POSETRANSFORMER: ENHANCING ACTION RECOGNITION VIA RGB-SKELETON FUSION WITH TRANSFORMER

Anonymous submission

ABSTRACT

The domain of human action recognition, pivotal in various applications, has witnessed significant advancements with deep neural networks and transformers. This paper introduces PoseTransformer, a novel method for action recognition, particularly relevant in the context of construction sites. PoseTransformer integrates a unique RGB-skeleton channel fusion model. Compared to traditional graph sequence methods, it employs fusion and skeleton adaptive coefficient to offer a more nuanced representation of the human skeleton, blending the self-attention mechanism with a transformer. Based on multi-person keypoint estimation, this model extracts the human skeleton across spatial windows, integrating it into original data. Subsequently, the integrated data is segmented into patches to create fixed-size tokens and processed through a transformer encoder. Comparative analysis demonstrates that PoseTransformer outperforms existing models with an accuracy of 0.955 on the KTH dataset and exhibits remarkable 10% improvement in accuracy when tested on diverse background settings, exemplified by the construction datasite, compared to the original RGB transformer models.

Index Terms— Action Recognition, Transformer, Pose Estimation, Deep Neural Networks, Skeleton Fusion

1. INTRODUCTION

1.1. Background

The digital age has witnessed an exponential increase in video content uploaded to the Internet, driven by advancements in video capturing technology, such as smartphones and cameras. This proliferation has sparked significant interest in understanding and analyzing video content, particularly the actions depicted within. Unlike static images, video information is dynamic and mutable, making the extraction of actionable insights particularly valuable. This is evident in its applications across automated surveillance, human-computer interaction, sports analytics, and real-time patient monitoring [1, 2]. Consequently, human action recognition in videos has emerged as a critical area of research, attracting substantial attention in recent years.

Previous studies in this field have explored various modalities for action feature representation, including single RGB sequences, optical flows, and human skeleton data. However, the transformative impact of the transformer architecture, initially renowned in natural language processing (NLP), has extended its efficacy to a broader spectrum of tasks, ranging from image classification to speech recognition. Drawing inspiration from this versatile architecture, our paper introduces a pure Transformer encoder derived architecture specifically tailored for action recognition in construction sites. This novel application effectively caters to scenarios with varying numbers of individuals, thereby advancing the field of action recognition.

1.2. Overview

This paper is structured as follows: Section 2 presents a comprehensive review of prior work in human action recognition, with a specific focus on construction site applications. Section 3 details our proposed method, which integrates a fusion coefficient and a skeleton adaptive coefficient with a Transformer-based self-attention mechanism, offering a refined representation of the human skeleton beyond conventional graph sequences. Section 4 discusses the experimental results obtained from the KTH dataset and a construction site dataset, underscoring the superiority of our approach in comparison to existing methodologies. Finally, Section 5 concludes the paper and outlines prospective avenues for future research.

2. RELATED WORK

The field of human action recognition has seen widespread applications across diverse domains, from human-computer interaction [3] and film and television animation [4] to rehabilitation training [5] and robotics [6]. Deep learning models, particularly those capturing human keypoints, have been central to these advancements. Notable works in this area include Banerjee et al. [7], who leveraged the DenseNet 201 architecture in a Convolutional Neural Network for classification, demonstrating its effectiveness across three challenging datasets: PPMI, Stanford 40, and BU-101. Chen et al. [8] introduced 'Watch Once Only' (WOO), an end-to-end pipeline for video action detection. Song et al. [9] innovated with a temporal context detector capable of extracting long-term context information efficiently, showcasing notable performance on the UCF101-24 and J-HMDB datasets. Angelini

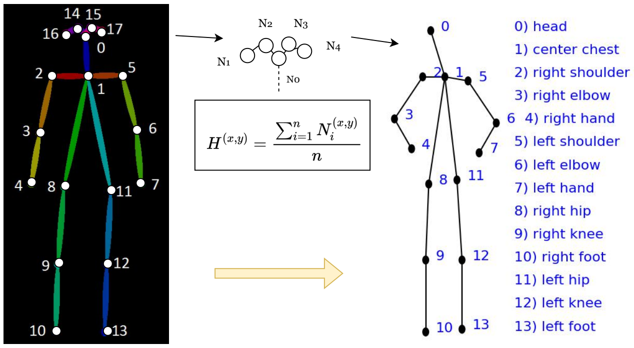

Fig. 1. The 14 keypoints utilized in our study. The lines represent the skeletal structure, while the points indicate the body parts of a person. Corresponding descriptions for each point are also provided.

et al. [10] proposed a novel 2D pose-based approach, which extracts low and high-level features from body poses, employing a combination of Long Short-Term Memory Neural Network and a 1D Convolutional Neural Network for pose-level human action recognition.

