PoseFusion: Multi-View Pose Integration for Comprehensive Action Recognition

Aniruddh Balram
University of Maryland
abalram1@umd.edu

Tharun Puthanveettil
University of Maryland
tvpian@umd.edu

Satish Vennapu
University of Maryland
satish@umd.edu

Sai Surya Sriramoju
University of Maryland
saisurya@umd.edu

Abhijay Singh
University of Maryland
abhijay@umd.edu *

Abstract

Human Action Recognition (HAR) is a critical domain with diverse applications, spanning human-computer interaction, surveillance, and augmented/mixed reality. This paper addresses the complex challenges posed by occlusions, viewpoint variations, and the synthesis of aggregate embeddings to enhance the accuracy of action recognition. Conventional methodologies rely on unified representations derived from multiple views, encountering limitations in scenarios characterized by partial occlusion. Our innovative approach strategically incorporates semantically rich partial information from each view, yielding a nuanced aggregate embedding conducive to robust human action recognition. This paradigm challenges the conventional assumption of a unified representation and showcases the adaptability of attention mechanisms. Comparative evaluations against single-view methodologies on occluded datasets underscore the limitations of traditional techniques in effectively handling occlusions. The proposed multi-view approach consistently surpasses its single-view counterpart, providing empirical evidence that aggregating partial or complete information from multiple views effectively addresses the shortcomings inherent in single-view-based techniques, particularly in challenging occluded scenarios. This study contributes substantive insights to the HAR field, challenging conventional embedding methodologies and laying the groundwork for the development of more nuanced and intricate spatio-temporal action recognition systems.

1. Introduction

Human Action Recognition (HAR), an actively researched area, holds intriguing applications in fields like human-computer interaction, surveillance, and AR/MR. An action is defined as a fundamental motion pattern, which can be as subtle as taking a step, while an activity encompasses a series of actions. Classifying actions based on motion patterns involves significant computational overhead. Neurobiological studies have indicated that only a few key points in the human body are necessary for action classification, commonly referred to as human pose [17]. Hence, accurate Human Pose Estimation (HPE) is a crucial precursor to action prediction.

Factors like viewpoint variation, varying lighting conditions, and occlusions from specific perspectives can introduce challenges in correlating frames, leading to distinctiveness in actions that are actually similar[11], [6]. Some studies have explored the use of 3D poses for pose estimation. However, obtaining in-the-wild image datasets with reliable 3D ground truth proves to be a formidable task. Additionally, 2D poses inherently exhibit ambiguity. A single 3D pose can project onto multiple similar 2D poses, making the deterministic embedding of 2D poses a complex undertaking [14].

Human action recognition can be inherently ambiguous when observed from multiple views due to partial visibility, occlusions, viewpoint disparities, and variations in lighting conditions. Despite these inherent challenges in images, humans possess the remarkable capacity to discern similar actions. While it is a crucial task for a computer to classify images that share semantic similarities under the same label, it is imperative to note that semantic similarity does not imply identical actions. Consequently, recognizing actions from a single image presents a non-trivial challenge.

Humans excel in recognizing actions from a single perspective due to their innate comprehension of the scene and the action. They possess the ability to mentally reconstruct the past and anticipate the future from a static vantage point, and can even extrapolate the action to encompass multiple views. Furthermore, humans tend to inspect an action from

*https://github.com/abhijaysingh/HPE-for-HAR

alternative angles to enhance their understanding, thereby amalgamating insights gleaned from diverse perspectives. This practice aids in aggregating a comprehensive understanding of the action.

The primary challenge addressed in this study pertains to the synthesis of an aggregate embedding conducive to accurate action recognition by leveraging non-occluded or partially occluded views of action. Existing research endeavors have often sought to establish a unified combined embedding for human poses from multiple views to facilitate activity recognition. However, our intuition posits that the pursuit of such a unified representation may be inherently flawed, particularly when one or more views encompass partial occlusion.

In this work, our objective is to devise an innovative approach that semantically incorporates partial information regarding the pose of the target from each view. This augmentation aims to generate an aggregate embedding capable of furnishing precise human action recognition, even in scenarios involving partial occlusion.

2. Background

Research in human action recognition was an integral part in the computer vision community since before the boom of AI. Early research included designing handcrafted features to train a linear classifier to predict action[11]. With the advent of deep learning, several methods were introduced for the task of action recognition. Initially, deep learning architectures such as CNNs[4] and RNNs[9] were used to classify human actions based on RGB/Depth image sequences or by a meaningful multi-modal fusion[18]. Sequences had to be short due to inherent vanishing gradient and exploding gradient problems in RNNs. With the introduction of Vision Transformer(ViT)[3], transformers could be integrated with CNNs to handle temporal information which overcame all the issues related to RNNs. For action recognition, as there can be changes in body scale, occlusion, viewpoint variation, and fast motion between frames; CNNs fail to capture the important features. Besides, for high-resolution videos; CNNs can be computationally burdening.

