ML-2SN: A Hybrid Two-Stream System for Sitting Posture Detection

Kehang Jia
khjia810@gmail.com
Cheng Cheng[†]
cc_412@outlook.com
Gaorui Zhang
dawangaoruizhang@gmail.com
Yixuan Yang
yxyang0708@gmail.com
Guangwei Huang
huangguangwei2809@outlook.com
Penghuan Wang
penghuanwang5@gmail.com

School of Future Science and Engineering
Soochow University
Jiangsu, CN

Abstract

Abnormal sitting postures often lead to neck, shoulder, back, and lumbar disorders and are becoming more prevalent across all age groups. Therefore, it is crucial to investigate intelligent monitoring technologies that can accurately recognize sitting postures in real-time. However, most of the previous studies in this field have focused on a single spatial or temporal domain, and these methods do not yield full information. In this paper, we propose a system called ML-2SN. In that system, we use LSTM to capture the latent temporal features in a set of consecutive skeletal key points as the temporal domain features of the model. We use MobileNetV3 to learn the connections among skeletal key points in the black background image of the skeleton as the spatial features of the model. To enhance the model's attention to action changes, we design the Mapping Block to attenuate the influence of spatial features on the model. Additionally, the ResLSTM Block is designed to enhance the effect of temporal features on the model. During model training, we use a novel label smoothing method (Action Label Smoothing) to attenuate the effect of action boundaries on the model. The system improves accuracy by recognizing human joints and filtering out ambient noise, effectively reducing the model's inference time. The experimental results show that the average detection accuracy of the system is 0.8894, which is 0.0192 better than the best existing method.

1 Introduction

With the increase of mental labor in modern society, long hours of work at the desk have become more and more normal. Abnormal sitting posture often leads to neck, shoulder, back,

[†]Corresponding author.

and lumbar disorders [2] [3], and it is becoming more prevalent in all age groups. However, due to the difficulty of effectively detecting sitting posture and maintaining a healthy sitting posture for a long time for occupational groups, real-time sitting posture detection and correction are of great importance. In previous studies, there are two main directions. One type of methods [4] [15] [6] [25] [18] [12] [14] uses a variety of sensors, such as pressure sensors, acceleration sensors, etc., to obtain information about various parts of the human body, and then build features and use models to classify them. The problem with this type of method is their poor convenience, low accuracy, and the small variety of sitting postures that can be detected. The other type of methods [1] [9] is to use images collected by visual sensors for detection. These works only focus on the human body shelter problem and do not utilize valid temporal information.

Compared to the limitations of sensors, sitting recognition methods using computer vision offer greater convenience, diverse sitting detection, and real-time feedback. Currently, traditional computer vision-based pose recognition methods usually use a single network model. The lack of ability of a single network model to extract key feature information leads to low recognition accuracy. Besides, most of the methods are unable to deal with the occlusion problem, and cannot solve the problem of occluding the lower body. The existing sitting detection systems are deficient in distinguishing the boundaries between actions and cannot effectively deal with the effect of action boundaries on the model. Aiming at the current sitting recognition methods is slow, has low accuracy, is susceptible to environmental influence, and action boundaries confuse the model, this research proposes a system that can detect poor sitting posture in real time to reduce the resulting health problems.

In this paper, we propose a new system called Mobile-LSTM Two-Stream Networks(ML-2SN) as an attempt to address these issues. Specifically, the system utilizes videos as input. This allows the model to extract action information and improve the accuracy of the model. Skeletal key points are localized to the original frames as well as a black background skeleton image is drawn. By using the skeletal information instead of the whole image input, we can reduce the amount of model computation while avoiding the influence of background noise. The information from the black background image of the skeleton is extracted as the spatial feature by using the MobileNetV3 model. However, direct use of this spatial feature will result in too much spatial information being input into the model, making the model unable to utilize the action information well. Therefore, we design a Mapping Block to solve the above problems. At the same time, the LSTM model is used to extract the information from the skeletal key points of consecutive frames as the temporal feature. We also designed a ResLSTM Block module to enhance the model's access to temporal information. Through the fusion of temporal and spatial information, the network can understand the temporal variation and spatial layout of poses better, improving the accuracy and robustness of pose detection. Notably, we use a novel label smoothing method (Action Label Smoothing) to attenuate the effect of action boundaries on the model. The following is our contribution:

- Separately, the skeleton is used as an input to the spatial domain, and the keypoint motion as an input to the temporal domain.