In more specialized domains, Ludl et al. [11] presented a modular simulation framework for training algorithms in various human-centered scenarios, achieving near-perfect performance with a recurrent neural network on real data using only simulated data based on motion capture data and 3D avatars. Kwon et al. [12] addressed the resource-intensive nature of frame-by-frame visual flow analysis with their trainable Motion Squeeze model, which efficiently transforms frame connections into motion features. Munro et al. [13] proposed a self-supervised alignment method to enhance motion recognition performance across diverse datasets and environments. Furthermore, Christoph et al. [14] achieved significant results in video classification and detection with their high spatial-temporal extensibility network.

In construction site applications, Fang et al. [15] utilized the high-precision Faster R-CNN network for far-field video analysis, focusing on helmet detection among workers. Kong et al. [16] predicted unsafe behavior using an improved Social-LSTM, proving the efficacy of computer vision in construction site safety. Additional research in this area has concentrated on worker safety, unsafe behaviours, surveillance, and risk reduction [17, 18, 19]

Collectively, these studies have delved deeply into aspects such as keypoints identification, human action recognition, construction safety identification, and other aspects in depth. However, the complex and variable nature of construction environments necessitates a more robust approach. Our work addresses this need by introducing an end-to-end action recognition framework for construction site workers, utilizing self-attention mechanisms and an innovative RGB-skeleton channel integration approach. This research provides an advancement in the field of intelligent applications in building construction.

3. PROPOSED APPROACH

This section details the architecture of our proposed approach. We first describe our method for extracting 14 keypoints from the 18 body keypoints typically recognized. The action transformer model, incorporating the fusion coefficient and skeleton adaptive coefficient, is then elaborated.

3.1. Human Key-points

Human keypoints recognition can be broadly categorized into two approaches: top-bottom and bottom-up [20, 21, 22]. The top-bottom method involves identifying the target detection box coordinates for each person in the input image, followed by pinpointing the human body keypoints. In contrast, the bottom-up approach first detects all keypoints in the input image, then calculates keypoints affinity for clustering, and finally assembles different keypoints into individual figures. The real-time multi-person key-point detection model [20] provides 18 body keypoints, 70 face keypoints, 42 hand keypoints, and 6 feet keypoints for each person. Our approach focuses on the 18 body keypoints for action recognition in construction sites. Specifically, we average five of these keypoints (left eye, left ear, right eye, right ear, and nose) to form a singular 'head' keypoint. The remaining 14 keypoints are then used in our analysis. Figure 1 illustrates the 14 human body keypoints detected by the real-time multi-person keypoints detection model.

3.2. Action Transformer Model

We introduce PoseTransformer, an innovative approach for RGB-skeleton fused action recognition. During preprocessing, we first extract a skeleton representation from an RGB image and identify human body areas based on this skeleton. The extracted skeleton information is then weight-combined with the input image. For images containing multiple individuals, each individual's RGB area with skeleton representation is segmented, reducing background impact on the analysis. Our approach uniquely applies a pure Transformer network with self-attention mechanisms to human action recognition, fusing RGB and skeleton information to enhance feature representation.

For input data, if a video frame is used, it is treated as an image. The process involves a fusion of RGB and skeleton data through specific fusion coefficients (f) and skeleton adaptive coefficients (γ), resulting in combined RGB-skeleton features. The transformer architecture in our approach is based on Vision Transformer [23]. An image is represented as $\mathbf{x} \in \mathbb{R}^{C \times W \times H}$, where C is the number of channels (3 for RGB images), and W and H are the width and height of the image, respectively (both set to 224 in this model). The image is divided into patches of size $\omega \times \omega$, each of which is mapped to a one-dimensional vector or 'token' (t_z). This process results in N tokens, where $N = (\frac{W}{\omega})^2$. The patches can be described as $\mathbf{x}_p \in \mathbb{R}^{N \times (C \cdot \omega^2)}$.

To maintain a consistent latent vector size (D=768) across all Transformer layers, we flatten the input patches and map them to D dimensions using linear projection. Each patch is represented as $\mathbf{P}_{\mathbf{x}}^i \in \mathbb{R}^{(C \cdot \omega^2) \times D_\omega}$.

The sequence of t_z is thus:

$$t_z = [\mathbf{P}_{\mathbf{x}}^1; \mathbf{P}_{\mathbf{x}}^2; \mathbf{P}_{\mathbf{x}}^3; ...; \mathbf{P}_{\mathbf{x}}^N] \in \mathbb{R}^{Pnum \times D_\omega} \tag{1}$$

In this study, we focus on five primary human actions relevant to construction sites: walking, working, standing, resting, and falling. The original image is processed to extract human skeletal features, and a skeleton diagram is generated based on the part index and connection rules shown in Figure 1. In this diagram, joints are represented by solid circles of a certain size, and bones by line segments of a certain thickness. The skeleton diagram includes detailed information about bone connections, positions, and pose features such as inclination.

