2023 IEEE Symposium Series on Computational Intelligence (SSCI)
Mexico City, Mexico. December 5-8, 2023

VIDEO-BASED SKELETON DATA ANALYSIS FOR ADHD DETECTION

Yichun Li[1], *Rajesh Nair*[2], *Syed Mohsen Naqvi*[1]

[1] Intelligent Sensing and Communications Research Group, Newcastle University, UK
[2] Cumbria, Northumberland, Tyne and Wear (CNTW), NHS Foundation Trust, UK

ABSTRACT

Attention Deficit Hyperactivity Disorder (ADHD) is a common neurobehavioral disorder in humans worldwide. While extensive research has focused on machine learning methods for ADHD detection and diagnosis. Most methods rely on high-cost equipment and trained staff for data collection, e.g., Magnetic Resonance Image (MRI) machine and Electroencephalography (EEG) patch. Therefore, low-cost sensors-based easy-to-process methods for ADHD detection by exploiting action and behavior symptoms are required. We present that skeleton-based action recognition has the potential to address the application due to the action-focused nature of ADHD. Hence, this work proposes a novel ADHD detection system with a privacy-mitigating skeleton-based action recognition framework by utilizing our new real multi-modal ADHD dataset. Compared to the conventional methods, the proposed method shows cost efficiency and significant performance improvement. This method also outperforms the conventional methods in accuracy and AUC on the real multi-modal dataset. Furthermore, our proposed method based on simple non-wearable sensors is widely applicable for ADHD screening.

***Index Terms*— ADHD diagnosis, skeleton, action-recognition, action classification**

1. INTRODUCTION

Attention Deficit Hyperactivity Disorder (ADHD) is a common neurobehavioral and neurodevelopmental disorder affecting 2-5% of school-age children worldwide, with a high rate of undiagnosed cases among adults [1]. Recently, machine learning detection and diagnosis methods based on Magnetic Resonance Imaging (MRI) [2] and Electroencephalography (EEG) [3] have achieved high accuracy of over 95% on related datasets [4], i.e., ADHD-200, but are limited by their expensive equipment and high operational costs [5]. Thus, there is a need for machine learning methods based on low-cost data categories, e.g., video and audio, to facilitate ADHD detection and primary diagnoses.

According to the Diagnostic and Statistical Manual of Mental Disorders-Fifth Edition (DSM-V), the supportive evidence shows that ADHD behavioral features such as fidgeting and restlessness in clinical notes are typically generalized rather than characterized by the repetitive and stereotyped movements [6]. Conventional clinical observation of ADHD behavioral symptoms is limited by the difficulty in accurately counting and extracting these characteristics [4, 7].

Action recognition methods have recently overcome manual data processing limitations by extracting action information, i.e., skeleton-joint, from raw videos. These methods are unaffected by privacy-related information and demonstrate remarkable robustness in dynamic environments and complex backgrounds[8]. This paper proposes a novel action recognition method based on the human skeleton-joint modality. Furthermore, it explores the potential of ADHD detection by identifying and analyzing raw video recordings. Our main contributions include: 1) to the best of our knowledge, the first multi-modal ADHD dataset based on real patients. Based on the characteristics of ADHD, a test focusing on ADHD actions and reaction ability is designed and implemented; 2) a novel ADHD detection system based on action recognition networks; 3) new classification criteria to provide detection results and analysis of ADHD behavioral characteristics. We verify the efficiency and feasibility of the system with ablation study results.

2. PROPOSED FRAMEWORK

2.1. Participants and Procedure

This study utilizes a real recorded multi-modal ADHD dataset consisting of 7 adult ADHD subjects diagnosed by medical consultants under DSM-V criteria and 10 neurotypical controls. The dataset includes 3 males and 4 females with ADHD and 9 males and 1 female in the control group. All subjects are provided by the CNTW-NHS Foundation Trust, while the control group volunteers are from Newcastle University.

The consecutive test focus on the ability of attention and responsiveness is provided for all participants. We prepare four continuous dialogue tasks: 1) a brief conversation between the subjects, controls, and the interviewer, approximately 10-20 minutes. The task consists of 21 questions selected from Diagnostic Interview for ADHD in Adults (DIVA), i.e., used by NHS medical consultants and professionals; 2) performing Cambridge Neuropsychological Test

Automated Battery (CANTAB) tasks, including Cambridge Gambling Task (CGT), Stop Signal Task (SST), Rapid Visual Information Processing (RVP), and Spatial Working Memory (SWM). This task takes about 40-50 minutes; 3) beep reaction task requires participants to respond to beeps of different lengths. This task takes 6 minutes; 4) watching videos, including a math video labelled 'boring' and a rally video labelled 'exciting'. This task takes 10 minutes.

