of both the spatial configuration and the temporal movement within the action sequence.

Adaptively fusing the outputs from the Spatial and Temporal Blocks, the DSTformer's architecture employs a series of N dual-stream-fusion modules. Within each module, spatial and temporal MHSAs are strategically stacked in alternating sequences, forming two parallel computational pathways. This configuration allows the DSTformer to synthesize the information from both streams, yielding a richly detailed motion representation.

The DSTformer processes the 3D heatmaps through these pathways, with each subsequent layer refining the feature embeddings. The final embedding layer, $E \in \mathbb{R}^{T \times J \times C_e(channel\_embeddings)}$, is then linearly transformed to estimate the 3D motion $\hat{X} \in \mathbb{R}^{T \times J \times C_{out}(channel\_output)}$. The outcome is a profound motion representation that encapsulates the depth and complexity of human actions as manifested in the 3D heatmaps.

3.3 *ViViT + 3D Heatmaps*

For this method we rely mainly on the tubelet embedding technique adopted by ViViT [2] and we test it on two model variants out of the four presented in ViViT.

3.3.1 Tubelet Embedding

The tubelet embedding process is a critical step in adapting the ViViT architecture for 3D heatmap inputs in skeletal action recognition. This method involves segmenting the input volume into non-overlapping, spatio-temporal tubelets that span across both spatial dimensions (height and width of the frame) and the temporal dimension (time across frames). These tubelets are then linearly projected into a higher-dimensional token space, akin to performing a 3D convolution. The result is a series of tokens that carry fused spatio-temporal information, providing a rich and detailed representation of the action sequences. This embedding technique enables the transformer to consider both the spatial arrangement of the joints and their temporal movement, which is crucial for capturing the dynamics of human motion.

3.3.2 Spatio-temporal Attention Model

The Spatio-temporal Attention Model processes the entire sequence of spatio-temporal tokens extracted from the tubelet embeddings. By forwarding these tokens through a transformer encoder, the model captures pairwise interactions and long-range dependencies across the entire sequence of actions. Each transformer layer within this model is capable of modeling interactions that cover the full extent of the video, learning from complex patterns that emerge over the course of the action sequence. This model variant is particularly powerful as it can discern intricate details and subtle variations within the movements, owing to its comprehensive attention mechanism that spans all tokens.

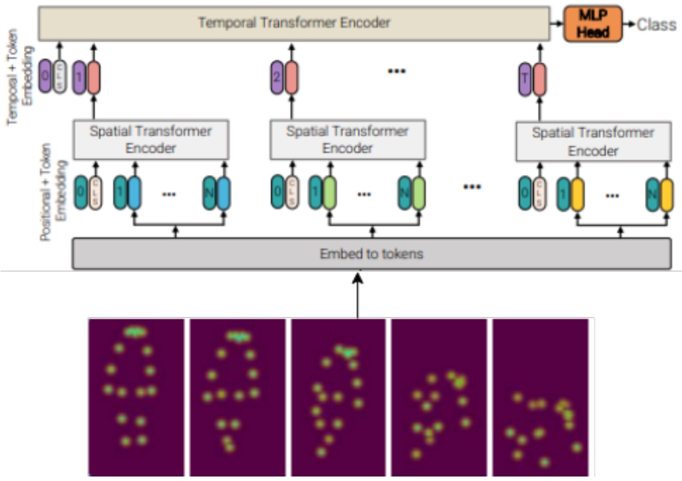
Temporal Transformer Encoder
MLP Head
Class
Temporal + Token Embedding
Spatial Transformer Encoder
Spatial Transformer Encoder
Spatial Transformer Encoder
Positional + Token Embedding
Embed to tokens

Fig. 2: Factorised Encoder [4]

3.3.3 Factorized Encoder

The Factorized Encoder variant of ViViT shown in figure 2 proposes an efficient two-stage architecture that processes spatial and temporal information separately. In the first stage, the spatial encoder focuses exclusively on modeling the interactions between tokens that are temporally aligned within the video sequence. Specifically, it generates a high-dimensional representation $h_i \in \mathbb{R}^d$ for each temporal step after processing through L_s layers.

Once the spatial information is encoded into these frame-level representations, they are concatenated to form $H \in \mathbb{R}^{n_t \times d}$, where n_t is the number of time steps. This concatenated representation is then fed into the temporal encoder, which consists of L_t transformer layers. The temporal encoder is responsible for modeling the interactions across different temporal indices, capturing the temporal dynamics of the skeletal movements.

This architecture mirrors a "late fusion" approach, first developing a deep spatial understanding before integrating the temporal dimension, leading to an efficient yet comprehensive analysis of the action sequences.

4 Experiment

The experiments were conducted to evaluate the performance of various models on the NTU RGB+D dataset, a comprehensive action recognition dataset consisting of 60 action classes and 56,880 video samples. The primary evaluation metric used is accuracy at one (*acc@1*), focusing on the model's ability to correctly identify the top-most likely action class.

4.1 Dataset and Evaluation Metrics

The NTU RGB+D dataset provides a challenging and diverse environment for testing action recognition models, offering a wide range of human actions captured in different modalities. Originally, this dataset has two variants: cross-subject (xsub) and cross-view (xview). For the purposes of our study, we specifically utilize the xsub dataset.

4.2 Training Setup and Purpose

The following implementations use 3D heatmap inputs with shape (frames (T), height (H), width(W)) = (48, 56, 56) for MotionBERT and (T, H, W) = (56, 56, 56) for ViViT.