The human body presents different relative sizes in different images for the original input. To adapt to varying human body sizes in different images, we introduce the skeleton adaptive coefficient γ, which adjusts bone thickness in relation to the image size and contrast ratio. γ is defined as:

$$\gamma = c \cdot \left[\sqrt{\frac{w_i \cdot h_i}{WH}} + \mu \cdot \Delta(I_{RGB} M^T) \right], \tag{2}$$

where w_i, h_i are the width and height of the external enclosing rectangle, c and μ represent basic pixel and contrast influence parameters, respectively. M is a grayscale projection vector, optimized to [0.81 0.12 0.07] for effective results. Δ denotes the difference between maximum and minimum values, and $I_{RGB} \in \mathbb{R}^{(w \cdot h) \times 3}$ is the pixel matrix of the original input. This adaptive mechanism ensures that joint size and bone line thickness are dynamically adjusted, reducing feature extraction errors due to size discrepancies.

The original RGB data is fused with the skeleton graph, leveraging both fusion coefficient f and skeleton adaptive coefficient γ. The fused tokens are derived using the token acquisition method outlined in formula (1). A special classification token (CLS-token), a vector of identical length to the tokens, is also incorporated.

As depicted in Figure 2, the fusion features of the combined RGB-skeleton are processed with the token sequence t_z. Each token $\mathbf{P}_{\mathbf{x}}^i$ is preceded by a learnable CLS-token $\mathbf{P}_0$. For each $\mathbf{P}_{\mathbf{x}}^i$, it is stacked using a positional embedding, which is expressed as $\mathbf{P}_{pos}^i \in \mathbb{R}^D$. Following this, the image sequence post-processing is output as the result of the final attention processing, as shown in formula (3):

$$\mathbf{P}^{(l)} = \begin{bmatrix} \mathbf{P}_0 \\ \mathbf{P}_{\mathbf{x}}^1 \\ \mathbf{P}_{\mathbf{x}}^2 \\ ... \\ \mathbf{P}_{\mathbf{x}}^n \end{bmatrix}^{(l)} + \begin{bmatrix} \mathbf{P}_{pos}^0 \\ \mathbf{P}_{pos}^1 \\ \mathbf{P}_{pos}^2 \\ ... \\ \mathbf{P}_{pos}^n \end{bmatrix}^{(l)} \in \mathbb{R}^{(N+1) \times D_\omega} \tag{3}$$

Here, the superscript l (ranging from 0 to L) on $\mathbf{P}^{(l)}$ denotes the token post-processing through the lth layer. The number of patches, denoted by n in Figure 2, is calculated based on the method described in formula (1). The Transformer Encoder comprises multiple stacked Encoder blocks, including Normalization blocks, Multi-Head Attention, and Dropout [24]. The MLP Head consists of the Linear layer, Tanh activation function layer, and Dropout layer.

The initial input $\mathbf{P}^{(0)}$ is fed into the Transformer encoder, and the Q, K, V values for each layer are calculated as follows:

$$\begin{aligned} Q_p^{(l,\, a)} &= W_Q^{(l,\, a)} LN(z_p^{(l-1)}), \\ K_p^{(l,\, a)} &= W_K^{(l,\, a)} LN(z_p^{(l-1)}), \\ V_p^{(l,\, a)} &= W_V^{(l,\, a)} LN(z_p^{(l-1)}), \end{aligned} \tag{4}$$

Here, $Q_p^{(l,a)}$ represents the query value of the pixel block after head a through layer l, and similarly for K and V. $W_Q^{(l,a)}, W_K^{(l,a)}, W_V^{(l,a)}$ denote the linear transformations post head a through layer l, and LN stands for layer Normalization. The model calculates the attention of each patch, with the attention mechanism in each layer calculated as follows:

$$\alpha_{(p)}^{(l,\, a)} = softmax(\frac{Q_p^{(l,\, a)}}{\sqrt{D_h}} \cdot {K^{(l,\, a)}}^T) \tag{5}$$

In this equation, D_h is the attention head size, and the activation function used is $softmax$. The attention vector for each head is:

$$S_p^{(l,\, a)} = \alpha_{(0)}^{(l,\, a)} V_{(0)}^{(l,\, a)} + \sum_{p'=1}^{p_{num}} \alpha_{p'}^{(l,\, a)} V_{p'}^{(l,\, z)}. \tag{6}$$

After concatenating the attention vectors of multiple heads, the output for each layer is obtained through a residual connection:

$$z'^{(l)}_p = W^o \begin{bmatrix} S_p^{(l,\, 1)} \\ \vdots \\ S_p^{(l,\, A)} \end{bmatrix} + z_p^{(l\, -\, 1)}, \quad W^o \in \mathbb{R}^{h \cdot D_h \times D} \tag{7}$$

$$z_p^{(l)} = MLP(LN(z'^{(l)}_p)) + z'^{(l)}_p, \tag{8}$$

Here, MLP represents the multi-layer perceptron linear classification head.