Skeletal graphs of humans are computationally efficient and maintain structural information of humans across frames without an effect from background variables[16]. Skeleton-based action recognition captures meaningful temporal sequences of human actions, outperforming CNN-based methods. Therefore, much of the research in human action recognition is skeleton based. Several techniques involving CNNs for skeletal data have been explored[1, 10].CNNs are popular for encoding spatial information, but since a skeleton is just a graph, CNNs aren't a natural fit for extracting features from a skeleton. Therefore, newer deep learning frameworks like Graph Convolutional Networks (GCNs) have become fairly popular in tackling action recognition with skeletal data.[19] utilizes GCNs for encoding both spatial and temporal information achieving better performance than CNN-based methods. For encoding temporal information, the state-of-the-art architecture is transformer[15]. Methods fusing transformer architecture to encode temporal information and GCN to encode graph information [20] outperforming all the previously explored methods. However, the real bottleneck of skeleton-based methods lies in the accuracy of pose estimation. Therefore, accurate human pose estimation is necessary for high quality action recognition.

Error-free estimation of human pose is a challenging problem. 2D human poses face challenges due to viewpoint variations, and noise and are a valid estimation from a single viewing direction. Recovering a 3-D pose from a 2D pose is ambiguous as 2D poses can correspond to varied 3D poses. Direct 3D pose estimation from single-view faces further challenges due to the lack of availability of in-the-wild datasets. State-of-the-art datasets for human pose estimation such as human3.6m[7] are collected under controlled environments. Curating in-the-wild datasets is a cumbersome task and may not generalize well to applications outside of which it was curated. Certain established models such as PoseNet[8] and OpenPose[2] produce reliable estimates of 3D poses but may produce estimates with low confidence under severe noise and occlusions. Errors in pose estimation can result in poor prediction of actions.

The impact of incorporating views from multiple cameras on action recognition models is a challenging aspect addressed in this study. While existing approaches, such as contrastive learning[12] and cross-view prediction[5], have demonstrated effectiveness in handling viewpoint variations in multiview action recognition, their reliance on image sequences limits their applicability. A notable gap exists in the exploration of recognizing actions from multiview skeletal pose data, where poses may be incomplete due to occlusions.

In response to this gap, our study proposes a novel architecture based on Graph Convolutional Networks (GCN) and Transformers. This architecture aims to aggregate high-dimensional embeddings derived from Multi-view 3D poses, specifically tailored to enhance action recognition in occluded scenarios. By leveraging the complementary information from multiple views, our approach seeks to address the challenges posed by occlusions in skeletal pose data, offering a promising avenue for improving the robustness of action recognition models in complex visual environments.

3. Methodology

This research endeavors to address the challenge of action recognition in scenarios where the target's pose is subject to partial occlusion in one or more views. The premise is that

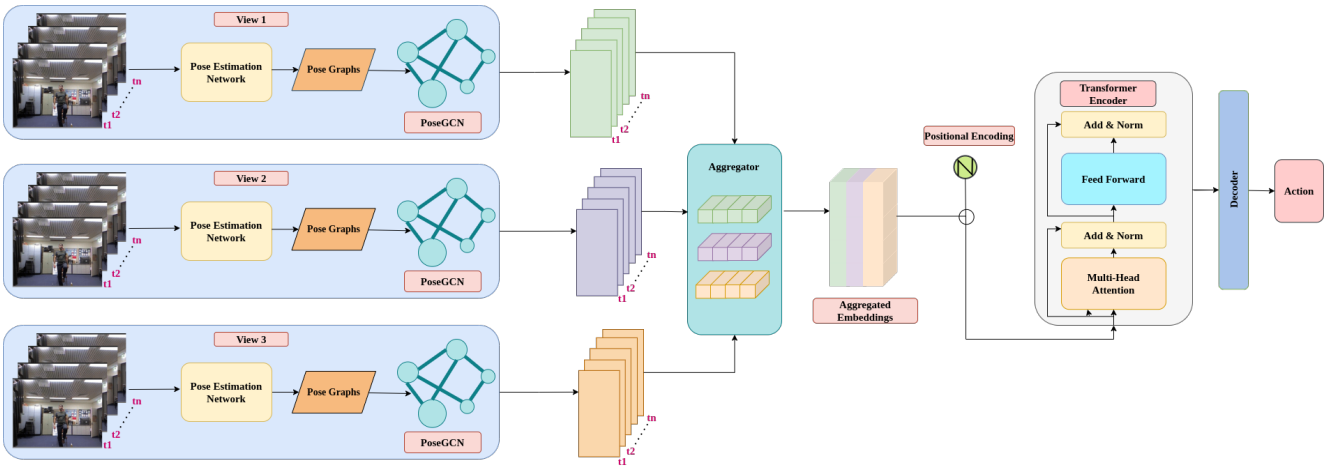


Figure 1. Process flow using Multi-View Pose Aggregation Network(MV-PAN)