Videos are recorded by 3 GoPro cameras which contain a front-faced Camera 1 to record facial information. Meanwhile, Cameras 2&3 record information on the left and right sides of the torso and limbs, respectively. The resolution of the three cameras is the same, i.e., 3840×2160 pixels.

2.2. Proposed ADHD Detection System

We propose an action-based analysis system, which can be used in ADHD detection with raw RGB videos and be a competitive approach to clinical, EEG, and FMRI-based approaches. As aforementioned, the proposed method is simple and efficient compared to conventional fMRI and EEG-based methods because the video signal is easy to obtain with the low equipment cost. The framework overview is shown in Fig. 1. Details of each proposed task will be covered in the following sections.

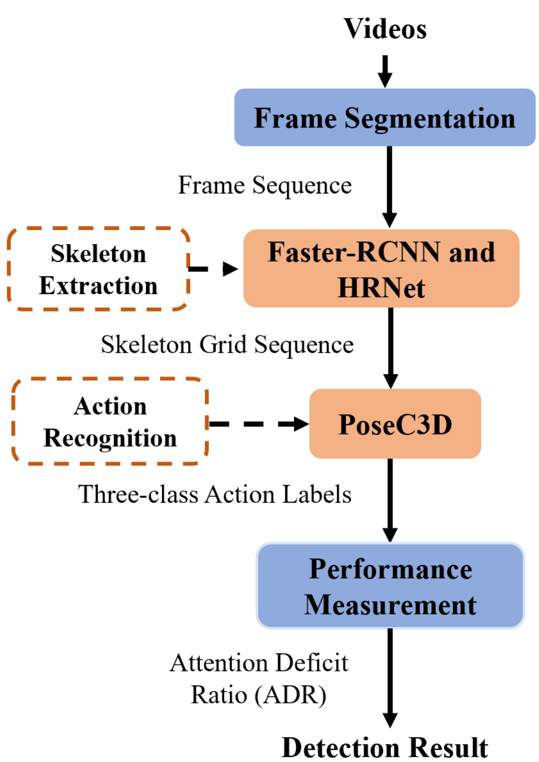

Fig. 1. Flow diagram of the proposed ADHD detection system. The dashed lines are the format of each task, while the solid lines point to the network and tasks of the system.

2.3. Skeleton Extraction and Action Recognition

In the frame segmentation task, the input video from our ADHD multi-modal dataset is decomposed into a frame sequence of 25 FPS. We use the *detector* and *estimator* to capture pose information in the frame sequence and record it as a human skeleton-joint grid sequence in the skeleton extraction task. Action-related information is extracted through these tasks without contextual nuisances, such as background variation and unrelated personnel interference [9].

In general, 2D poses are of better quality and higher accuracy than 3D poses [10], which is crucial for applications related to medical detection and diagnosis. In this work, a ResNet50-based Faster-RCNN network is used as the *detector* [11]; the pose *estimator* is a pre-trained HRNet because they achieve the state-of-the-art results on the commonly used human posture estimation dataset: MPII (top) and COCO (bottom) datasets [11]. As shown in Fig. 2, we use this combination of detector and estimator to capture standard benchmarks such as COCO-keypoints of subjects and controls in a sitting position[8]. The 17 joint points are detected and used in action and pose tasks. Skeleton-joint grid sequence information is stored in a series of coordinate triplets (x, y, c), where c is related to the number of joints, height, and weight at each frame. (x, y) is the corresponding coordinates of the c [8].