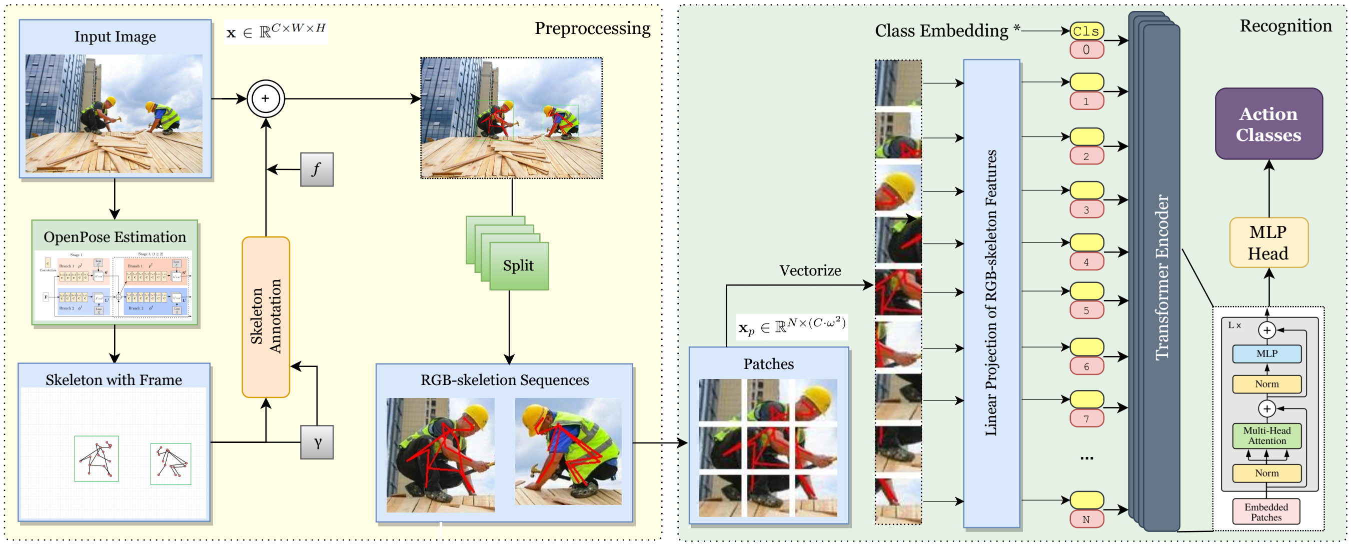

Fig. 2. Schematic overview of the PoseTransformer model for human action recognition. The process involves obtaining RGB-skeleton sequences based on the number of individuals in the input image. Each sequence is then processed for action recognition, allowing us to categorize the actions of each individual.

4. EXPERIMENTS AND ANALYSIS

4.1. Datasets

This section elaborates on the datasets used to evaluate the performance of our proposed PoseTransformer approach: the publicly available KTH human action dataset and a specially compiled construction site dataset. We describe the KTH dataset and outline the preprocessing method for the construction site dataset images.

4.1.1. KTH Dataset

The KTH dataset [25], renowned for its wide use in action recognition research, comprises 2391 video samples. It covers six distinct action categories: boxing, handclapping, handwaving, jogging, running, and walking. These actions are performed by 25 individuals across four different settings. Each action category includes around 100 videos, lasting between 20 to 30 seconds, summing up to approximately 600 video sequences in total. Selected frames from the six action categories are displayed in Figure 3.

4.1.2. Construction Site Dataset (ConSi-5)

Due to the scarcity of publicly available datasets focused on construction sites, we compiled a dataset comprising images from construction sites and web sources with actions similar to those observed in construction environments. This dataset consists of 7349 labeled images, with 5147 designated for training and 2202 for testing. The actions categorized and labeled include walking, working, standing, resting, and falling.

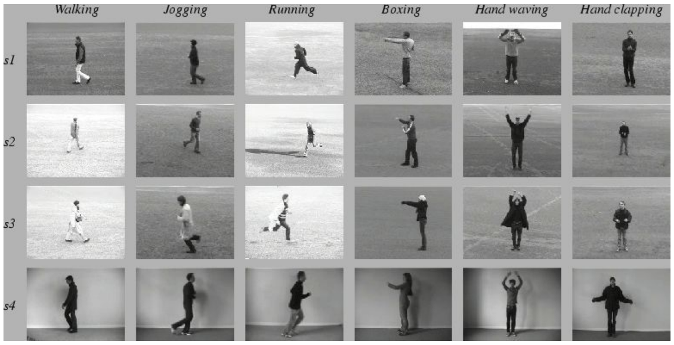

Fig. 3. Sample frames from the KTH dataset, showcasing the variety of actions. [25]

To illustrate the dataset's diversity, we selected nine representative images from each action category, resized them to 40×40 pixels, and arranged them as shown in the first row of Figure 4. It is worth noting that at construction sites, scenes with a single person are rare. Therefore, the images from construction sites include both multiple persons and a single person performing different action types. For analysis, images with multiple individuals were segmented to focus on single-person actions. The second row of Figure 4 visualizes skeletal structures for the five actions using coordinate information of 14 key nodes, including the trunk, limbs, and head. The skeletal structure varies across images due to differences in camera distance, necessitating the use of the skeleton adaptive coefficient γ introduced in Section 3 for effective feature extraction.