- For the first time in the field of sitting detection, we mix MobIleNetV3 and LSTM models. The high accuracy is guaranteed while maintaining the real-time performance of the system.

- We design the Mapping Block and ResLSTM Block modules to enable the model to better utilize the action information.

- We propose Action Label Smoothing methods to attenuate the obfuscation of the model by action boundaries.

2 Related Work

Detection of Sitting Posture Sitting Posture Recognition aims to extract and analyze the posture of the human body from the input RGB image or video to determine whether the human body is in a healthy or poor sitting posture. An early study [1] built a simple convolutional network for classification after using OpenPose to extract skeletal key points. Using skeletal information as input is effective and avoids the effect of background noise. However, he only used a 19-layer CNN model, which is not able to fully extract the effective information. Recently, SPRNet [5] used Vision Transformer and introduced Convolutional Stems and External-Attention. His advantage is that the model can obtain a more comprehensive sample relationship. A recent work [22] proposes to extract shallow features by the VGG network, and add constant mapping by the ResNet residual network to extract multi-dimensional features of the image; finally, the fused feature information is input into the Vision Transformer network for sitting posture recognition. This method makes the localization of key feature regions more complete and accurate. But agreed that the dynamic changes in human posture are not taken into account. PatchesCA [23] is a novel and efficient patch channel attention module that plugs into MobileNetV3-large. It helps to reduce the parameters and increase the accuracy. However, several of the above methods do not take into account the fact that dynamic changes in human posture can also help recognize posture. Later, an abnormal sitting recognition method based on multi-scale spatio-temporal features of the skeletal map [11] solved this problem. It incorporates the spatial dimension, temporal dimension, and whole-body skeletal features of the human body in abnormal sitting postures. Although it effectively utilized the information contained in the video. But he couldn't solve the masking problem. In contrast, our model fully integrates temporal and spatial information and uses skeletal information as input to avoid the effect of background noise.

Human Pose Estimation Although this study focuses on sitting detection, inspiration can still be drawn from human pose recognition. An adaptive neural network [24] through the two neural network structures of VA-RNN and VA-CNN. And designed a dual-stream scheme, end-to-end recognition of skeleton data is realized. The effect of view changes is mitigated and superior performance is achieved. This study only uses CNN to extract spatial features which cannot capture the features better. One study [10] designed a novel skeleton transformer module that automatically rearranges and selects important skeleton joints. Raw skeleton coordinates as well as skeleton motion are fed directly into CNN for label prediction. This method effectively integrates both skeletal information and skeletal motion. There was a recent job [19] proposed a novel multimodal two-stream 3D network framework, which can exploit complementary multimodal information to improve the recognition performance, and construct two discriminative video representations under depth and pose data modalities. This framework can exploit complementary multimodal information to improve the recognition performance. However, the introduction of multiple data modalities and the two-stream 3D CNN structure increases the computational complexity of the model, leading to an increase in inference time. It can be concluded from the previous research that extracting effective spatial-temporal information is very important for video-based action recognition. Therefore, in this paper, the same is performed.