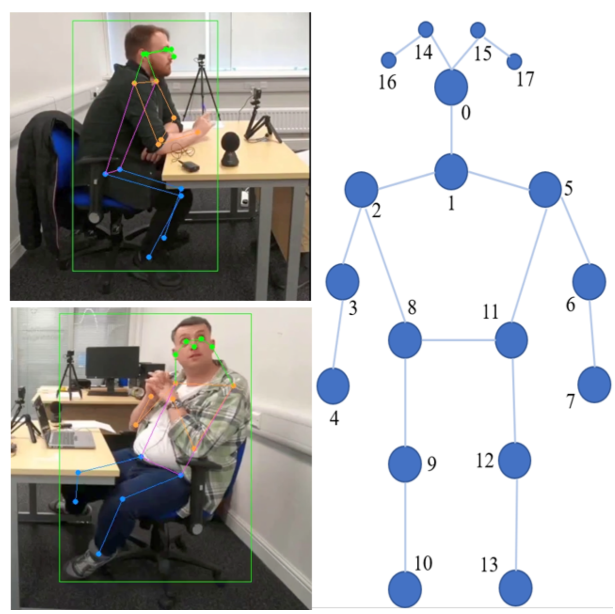

Fig. 2. Skeleton extraction results from left and right side cameras in our dataset and the COCO-Keypoints information.

2.4. Performance Measurement

To the ADHD typical symptoms, the actions of subjects and controls in our dataset mainly contain three categories: still position, small ranges of limb fidgets, and large rotations of torso movements. The existing action recognition evaluation criterion cannot be applied to ADHD-specific classification and evaluate its action frequency characteristics. We propose a novel Hyperactivity Score (HS) and a measurement named Attention Deficit Ratio (ADR) as the evaluation criterion for action classification of ADHD symptoms detection. They focus on the action change frequency of the subjects and controls during the test, which is also defined as the model's ability to focus on the movement or posture. The Hyperactivity Score (HS) is calculated as :

$$HS_n = \begin{cases} HS_{n-1} + 1, & \text{if } l_n = l_{n-1} \\ HS_{n-1} - 1, & \text{otherwise} \end{cases} \quad (1)$$

where n denotes the number of labels, HS_n denotes the score of n labels, l_n denotes the nth label in the label sequence. HS increases if the action is consistent in the continuous time frame. Otherwise, it is reduced.

According to the effect of video length on HS, we normalize the results by the ratio of HS to the n labels and denote it as ADR, which is calculated as:

$$ADR(\%) = \frac{HS_n \cdot 100}{n} \quad (2)$$

We use ADR_L and ADR_R, i.e., ADR measures of left and right viewpoints recording for two cameras, respectively. The final ADR is the average ADR_L and ADR_R.

The detection results R are obtained by binary classification of the ADR results of all participants using a determined threshold T. The diagnosis result is calculated as :

$$R = \begin{cases} \text{ADHD}, & \text{if } ADR < T \\ \text{Control}, & \text{otherwise} \end{cases} \quad (3)$$

The performance of the proposed ADHD detection system is also evaluated by the standard measurements, e.g., accuracy, sensitivity, precision, and the area under curve (AUC).

3. EXPERIMENTS

3.1. Dataset Preparation

We use our real multi-modal ADHD dataset for the proposed ADHD diagnosis system. Especially to recognize ADHD symptom-related actions, a three-classes-action ADHD dataset is used for training and testing in action recognition.

The ADHD detection dataset contains the left and right body information recorded by two side cameras. The whole dataset contains 34 videos. In the action recognition part, we divide the subjects' actions in the sitting state into three categories, i.e., still-position (Action 1), which contains 88 video clips, limb-fidgets (Action 2) with 110 clips, and torso movements (Action 3) with 101 clips. Each of the clips is between 10-15 seconds. The training, validation, and testing data split is 7/1/3, respectively.

The input frame is reduced from 3840×2160 to 1080×920 and down-sampled from 32 to 25 FPS to minimize the computation cost. 2D-Poses are captured and estimated by the top-down estimator from RGB inputs, as shown in Fig. 2. Actions are labeled per 50 frames in the training and detection steps.

3.2. Experiment Setup

We exploit a 3D-CNN structure (PoseC3D) as the main core network [8]. Different from commonly used GCN methods in skeleton-based action recognition, PoseC3D is a novel backbone that takes the 2D-Poses as the heatmap stacks of skeleton joints rather than graph coordinates. On the temporal dimension, the heatmap sequence of different time steps consists of a 3D-dimension heatmap volume. PoseC3D is more robust to the upstream pose estimation and temporal actions due to the 3D structure of heatmap [8]. Compared with grid-based GCN methods, the interoperability of PoseConv3D makes it easier to process human skeletons in the multi-modality and the multi-modal fusion, potentially used in ADHD detection and diagnosis. Meanwhile, the PoseC3D performs better on most existing action detection datasets, such as UCF101, NTURGB-D, FineGYM, etc. [12].