The third row of Figure 4 displays images post-fusion us-

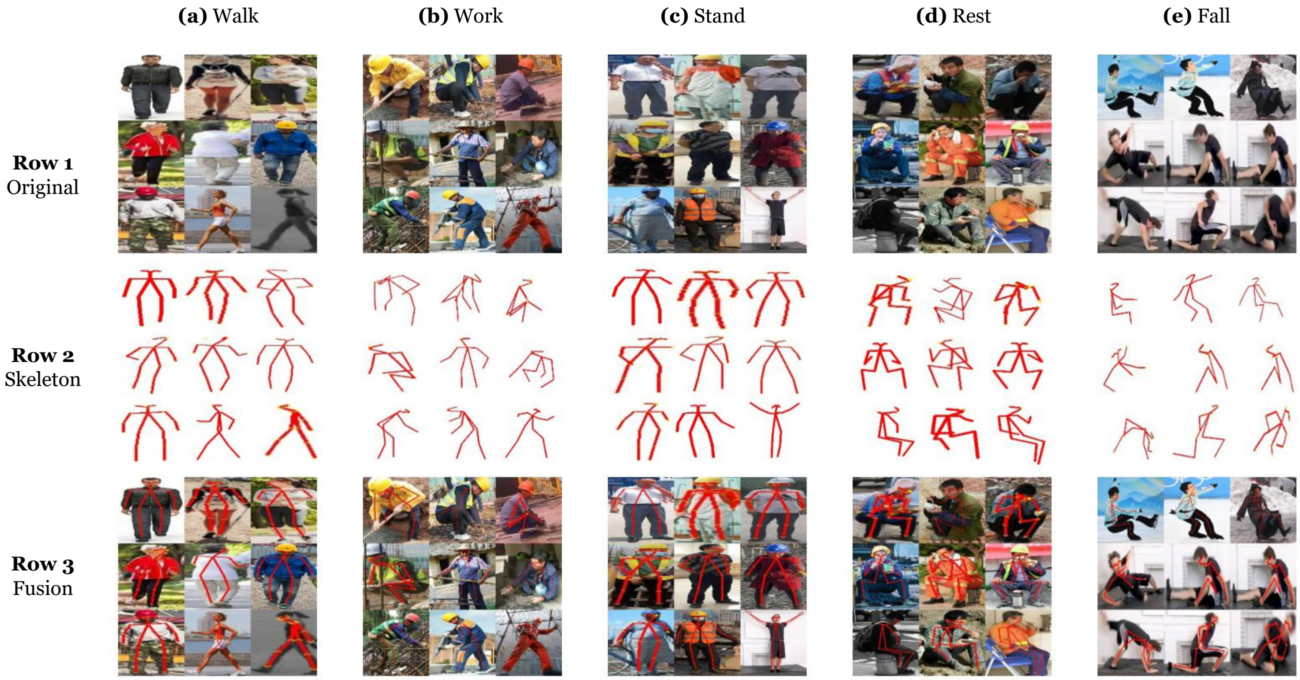

Fig. 4. Representative images of five action types from the construction site dataset. The first row shows original images, the second row depicts the corresponding skeleton structures, and the third row presents the fused images with skeletons. Columns (a) to (e) correspond to the actions of walking, working, standing, resting, and falling, respectively.

ing the fusion coefficient f and the skeleton adaptive coefficient γ. These fused images are segmented based on the minimum bounding rectangle shape to isolate individual figures. Similar to the original images, we randomly selected nine fused images, resized them to 40×40 pixels, and compiled them as shown in the third row of Figure 4.

4.2. Experimental Settings

This section presents the evaluation of our proposed PoseTransformer approach on both the KTH dataset and a construction site dataset. The experiments were conducted on Ubuntu 20.04, utilizing PyTorch version 1.9 and CUDA version 10.1. In both datasets, the data was split into 75% for training and 25% for testing.

The experimental setup for each trial included 100 training epochs, with a batch size of 16 and an initial learning rate of 0.001. We employed a cosine decay strategy for learning rate adjustment and opted for stochastic gradient descent with momentum for optimization. Considering the limited size of many human action recognition datasets, particularly in construction site scenarios, data augmentation is crucial. We adopted random horizontal flipping for image augmentation, carefully considering the impact of skeleton orientation on the recognition of actions, particularly the 'falling' action in our dataset.