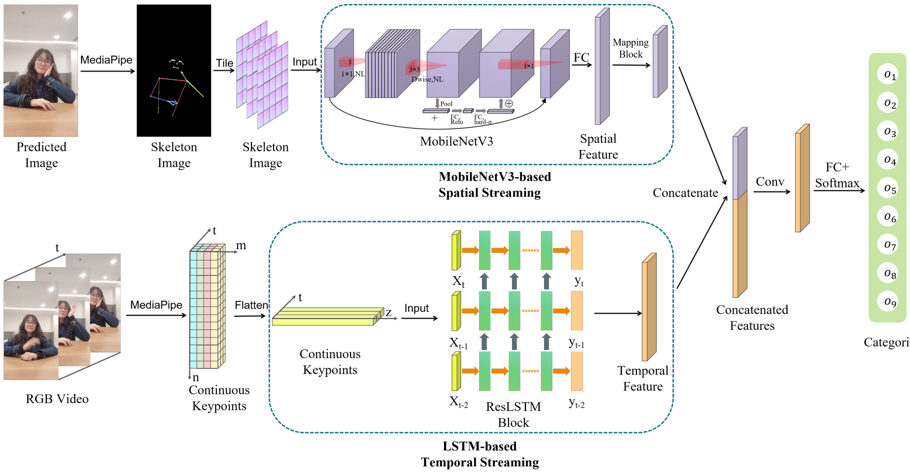

Figure 1: We propose ML-2SN for solving sitting detection. Our system mainly consists of (i) MobileNetv3-based spatial streaming, and (ii) LSTM-based temporal streaming. The system extracts a series of video frames. Different pose data is fed into the system after performing key point extraction. The spatial as well as temporal features obtained are concatenated and then convolutional feature extraction is performed. The model output is transformed into a probability distribution by FC + softmax for the task of sitting pose classification.

3 Method

In this study, we design a system of hybrid two-stream networks (Fig.1) for sitting recognition. Our approach involves several steps: firstly, raw video frames are passed through MediaPipe [13] for keypoint detection. and the skeleton is drawn on a black background image of equal size. Secondly, this processed set of data is fed into the model framework. In the spatial streaming, the black background image with the skeleton drawn is fed into MobileNetV3-Small to get features. MobileNetV3-Small [8] is an efficient, automatic search-optimized network model with excellent performance and adaptability. In temporal streaming, a motion vector is created by taking several consecutive frames of skeletal joints. This is fed into the LSTM network to get the temporal stream features. Finally, the features obtained through the two streams are concatenated together and fed into a classifier module consisting of a convolutional block and a fully connected layer. Meanwhile, we designed a unique label smoothing function during the training process to enhance the model generalization ability and improve the model calibration [16].

3.1 Skeletal Keypoints Detection

We use MediaPipe for keypoint detection on the captured video frames. It offers 32 key points including face, shoulders, knuckles, hips, knees, ankles, and so on. Since this work focuses on sitting detection, the areas below the hips are not usually captured by the camera. Therefore, we only take the key points above the hips to obtain sitting posture data. The key point i is described as $[x_i, y_i, z_i, v_i]$. where x, y,z describe the coordinates of the keypoint i and v describes the visibility of the point. After that, two operations are performed on the key points of consecutive frames. One is to combine the data of consecutive frames and spread them into a one-dimensional array. The other is to draw the key points on a black background image of equal size and connect some key points to draw them in the form of a human skeleton. Finally, resize this image to 720*1080.

3.2 MobileNetV3-based Spatial Streaming

MobileNetV3 is a lightweight deep neural network that improves the efficiency and performance of the model by introducing optimization features such as the h-swish activation function and the SE (Squeeze-and-Excitation) module, which utilizes the attentional mechanism to capture the key information of the image. In our experiments, we found that the feature vectors extracted from the MobileNetV3 backbone network contain a lot of redundant information, which may interfere with the model's recognition of important features. Therefore, we designed the Mapping Block to map the spatial feature size from the high latitude space to the low dimensional space. The structure of the Mapping Block is shown in Fig. 2(a), which contains the convolutional layer, the regularization layer, and the activation layer. The Mapping Block obtains the mapped low-dimensional feature representation by tandem operation of these three components. Mapping Block makes the model more focused on the changes between actions. In the sitting detection task, it is more important to focus on the dynamic changes in the sitting posture, while the focus on spatial details may lead to overfitting or sensitivity to noise. Specifically, a black background image [3, 720, 1280] drawn with a skeleton is input to the MobileNetV3 network. After the features are extracted by the MobileNetV3 backbone network, these feature vectors have the shape of [batch_size, 1000]. Where batch_size is a hyperparameter for the number of data samples we process in each training iteration. Next, these feature vectors are subjected to a series of operations such as convolution through a Mapping Block. Finally, they are converted into feature vectors of shape [batch_size, 256].