a unified embedding, intended to be universally applicable for pose estimates from multiple views, may prove inadequate for action recognition tasks in such instances. Consequently, the intuitive approach we proposed is wherein the aggregation of embedding representations from each view is conducted in a sequential manner. This is deemed essential to enhance the efficacy of action recognition in the presence of occluded pose information.The key steps in the process flow of our proposed pipeline (as shown in Fig.1) are as follows:

1. Data Pre-Processing
2. Spatial Encoding using GCNs
3. Aggregation of View-Specific Embeddings
4. Temporal Encoding with Transformer Encoder
5. Softmax-based Multi-Class Classification

3.1. Data Processing

Dataset: NTU RGB+D dataset containing 56,880 samples was used. It includes 60 action classes, including daily behaviors and health-related actions, performed by 40 participants.

3.1.1 Offline Preprocessing

The primary preprocessing step involves sampling the action sequence video to obtain a sequence of frames corresponding to the action classes of interest. Subsequently, we employ a Human Pose Estimation algorithm, such as the one provided by Mediapipe, to extract keypoints specific to each frame. We store the resulting sequence of keypoints locally for each view within a given action sequence, facilitating their subsequent utilization in the model training pipeline.

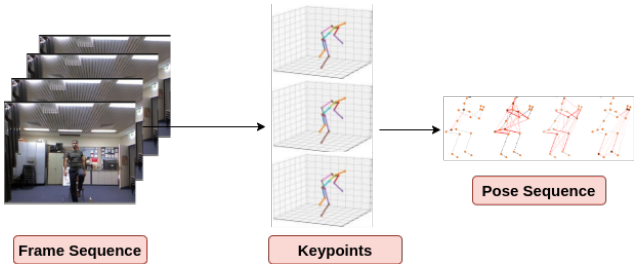


Figure 2. Offline preprocessing pipeline

3.1.2 Data Batching

The data is passed to the multiview action classifier in mini-batches, where each batch includes multiple pose graph sequences. Within a batch, each sequence corresponds to a set of pose graphs captured from different views. The sequences are structured to preserve temporal relationships. To ensure consistency across sequences within a batch, padding is applied to standardize sequence lengths. During training, these batches of pose graph sequences are loaded and processed concurrently, harnessing parallelism for efficient model training.

3.2. Spatial Encoding using GCN

In the initial phase of our methodology, we employ a Graph Convolutional Network (GCN)-based model to perform spatial encoding. This model is designed to encode spatial information derived from the sequence of key points observed in each view. The primary focus of this stage is to capture the inherent spatial dynamics present in the observed poses from individual views and create a view-specific embedding. The GCN is strategically utilized to ensure an effective representation of spatial relationships within the key point sequences.[21]. In Fig.3, the PoseGCN

is depicted in action, showcasing the creation of view-specific embeddings. These embeddings are generated by the PoseGCN from the pose graphs, which, in turn, are derived from the estimated body pose keypoints and their corresponding edge indices (representing links between body joints).

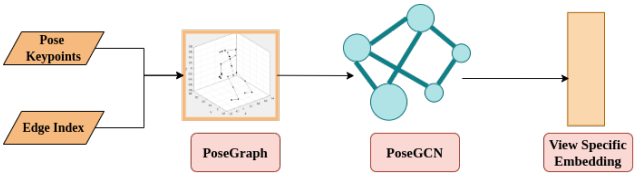


Figure 3. Synthesis of view-specific embedding using PoseGCN

3.3. Aggregation of View-Specific Embeddings

Following the spatial encoding stage, the embeddings that encode spatial information from each view undergo a systematic process known as semantic aggregation. This aggregation process is designed to systematically combine the view-specific embeddings, aiming to generate a comprehensive embedding that provides a more holistic representation of the target's pose across multiple views for a given instant. The consideration of partial pose information from individual views is integral to this aggregation, ensuring a robust and inclusive representation. Three aggregation techniques were employed and tested for performing semantic aggregation.

1. **Mean/Average Aggregation:** The mean of the three embeddings is computed, resulting in a new embedding of size n.

- Let E_1, E_2, E_3 represent the three embeddings of size n. The mean aggregation is defined as:

$$\text{Mean Aggregation}(E_1, E_2, E_3) = \frac{E_1 + E_2 + E_3}{3}$$

where the output size remains n.