4.3. Results and Discussion

4.3.1. Results on KTH Dataset

For the KTH dataset, three comparative experiment sets were conducted. The first set employed conventional deep learning algorithms, such as VGGNet or LeNet, for RGB image classification (denoted as 'RGB' in Figures 5 and 6). The second set used a Transformer model with the self-attention for RGB images (denoted as 'Transfm. (RGB)'). The final set incorporated our proposed RGB-skeleton fusion approach with self-attention Transformer (denoted as 'Transfm. (RGB-skeleton)'). The training epoch for each experiment was set to 100. The results are presented in Figure 5.

As evident in Figure 5, the PoseTransformer (RGB-skeleton) approach yielded the highest accuracy (0.955), surpassing the RGB-only method (0.940 accuracy) and the LeNet-based approach without skeleton fusion (0.891 accuracy). The training loss trends corroborate these findings, indicating the effectiveness of RGB-skeleton fusion and the potential of Transformer models in feature recognition.

Based on these observations, it becomes evident that the fusion of RGB information with skeleton data can effectively capture the action features present in images. Furthermore, this fusion process facilitates feature extraction at the pixel level. Notably, the comparison between RGB and Transformer (RGB) models, as depicted in Figure 5, reveals a significant difference of approximately 5%. This discrepancy

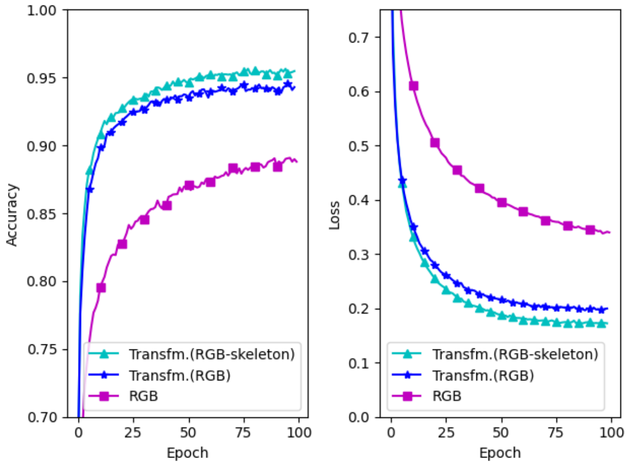

Fig. 5. Performance comparison on the KTH dataset.

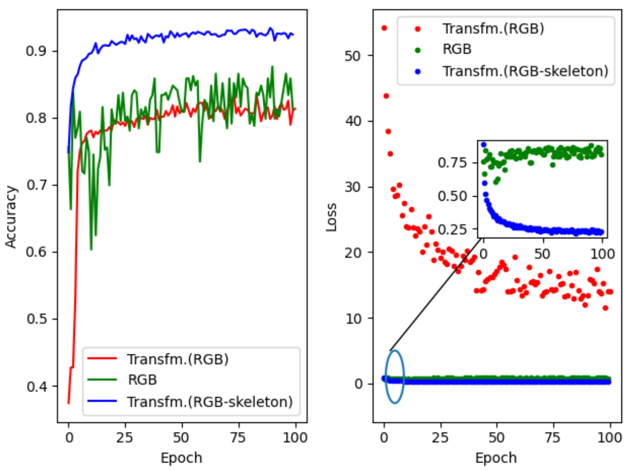

Fig. 6. Performance evaluation on ConSi-5 dataset.

highlights the substantial potential of the transformer model with self-attention in recognizing and extracting features.

4.3.2. *Comparative Analysis*

A comparative analysis with existing methods on the KTH dataset further validates our approach. The PoseTransformer achieved an average accuracy of 0.955, outperforming recent methods such as 3D convolutional networks, Deep brief networks, 2-D wavelet transforms, Differential recurrent networks, and others. Table 2 summarizes the recognition accuracy of our proposed approach and the recent methods from state-of-the-art using other deep learning models.

Table 1. Performance Comparison with Other Approaches

Method	Accuracy
Proposed Approach	0.955
3D Convolutional [26]	0.903
Deep Brief Network [27]	0.948
2-D Wavelet Transform [28]	0.943
Differential Recurrent Network [29]	0.936
Recurrent Neural Network [30]	0.943
Optical Flow [31]	0.950

This comparative analysis demonstrates the superior performance of the PoseTransformer, highlighting its effectiveness in accurately recognizing human actions in videos.

4.3.3. *Results on Construction Site Dataset*

The primary objective of this study was to achieve effective action recognition in construction sites. In this context, we applied our proposed PoseTransformer approach to the construction site dataset outlined in section 4.1.2. The experimental outcomes are illustrated in Figure 6.