Different from the original implementation [8], the first convolution layer of our PoseC3D network is changed to $17 \times 25 \times 56 \times 56$ kernels with 1×1 stride to fit the size of our input data format. The training epochs for the action classification are 30, and the learning rate is 4×10^{-3}. All the experiments are run on a workstation with four Nvidia GTX 1080Ti GPUs and 16 GB of RAM.

3.3. Time-Action based Diagnosis Results and Comparisons

According to DSM-V, some symptoms of hyperactivity-impulsivity are observable in ADHD adults, such as difficulty in sitting still, fidgeting legs, tapping with a pen, etc., and those actions are not characterized by repetitive stereotypical movements [6]. However, it is hard to manually record irregular, high-frequency, and small-range actions during the traditional diagnostic process. Through our system, the skeleton-based poses and actions of each participant are fully captured and visualized. Fig. 3 shows the action recognition results timeline bar chart from a randomly selected subject and control.

Through Fig. 3, it can be easily observed that the action change frequency for the ADHD subject is significantly higher than the control. We further provide the ADR performance of 7 subjects and 10 controls as shown in Table 1.

From Table 1, the average ADR_{Avg} for 7 subjects and 10 controls are 71.7% and 79.8%, respectively. The average ADR_{Avg} of all 17 participants is 76.5%. Therefore, 76.5% is adapted as the threshold for ADHD detection.

Table 1. The Attention Deficit Ratio (ADR) comparisons for the overall subjects and controls. And 'S', 'C', 'F', 'M' indicate subject, controls, female, and male, respectively. Each result is the average of 5 experiments.

Samples	S2 (M) C1 (M)	S6 (M) C3 (M)	S9 (F) C4 (M)	S10 (F) C5 (M)	S12 (F) C7 (M)	S13 (M) C8 (M)	S14 (F) C11 (M)	C15 (M)	C16 (F)	C17 (M)
ADR$_{Avg}$(%)	48.9 80.8	75.7 79.9	74.6 83.9	71.9 81.7	72.6 67.8	72.1 84.0	76.1 81.3	86.6	65.6	86.5

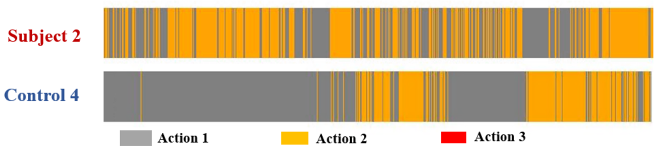

Fig. 3. Action change timeline chart of a randomly selected subject and control. Action 1 is still-position, Action 2 is limb-fidgets, and Action 3 is torso movements.

In the next experiment, we compare the proposed method with the other two commonly used skeleton-based action recognition networks (ST-GCN, MS-G3D) in our action recognition task. To ensure the fairness of the experiment, we use the same configurations, i.e., estimator, detector, learning rate, and evaluation matrix, for ST-GCN and MS-G3D as the proposed PoseC3D. Based on this basic framework, we adapt two popular 3D structure networks,i.e., C3D and R3D, to the action recognition task. Different from skeleton-based methods, these two networks both use the raw RGB frame sequence as input. The training epochs for the action classification are 80, and the learning rate is empirically set to 1×10^{-9}.

We further calculate the precision, sensitivity, accuracy, and AUC of four comparison networks: ST-GCN, MS-G3D, C3D, and R3D [12, 13, 14, 15], and our proposed PoseC3D framework in Table 2.

Table 2. ADHD detection system performance with different neural networks.

	Precision(%)	F1(%)	Accuracy(%)	AUC
R3D [15]	58.8	74.0	58.8	0.50
C3D [14]	85.7	70.6	70.6	0.71
ST-GCN [12]	100.0	75.0	76.4	0.70
MS-G3D [13]	85.7	70.6	70.6	0.72
PoseC3D	**100.0**	**88.9**	**88.2**	**0.83**

From Table 2, the proposed PoseC3D is significantly higher than the C3D, R3D, ST-GCN, and MS-G3D in precision, accuracy, F1 Score, and AUC. The Posec3d takes advantage of combining the skeleton grid and heatmap, which leads to improved performance on action recognition tasks, indicates a clear differentiation between ADHD subjects and controls in diagnosis outcomes, as well as improved detection accuracy.