2. **Linear Layer Aggregation:** The three embeddings are concatenated into a single-column vector, which is then passed through a linear layer to obtain a new embedding of size n.

- Concatenate the embeddings into a column vector C of size $3n$:

$$C = \begin{bmatrix} E_1 \\ E_2 \\ E_3 \end{bmatrix}$$

- Pass the concatenated vector through a linear layer with weight matrix W of size $n \times 3n$ and bias vector b of size n:

$$\text{Linear Layer Concatenation}(E_1, E_2, E_3) = W \cdot C + b$$

where the output size is n.

3. **Self-Attention Aggregation:** The embeddings are concatenated and fed into a self-attention network. The output of the self-attention mechanism is projected using a linear layer, yielding a final aggregated embedding of size n.

- Concatenate the embeddings into a matrix M of size $n \times 3$:

$$M = \begin{bmatrix} E_1 & E_2 & E_3 \end{bmatrix}$$

- Pass M through a self-attention mechanism, resulting in A of size $n \times 3$:

$$A = \text{Self-Attention}(M, M, M)$$

where, each of the Ms, in order represents Query, Key, and Value respectively.

- Project the output using a linear layer with weight matrix W of size $n \times n$ and bias vector b of size n:

$$\text{Self-Attention Aggregation}(E_1, E_2, E_3) = W \cdot A + b$$

where the output size is n.

3.4. Temporal Encoding using Transformers

The temporal encoding stage involves the application of higher-dimensional positional encoding to the aggregated embeddings at each time instant. This process enhances the temporal information within the embeddings. Subsequently, the positionally encoded aggregated embeddings traverse through a Multi-head attention-based Transformer encoder layer. This sophisticated layer is specifically chosen to obtain a single higher-dimensional temporal embedding representing the flow of the target's poses over time within the action sequence. The transformer architecture is leveraged for its effectiveness in capturing complex temporal dependencies and patterns.

3.5. Softmax-based Multi-Class Classification

The resulting higher-dimensional temporal encoding of aggregated spatial embeddings becomes the focal point for the final stage of our methodology—softmax-based multi-class classification. In this stage, we execute a multiclass classification leveraging the enriched information captured through sequential aggregation. The comprehensive embedding obtained from the previous stages proves instrumental in facilitating robust action recognition. This final stage is particularly crucial for scenarios marked by partial occlusion of the target's pose in one or more views, where the aggregated information ensures a resilient and accurate classification outcome.

This proposed methodology integrates spatial information aggregation with advanced temporal encoding techniques to bolster action recognition in multi-view environments, particularly when posed with challenges such as partial occlusion. The sequential fusion of view-specific embeddings ensures a more resilient and comprehensive rep-

resentation for performing action recognition for partially occluded views of the target pose.

4. Experimental Setup

The dataset consisted of 1868 samples per view, distributed across training, validation, and test sets. Within each set, 1401 samples were designated for training, 351 for validation, and 116 for testing. Due to resource constraints, the training process focused on a specific subset comprising 6 action classes, carefully selected from a larger pool of 60 action classes available in the NTU dataset[13] (Table 1). All subsequent evaluations and analyses were conducted exclusively within the context of this constrained subset.

Class	Action
0	Standing Up
1	Falling
2	Hopping
3	Wearing shoes
4	Taking off shoes
5	Jump up

Table 1. Action Classes

The training regimen spanned 50 epochs, utilizing a batch size of 32. An initial learning rate of 0.0001 was employed, complemented by a ReduceOnPlateau scheduler to optimize the training process.

5. Results

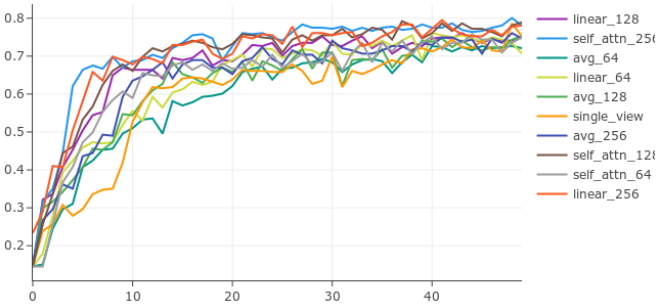
5.1. Evaluation Metrics

In evaluating the efficacy of the proposed human action recognition framework, treated as a multi-class classification problem, the key performance metrics employed to assess the model's accuracy, robustness, and generalization capabilities are as follows:

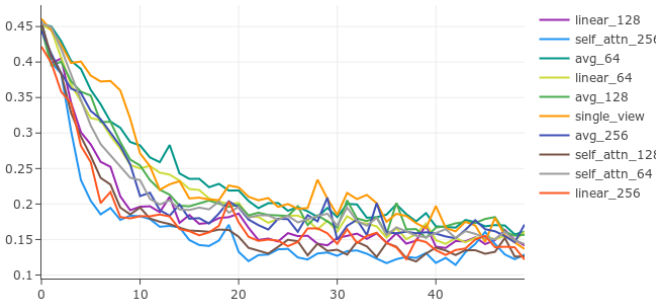
1. **Accuracy:** Accuracy is the main statistic for assessing the classification performance of the model. It displays how accurate action predictions have been overall across all classes.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

2. **Confusion Matrix:** To provide a thorough examination of the model's predictions, a confusion matrix will be created. It will highlight true positives, false positives, true negatives, and false negatives for each action type. This makes it easier to pinpoint specific areas where the model may struggle or perform well.
3. **ROC Curve:** The Receiver Operating Characteristic (ROC) curve illustrates the trade-off between sensitivity and specificity across various threshold values. It is valuable for understanding the model's discrimination performance.
4. **Sample-Weighted Precision, Recall, and F1 Score:** Sample-weighted precision, recall, and F1 score take into account class imbalances by assigning weights based on the number of samples available in each class. These metrics provide a more nuanced evaluation, particularly in scenarios where classes have unequal representation.



(a) Validation accuracy per epoch



(b) Validation loss per epoch

Figure 4. Validation Accuracy and Loss of MV-PAN on Occluded dataset

$$\text{Sample-Weighted Precision} = \frac{\sum_i n_i \cdot \text{Precision}_i}{\sum_i n_i}$$

$$\text{Sample-Weighted Recall} = \frac{\sum_i n_i \cdot \text{Recall}_i}{\sum_i n_i}$$

$$\text{Sample-Weighted F1 Score} = \frac{\sum_i n_i \cdot \text{F1 Score}_i}{\sum_i n_i}$$

where n_i represents the number of samples in class i.

5.2. Evaluation

Fig.4a and Fig.4b present the Accuracy and Loss curves corresponding to each model listed in Table 3. The consistent convergence evident in these graphs reaffirms the robust learning capacity of the proposed model. Despite variations in performance based on the choice of aggregation

Technique	Precision			Recall			F1			Accuracy		
	64	128	256	64	128	256	64	128	256	64	128	256
Mean	0.781	0.792	**0.789**	0.775	0.775	**0.793**	0.771	0.768	**0.773**	0.775	0.775	**0.793**
Linear	0.788	**0.865**	0.808	0.775	**0.862**	0.801	0.776	**0.863**	0.792	0.775	**0.862**	0.801
Self-Attention	0.818	0.826	**0.867**	0.784	0.819	**0.853**	0.786	0.814	**0.852**	0.784	0.819	**0.853**

Table 2. Normal Dataset: Metrics-based comparison based on the different combinations of aggregation techniques & embedding size.

Technique	Precision			Recall			F1			Accuracy		
	64	128	256	64	128	256	64	128	256	64	128	256
Mean	**0.745**	0.736	0.712	**0.750**	0.741	0.724	**0.743**	0.721	0.704	**0.750**	0.741	0.724
Linear	0.782	0.789	**0.820**	0.775	0.784	**0.819**	0.776	0.782	**0.819**	0.775	0.784	**0.819**
Self-Attention	0.774	**0.816**	0.801	0.775	**0.819**	0.801	0.766	**0.817**	0.800	0.775	**0.819**	0.801

Table 3. Occluded Dataset: Metrics-based comparison based on the different combinations of aggregation techniques & embedding size.