As demonstrated in Figure 6, the test accuracy and training loss curves over epochs for three different models show distinct trends. Notably, both the Transformer (RGB) and Transformer (RGB-skeleton) models converge around the 30-epoch mark, with the Transformer (RGB-skeleton) model achieving an accuracy 10% higher than the Transformer (RGB) counterpart. This highlights the effectiveness of incorporating the Transformer model in enhancing action recognition accuracy. However, the RGB model's accuracy fluctuates around 0.8, indicating less stable convergence and inferior feature extraction capabilities compared to Transformer-based models.

The Transformer (RGB) model starts with a relatively low accuracy of 0.3, while both the Transformer (RGB-Skeleton) and the pure RGB methods begin at an accuracy level of around 0.75. The lower initial performance of the Transformer (RGB) model is attributed to its inability to integrate RGB and skeleton information, limiting its feature extraction capability in the complex and color-rich backgrounds typical of construction sites. After being split into patches, the Transformer (RGB) model struggles to effectively extract features. Conversely, methods that either use pure RGB or fuse RGB with skeleton information exhibit more robust feature extraction capabilities.

Our findings indicate that while the Transformer model enhances the overall recognition ability, the RGB-skeleton fusion notably improves feature extraction. The combination of these approaches results in superior performance across various aspects, particularly in the challenging environment of construction sites, where feature extraction is more difficult.

Comparing Figures 6 and 5, it is evident that the accuracy on the construction site dataset is lower than that on the KTH dataset. This discrepancy is likely due to the smaller number of images in the construction site dataset and the diversity in background settings, which could negatively impact accuracy.

5. FUTURE WORK

Our future research will be directed towards refining the PoseTransformer model to achieve more precise and nuanced recognition of various action categories. Our primary focus will be on applications in complex and dynamically changing environments, such as construction sites.

Our experimental results have shown that the fusion of features provided by pose estimation and the classification capabilities of the Transformer model can effectively complement each other, mitigating their respective shortcomings. However, we recognize that there is room for exploration in terms of alternative fusion methods.

In the forthcoming research, we plan to delve deeper into the fusion of multimodal data, exploring different techniques to optimize the integration process. Simultaneously, we aim to fine-tune the Transformer architecture to enhance the model's adaptability and robustness. Our overarching objective is to develop a model that not only excels in controlled environments, as demonstrated in the KTH dataset, but also thrives in the unpredictable and varied settings encountered in real-world scenarios.

6. CONCLUSION

This study introduced PoseTransformer, a transformative approach to human action recognition using a pure Transformer network augmented with self-attention mechanisms. Our novel method combines RGB and skeleton information, effectively utilizing fusion and skeleton adaptive coefficients to significantly enhance feature extraction. The PoseTransformer demonstrated exceptional performance on the KTH dataset, achieving a classification accuracy of 0.955, thereby outperforming current state-of-the-art methods. In the more challenging context of construction site datasets, where the complexity of the environment and diverse backgrounds present significant challenges, our approach still marked a 10% improvement in accuracy over traditional models. Notably, the integration of RGB with skeleton information proved crucial in these environments, compensating for the Transformer (RGB)'s lower performance due to its inability to incorporate skeleton data.

7. REFERENCES

[1] Maide Bucolo, Arturo Buscarino, Carlo Famoso, Luigi Fortuna, and Mattia Frasca, "Control of imperfect dynamical systems," *Nonlinear Dynamics*, vol. 98, pp. 2989–2999, 2019.

[2] Hehe Fan and Mohan Kankanhalli, "Motion= video-content: Towards unsupervised learning of motion representation from videos," in *ACM Multimedia Asia*, pp. 1–7. 2021.

[3] Imen Jegham, Anouar Ben Khalifa, Ihsen Alouani, and Mohamed Ali Mahjoub, "Safe driving: Driver action recognition using surf keypoints," in *2018 30th International Conference on Microelectronics (ICM)*. IEEE, 2018, pp. 60–63.

[4] Wafa Lejmi, Anouar Ben Khalifa, and Mohamed Ali Mahjoub, "Fusion strategies for recognition of violence actions," in *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*. IEEE, 2017, pp. 178–183.

[5] G Anitha and S Baghavathi Priya, "Posture based health monitoring and unusual behavior recognition system for elderly using dynamic bayesian network," *Cluster Computing*, vol. 22, pp. 13583–13590, 2019.

[6] Paul Duckworth, Muhannad Alomari, Yiannis Gatsoulis, David C Hogg, and Anthony G Cohn, "Unsupervised activity recognition using latent semantic analysis on a mobile robot," in *IOS Press Proceedings*, 2016, number 285, pp. 1062–1070.

[7] Avinandan Banerjee, Sayantan Roy, Rohit Kundu, Pawan Kumar Singh, Vikrant Bhateja, and Ram Sarkar, "An ensemble approach for still image-based human action recognition," *Neural Computing and Applications*, vol. 34, no. 21, pp. 19269–19282, 2022.

[8] Shoufa Chen, Peize Sun, Enze Xie, Chongjian Ge, Jiannan Wu, Lan Ma, Jiajun Shen, and Ping Luo, "Watch only once: An end-to-end video action detection framework," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8178–8187.