- **MotionBERT**: Trained to establish baseline performance using 3D heatmaps with state-of-art transformer for skeletal action recognition.
- **MotionBERT with Augmentation**: Trained with data augmentations (Rotation, Gaussian noise and temporal shift) to enhance robustness against input variations and try to improve the model performance.
- **ViViT (Model 1 and 2)**: Trained to compare the efficiency and effectiveness of different architectural designs.
- **Bigger ViViT (Model 1)**: Trained with bigger architecture to investigate its impact on performance.
- **Smaller Tublelet ViViT (Model 1)**: As mentioned in the original paper [2] using smaller tubelet size (for the temporal dimension only) can increase the model's performance. In this training a smaller tubelet size is used (4,8,8) compared to the originally tubelet size used of (8,8,8) .
- **Vanilla Transformer** [1]: Served as a control to benchmark the specialized architectures.
- **Bigger Vanilla Transformer**: An upscaled model to explore the relationship between model size and performance.
- **Vanilla Transformer with Augmentation**: To examine the impact of data augmentations (Rotation, Gaussian noise and temporal shift).

The trainings with different data representations aimed to highlight the advantages of 3D heatmaps in skeletal action recognition.

Model	**2D Keypoints**	**3D Keypoints**	**3D Heatmaps**
MotionBERT	87.7%	74.2%	87.8%
MotionBERT + Aug	-	-	87.2%
ViViT (M1)	-	-	81.3%
Big ViViT	-	-	79.6%
Small Tub ViViT	-	-	82.3%
ViViT (M2)	-	-	80.9%
Vanilla	78.8%	76.7%	82.2%
Big Vanilla	-	-	83.4%
Vanilla + Aug	-	-	84.8%

Table 1 Training results on the NTU RGB+D dataset.

Below is a table representing the number of parameters for each model:

Model	**Number of Parameters in 10^6**
MotionBERT	63.5
MotionBERT + Aug	63.5
ViViT (M1)	8.65
Big ViViT	63.7
Small Tub ViViT	7.6
ViViT (M2)	44.4
Vanilla	9.5
Big Vanilla	94.7
Vanilla + Aug	94.7

Table 2 Number of parameters for each model used.

5 Discussion

The experiment results, as detailed in Tables 1 and 2, offer critical insights into the efficacy of 3D heatmaps in skeletal action recognition, particularly when compared against the current state-of-the-art model, MotionBERT with 2D keypoints.

5.1 Performance Analysis

1. Minor Improvement with MotionBERT: The objective to surpass the state-of-the-art accuracy of MotionBERT using 2D keypoints was met with a marginal improvement of just 0.1% when using 3D heatmaps. This modest increase suggests that while 3D heatmaps provide a richer representation of spatial-temporal data, MotionBERT's architecture might not fully exploit the additional depth of information available in 3D heatmap inputs.

2. Performance of ViViT: Despite being a sophisticated video vision transformer, ViViT showed lower performance compared to the state-of-the-art DSTformer architecture of MotionBERT. Even with the introduction of tubelet embeddings, which are designed to enhance spatial-temporal understanding, ViViT did not manage to outperform MotionBERT.

3. Impact of 3D Heatmaps on the Vanilla Transformer: A significant revelation from the experiments was the performance of the Vanilla Transformer. The introduction of 3D heatmaps led to a notable increase in accuracy by 3.4%. This substantial improvement underscores the impact of 3D heatmaps in enhancing the model's ability to understand and classify complex human actions. The simpler architecture of the Vanilla Transformer, compared to more specialized models like MotionBERT and ViViT, surprisingly benefited more from the depth and richness of 3D heatmaps. This finding suggests that even basic transformer models can achieve significant gains in action recognition tasks when equipped with more sophisticated input representations.

5.2 Model Complexity and Efficiency

The number of parameters, as shown in Table 2, indicates a wide variation in model complexity. While MotionBERT, with its larger size, could only marginally exploit the 3D heatmap advantage, the Vanilla Transformer demonstrated that even less complex models could significantly benefit from enhanced input data. This observation points towards the potential of achieving efficient yet effective skeletal action recognition models by carefully balancing architectural complexity with the nature of input data.

6 Conclusions

This project embarked on an exploration of 3D heatmaps in skeletal action recognition, comparing their effectiveness across various transformers. While the integration into state-of-the-art models like MotionBERT yielded only a slight improvement, a notable enhancement was observed in simpler models like the Vanilla Transformer. This outcome highlights the dependency of 3D heatmap efficacy on the specific architecture of the models.

Our experiments have demonstrated that while 3D heatmaps hold promise, their full potential is yet to be harnessed in complex action recognition models. Future work should thus focus on developing and fine-tuning model architectures that can more effectively utilize the spatial-temporal depth provided by 3D heatmaps.

References

[1]Shazeer Ashish Vaswaniand Noam et al. "Attention Is All You Need". In: *NeurIPS*. 2017.

[2]Arnab Anurag et al. "ViViT: A Video Vision Transformer". In: *ICCV*. 2021.

[3]Haodong Duan et al. "Revisiting Skeleton-based Action Recognition". In: *CVPR*. 2022.

[4]Zhu Wentao et al. "MotionBERT: A Unified Perspective on Learning Human Motion Representations". In: *ICCV*. 2023.