3.4. Ablation Study

In the ablation study experiment, the original ADHD-3 dataset is shattered and labelled as ADHD subjects and controls. The C3D-1 is a diagnostic discriminant network trained on this binary ADHD classification dataset. Apart from the binary ADHD classification dataset, PoseC3D adds the skeleton information extraction task. The PoseC3D-2 and C3D-2 are action recognition networks with and without skeleton extraction task trained on the three-class action dataset, as mentioned in Section 3.1, respectively. It is highlighted that the action recognition task and the ADR task are closely related and cannot be separated. The AUC results are shown in Table 3:

Table 3. Ablation study results with AUC.

Task	Network	AUC
-	C3D-1	0.50
Skeleton	PoseC3D-1	0.59
Action+ADR	C3D-2	0.71
Skeleton + Action + ADR	**PoseC3D-2**	**0.83**

According to the results of the ablation study, firstly, the action recognition module plays an important role in the overall diagnostic system, which significantly improves diagnostic accuracy by extracting and classifying action features. Secondly, the skeleton extraction task improves the detection accuracy on the basis of the action recognition task by its robustness for the impact of environmental interference.

4. CONCLUSION

This paper proposed an ADHD detection system based on a skeleton-joints modality action recognition framework. A novel measure named ADR was proposed to evaluate the attention deficit performance of the action recognition results. The experimental results demonstrated that our system outperforms state-of-the-art methods regarding precision, accuracy, and AUC with high efficiency. Our systems are cost-effective and easily integrated into clinical practice. In future work, we plan to expand the dataset to cover a real-world patient distribution and record more multi-modal data such as EEG and fMRI for fusion and evaluation of related results.

Furthermore, we will focus on the effectiveness of deep learning models, particularly those based on graph convolutional networks and spatial-temporal architectures, to achieve superior results in action recognition tasks, thereby enabling the development of more efficient diagnostic systems for various applications.

5. ACKNOWLEDGEMENT

The authors would like to thank Yi Li for his help with proofreading.

6. REFERENCES

[1] C. Nash, R. Nair, and S. M. Naqvi, "Machine learning and ADHD mental health detection-a short survey," in *International Conference on Information Fusion (FUSION), 2022*.

[2] Y. Tang, J. Sun, C. Wang, Y. Zhong, A. Jiang, G. Liu, and X. Liu, "Adhd classification using auto-encoding neural network and binary hypothesis testing," *Artificial Intelligence in Medicine*, vol. 123, pp. 102209–102230, 2022.

[3] Y. Tang, X. Li, Y. Chen, Y. Zhong, A. Jiang, and X. Liu, "High-accuracy classification of attention deficit hyperactivity disorder with L2,1-Norm linear discriminant analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020*.

[4] N. A. Khan, S. A. Waheeb, A. Riaz, and X. Shang, "A novel knowledge distillation-based feature selection for the classification of ADHD," *Biomolecules, 2021*, vol. 11, no. 8, pp. 1093, 2021.

[5] H. Loh, C. Ooi, P .D. Barua, E .E. Palmer, F. Molinari, and U. Acharya, "Automated detection of ADHD: current trends and future perspective," *Computers in Biology and Medicine*, vol. 146, pp. 1–18, 2022.

[6] Fifth Edition et al., "Diagnostic and statistical manual of mental disorders," *Am Psychiatric Assoc*, vol. 21, no. 21, pp. 591–643, 2013.

[7] L. Dubreuil-Vall, G. Ruffini, and J. A. Camprodon, "Deep learning convolutional neural networks discriminate adult ADHD from healthy individuals on the basis of event-related spectral EEG," *Frontiers in Neuroscience*, vol. 14, pp. 251, 2020.

[8] H. Duan, Y. Zhao, K. Chen, D. Lin, and B Dai, "Revisiting skeleton-based action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022*.

[9] A. Sabater, L. Santos, J. Santos-Victor, A. Bernardino, L. Montesano, and A.C Murillo, "One-shot action recognition in challenging therapy scenarios," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021*.

[10] J. Liu, N. Akhtar, and A Mian, "Skepxels: Spatio-temporal image representation of human skeleton joints for action recognition.," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019*.

[11] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019*.

[12] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H Lu, "Skeleton-based action recognition with shift graph convolutional network," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020*.

[13] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020*.

[14] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *IEEE International Conference on Computer Vision (ICCV), 2015*.

[15] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018*.