[9] Lin Song, Shiwei Zhang, Gang Yu, and Hongbin Sun, "Tacnet: Transition-aware context network for spatio-temporal action detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11987–11995.

[10] Federico Angelini, Zeyu Fu, Yang Long, Ling Shao, and Syed Mohsen Naqvi, "2d pose-based real-time human action recognition with occlusion-handling," *IEEE Transactions on Multimedia*, vol. 22, no. 6, pp. 1433–1446, 2019.

[11] Dennis Ludl, Thomas Gulde, and Cristóbal Curio, "Enhancing data-driven algorithms for human pose estimation and action recognition through simulation," *IEEE transactions on intelligent transportation systems*, vol. 21, no. 9, pp. 3990–3999, 2020.

[12] Heeseung Kwon, Manjin Kim, Suha Kwak, and Minsu Cho, "Motionsqueeze: Neural motion feature learning for video understanding," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*. Springer, 2020, pp. 345–362.

[13] Jonathan Munro and Dima Damen, "Multi-modal domain adaptation for fine-grained action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 122–132.

[14] Christoph Feichtenhofer, "X3d: Expanding architectures for efficient video recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 203–213.

[15] Qi Fang, Heng Li, Xiaochun Luo, Lieyun Ding, Hanbin Luo, Timothy M Rose, and Wangpeng An, "Detecting non-hardhat-use by a deep learning method from far-field surveillance videos," *Automation in construction*, vol. 85, pp. 1–9, 2018.

[16] Ting Kong, Weili Fang, Peter ED Love, Hanbin Luo, Shuangjie Xu, and Heng Li, "Computer vision and long short-term memory: Learning to predict unsafe behaviour in construction," *Advanced Engineering Informatics*, vol. 50, pp. 101400, 2021.

[17] Seyed Meysam Khoshnava, Raheleh Rostami, Rosli Mohamad Zin, Arunodaya Raj Mishra, Pratibha Rani, Abbas Mardani, and Melfi Alrasheedi, "Assessing the impact of construction industry stakeholders on workers' unsafe behaviours using extended decision making approach," *Automation in Construction*, vol. 118, pp. 103162, 2020.

[18] Bogyeong Lee and Hyunsoo Kim, "Measuring effects of safety-reminding interventions against risk habituation," *Safety science*, vol. 154, pp. 105857, 2022.

[19] Bader Al Mawli, Mubarak Al Alawi, Ashraf Elazouni, and Abdullah Al-Mamun, "Construction smes safety challenges in water sector in oman," *Safety science*, vol. 136, pp. 105156, 2021.

[20] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7291–7299.

[21] Pichao Wang, Wanqing Li, Philip Ogunbona, Jun Wan, and Sergio Escalera, "Rgb-d-based human motion recognition with deep learning: A survey," *Computer vision and image understanding*, vol. 171, pp. 118–139, 2018.

[22] Tanmay Nath, Alexander Mathis, An Chi Chen, Amir Patel, Matthias Bethge, and Mackenzie Weygandt Mathis, "Using deeplabcut for 3d markerless pose estimation across species and behaviors," *Nature protocols*, vol. 14, no. 7, pp. 2152–2176, 2019.

[23] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[25] Christian Schuldt, Ivan Laptev, and Barbara Caputo, "Recognizing human actions: a local svm approach," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.* IEEE, 2004, vol. 3, pp. 32–36.

[26] Majd Latah, "Human action recognition using support vector machines and 3d convolutional neural networks," *Int. J. Adv. Intell. Informatics*, vol. 3, no. 1, pp. 47–55, 2017.

[27] Mehrez Abdellaoui and Ali Douik, "Human action recognition in video sequences using deep belief networks.," *Traitement du Signal*, vol. 37, no. 1, 2020.

[28] Khawlah Hussein Ali and Tianjiang Wang, "Learning features for action recognition and identity with deep belief networks," in *2014 International Conference on Audio, Language and Image Processing*. IEEE, 2014, pp. 129–132.

[29] Vivek Veeriah, Naifan Zhuang, and Guo-Jun Qi, "Differential recurrent neural networks for action recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4041–4049.

[30] Moez Baccouche, Franck Mamalet, Christian Wolf, Christophe Garcia, and Atilla Baskurt, "Sequential deep learning for human action recognition," in *Human Behavior Understanding: Second International Workshop, HBU 2011, Amsterdam, The Netherlands, November 16, 2011. Proceedings 2*. Springer, 2011, pp. 29–39.

[31] Ning Zhang, Zeyuan Hu, Sukhwan Lee, and Eungjoo Lee, "Human action recognition based on global silhouette and local optical flow," in *International Symposium on Mechanical Engineering and Material Science (ISMEMS 2017)*. Atlantis Press, 2017, pp. 1–5.