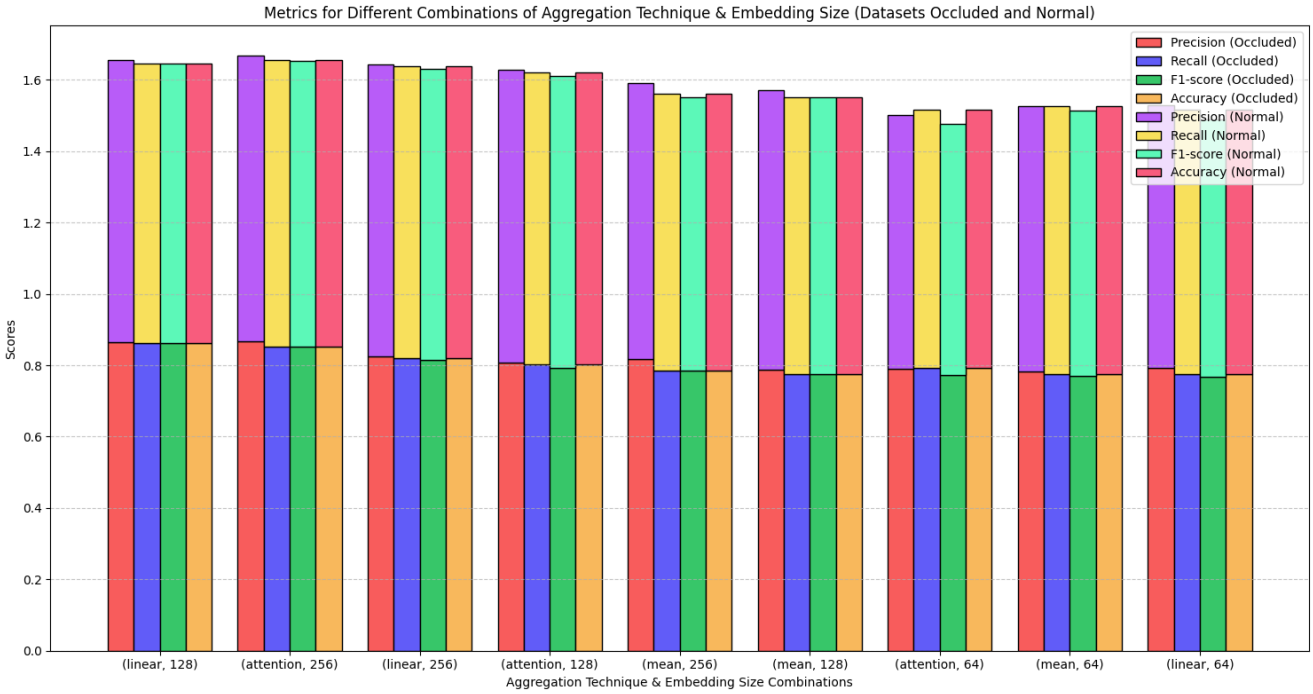
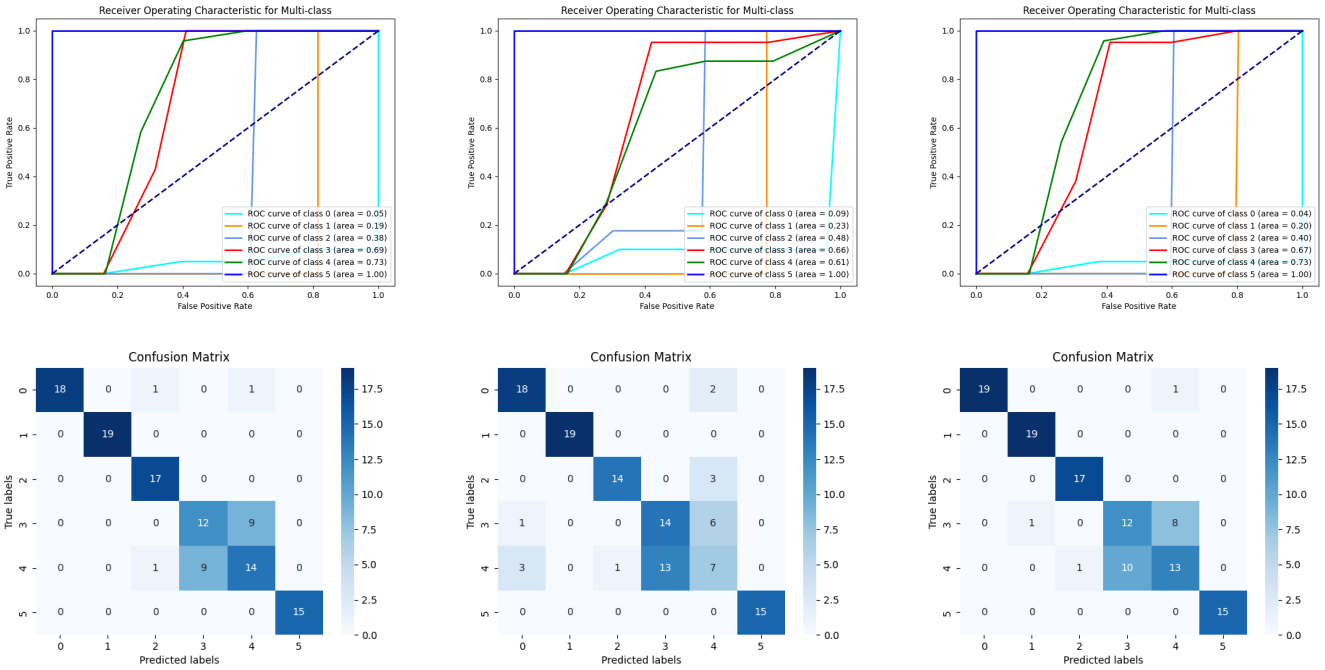


Figure 5. Metrics-based comparison between Occluded and Normal dataset based on the different combination of aggregation techniques and embedding size.(ascending order of accuracy(normal)

techniques and embedding size, the model demonstrates a notable ability to comprehend, learn, and discriminate between the underlying characteristics of each action class.

In the occluded dataset, as presented in Table 3, the Linear aggregation technique with an embedding size of 256 demonstrated marginal improvement over the Self-Attention-based aggregation technique. Conversely, the Mean aggregation technique exhibited the poorest performance. This observed behavior can be attributed to the inherent nature of Linear and Attention-based aggregation techniques, which operate based on learned weights, as opposed to Mean Aggregation, which yields an absolute embedding. It is reasonable to assume that the Linear and Self-Attention-based techniques adapt and extract information



(a) ROC and Confusion Matrix for linear-256 (b) ROC and Confusion Matrix for mean-64 (c) ROC and Confusion Matrix for self-attention-128

Figure 6. Comparison between ROC Curves and AUC Curves of the top-3 best performing combination of Aggregation techniques and Embedding size.

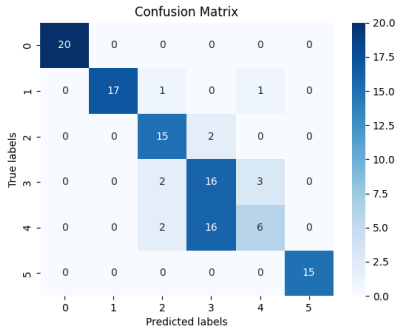


Figure 7. Single-view Confusion Matrix

from unoccluded views to compensate for the information loss from occluded views.

Dataset	Precision	Recall	F1	Accuracy
Occluded	0.7848	0.7672	0.7536	0.7672
Original	0.8743	0.8707	0.8692	0.8707

Table 4. Metrics-based comparison of single-view on original and occluded datasets.

From Fig.5, the impact of hyperparameter combinations on model performance is evident across the original and occluded datasets. In the occluded dataset, the *(mean, 64)* combination achieves a balance between precision and recall, while *(linear, 128)* prioritizes precision at the expense of recall. Notably, *(attention, 256)* consistently outperforms, leveraging the larger embedding size for focusing on relevant features in occluded data. Conversely, simplicity proves effective for *(mean, 64)*, and *(linear, 128)* maintains consistent high performance in the normal dataset. The attention mechanism *(attention, 256)* excels, showcasing its ability to capture complex relationships.

In examining specific models, the one emphasizing accuracy in the normal dataset *(mean, 64)* encounters a significant drop in accuracy when faced with occluded data, indicating sensitivity to challenges introduced by obstructed views. Similarly, the model prioritizing precision in the normal dataset *(linear, 128)* faces difficulties in the occluded dataset, achieving a balance between precision and recall but at the cost of overall accuracy. In contrast, the model with the highest performance in the normal dataset (attention, 256) demonstrates resilience in the occluded dataset, maintaining relatively high precision, recall, and accuracy. This underscores models' sensitivity to occlusions and the trade-offs between precision and recall. The attention mechanism proves crucial in ensuring consistent performance across datasets, emphasizing the importance

of careful hyperparameter tuning and model optimization.

Analysis of Fig.6 suggests that while the *(linear, 256)* combination achieved an overall higher accuracy and F1 score, the *(attention-128)* combination exhibited a more balanced performance. Particularly for similar action classes such as *wearing shoes* and *taking off shoes*, the attention-based aggregation proved to be more robust to occlusion. Notably, when compared to the Linear and Attention-based aggregation techniques, even the best-performing Mean aggregation combination *(mean-64)* was observed to generate more false positives and fewer true positives.

As depicted in Fig.7, the performance of the single-view action recognition model notably diminishes in the presence of occlusion compared to the Multi-view action recognition model. This observation underscores the efficacy of the proposed Multi-view pose aggregation model, which demonstrates the capacity to glean partial or complete information from non-occluded views to mitigate any information loss resulting from occlusion in another view.

Table 5 illustrates the model's performance when one view is occluded while the other two remain unobstructed. Consistent performances across all metrics are observed irrespective of the occluded view. This again suggests that the proposed pipeline adeptly extracts information from the non-occluded views to compensate for the loss of information in the occluded ones.

The examination of the table 4 reveals that the performance of the Multi-View Pose Aggregation Network (MV-PAN) in the occluded dataset attains an 82% accuracy and 82% F1 score. In contrast, its single-view counterpart achieves 75% accuracy and 76% F1 score. This reaffirms the efficacy of our proposed pipeline in aggregating partial information from multiple views to enhance decision-making capabilities especially in scenarios with partial visibility or occlusion.

Occluded View	Precision	Recall	F1	Accuracy
View 1	0.8296	0.8276	0.8276	0.8276
View 2	0.8178	0.819	0.8174	0.819
View 3	0.8346	0.8362	0.8348	0.8362

Table 5. Generalization across different views

6. Conclusion

In conclusion, our investigation into Human Action Recognition (HAR) addressed a pivotal challenge in crafting a robust aggregate embedding for precise action recognition, particularly in scenarios marked by partial occlusion and varying perspectives. Our work is grounded in the recognition that a unified representation of human poses from multiple views might be inherently flawed when confronted with partial occlusion. Our innovative approach overcomes this challenge by incorporating partial information from each view, fostering a nuanced aggregate embedding.

The obtained results validate the efficacy of our approach in mitigating the challenges posed by occlusions and viewpoint variations. The nuanced examination of hyperparameter combinations highlighted the adaptability of attention mechanisms and the inherent trade-offs between precision, recall, and accuracy. Our model demonstrated resilience in maintaining robust performance across diverse datasets, underscoring the practical applicability of our proposed method.

Significantly, the comparative assessment against a single-view approach on the occluded dataset emphasized the limitations of traditional techniques in handling occlusions. Our proposed multi-view approach consistently outperformed its single-view counterpart with an average score of 82% accuracy for 6 action classes, reinforcing our assertion that the aggregation of partial or full information from multiple views can effectively address the deficiencies of single-view-based techniques, particularly in scenarios where occlusion poses a significant challenge.

The findings of this study contribute to the broader field of HAR, shedding light on the intricacies of managing partial information for action recognition. By challenging the conventional notion of a unified embedding and proposing an approach that embraces the inherent ambiguity of occluded views, our work sets the stage for more nuanced and intricate spatio-temporal action recognition systems. This innovative perspective opens avenues for future research aimed at refining human pose estimation and action recognition in real-world scenarios characterized by prevalent occlusions and viewpoint variations.

References

[1] Carlos Caetano, Jessica Sena, François Brémond, Jefersson A. dos Santos, and William Robson Schwartz. Skelemotion: A new representation of skeleton joint sequences based on motion information for 3d action recognition. *CoRR*, abs/1907.13025, 2019. 2

[2] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *CoRR*, abs/1812.08008, 2018. 2

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. 2

[4] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. *CoRR*, abs/1604.06573, 2016. 2

[5] Gaurvi Goyal, Nicoletta Noceti, and Francesca Odone. Cross-view action recognition with small-scale datasets. *Image and Vision Computing*, 120:104403, 2022. 2

[6] Chih-Hui Ho, Pedro Morgado, Amir Persekian, and Nuno Vasconcelos. Pies: Pose invariant embeddings. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[7] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014. 2

[8] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Convolutional networks for real-time 6-dof camera relocalization. *CoRR*, abs/1505.07427, 2015. 2

[9] Inwoong Lee, Doyoung Kim, Seoungyoon Kang, and Sanghoon Lee. Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1012–1020, 2017. 2

[10] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. *CoRR*, abs/1804.06055, 2018. 2

[11] Greg Mori, Caroline Pantofaru, Nisarg Kothari, Thomas Leung, George Toderici, Alexander Toshev, and Weilong Yang. Pose embeddings: A deep architecture for learning to match human poses. In *arXiv*, 2015. 1, 2

[12] Ketul Shah, Anshul Shah, Chun Pong Lau, Celso M. de Melo, and Rama Chellapp. Multi-view action recognition using contrastive learning. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3370–3380, 2023. 2

[13] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. NTU RGB+D: A large scale dataset for 3d human activity analysis. *CoRR*, abs/1604.02808, 2016. 5

[14] Jennifer J. Sun, Jiaping Zhao, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, and Ting Liu. View-invariant probabilistic embedding for human pose, 2020. 1

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. 2

[16] Cailing Wang and Jingjing Yan. A comprehensive survey of rgb-based and skeleton-based human action recognition. *IEEE Access*, 11:53880–53898, 2023. 2

[17] Yang Wang and Greg Mori. Hidden part models for human action recognition: Probabilistic versus max margin. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(7):1310–1323, 2010. 1

[18] Haoran Wei and Nasser Kehtarnavaz. Simultaneous utilization of inertial and video sensing for action detection and recognition in continuous action streams. *IEEE Sensors Journal*, 20(11):6055–6063, 2020. 2

[19] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. *CoRR*, abs/1801.07455, 2018. 2

[20] Qipeng Zhang, Tian Wang, Mengyi Zhang, Kexin Liu, Peng Shi, and Hichem Snoussi. Spatial-temporal transformer for skeleton-based action recognition. In *2021 China Automation Congress (CAC)*, pages 7029–7034, 2021. 2

[21] Huanyu Zhou, Qingjie Liu, and Yunhong Wang. Learning discriminative representations for skeleton based action recognition, 2023. 3