3.3 LSTM-based Temporal Streaming

LSTM introduces forgetting gates, input gates, and output gates to control the information flow and better capture and utilize long-term dependencies in time series data. Meanwhile, we designed a module (ResLSTM Block) for enhancing the temporal information. The structure is shown in Fig. 2(b). It combines the fully connected layer, the regularization layer, and the activation layer, and adds a residual connection between these layers. The fully-connected layer contains the process of upscaling and then downscaling. This design helps the model to better capture and utilize the temporal information and improves the model's utilization of action information. ResLSTM Block helps the model to be more stable and improves the model's generalization ability while ensuring that the model obtains more temporal information. Specifically, we input the skeletal key points [batch_size, 100, time_steps] of continuous frames into the convolution block to get the data of size [batch_size, 256, time_steps]. The feature vector of [batch_size, sequence_size, 512] is obtained after LSTM. Where sequence_size is a hyperparameter, which is the number of frames of the image we input to the model. After the fully connected layer in ResLSTM Block is upscaled to [batch_size, sequence_size, 1024] and downscaled to [batch_size, sequence_size, 512], and a residual connection is added between the first and last of ResLSTM Block. We used 2 ResLSTM Blocks in our system.

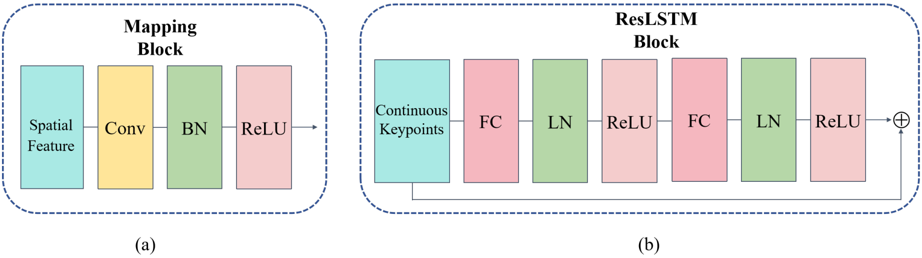

Figure 2: Figure (a) shows the details of the Mapping Block. It map the spatial feature size from the high latitude space to the low dimensional space. With the Mapping Block, the model can reduce redundant information. While figure (b) shows the details of the ResLSTM Block. In the ResLSTM Block, residual are added to improve stability. Throughout the process, the features go through a process of dimensionality upgrading and then dimensionality downgrading. So that the model gets more information about the change of action.

3.4 Action Label Smoothing

It is always known that the actions in video sequential frames are continuous and the action boundaries are not well defined. As a result, features near the action boundaries have high similarity compared to other classification tasks, which may confuse the model. So in this study, we design a label smoothing technique to regularize the model. In general, the data we feed into the LSTM model will not be overly long time-series data, i.e., they will not contain too many changes in actions. We make the t-frame images of the input model labeled as follows:

$$L = [l_1, l_2, ..., l_t] \tag{1}$$

where l_t is the label of the tth frame image, and t is a hyperparameter.

The label vector y_{ij} for conventional one-hot coding is known as follows:

$$y_{ij} = \begin{cases} 1, j = l_i \\ 0, j \neq l_i \end{cases} \tag{2}$$

where y_{ij} denotes the probability of the j label of the ith frame of the image. And $0 \leq i \leq t$, $0 \leq j \leq 8$.

Then our Action Label Smoothing operation is defined as follows. Let the function $c_n(L)$ be the count of the number of occurrences of label n in a given L list, denoted as follows:

$$c_n(L) = \sum \mathbb{I}(L_m = L_n) \tag{3}$$

Where $\mathbb{I}$ is the indicator function, if $L_m = L_n$ then take 1, the rest is 0. Where $0 \leq m \leq t$. Then L of each frame of the statistics for $count(L) = [c_1, c_2, ..., c_n]$.

Then the new image label is obtained as:

$$\hat{L} = \left[L_{i_1}, L_{i_2}, L_{i_3}\right] \tag{4}$$

Where $\hat{l}_1 = argmax\left(count\left(L\left[0 : \frac{t}{2}\right]\right)\right)$, $\hat{l}_2 = argmax\left(count\left(L\left(\frac{t}{2} : t-1\right]\right)\right)$, $\hat{l}_3 = l_t$. The function $argmax$ is the index to take the maximum value of $count(L)$. The above operations get the label with the highest number of occurrences of L in the frames $\left[0 : \frac{t}{2}\right]$ and $\left(\frac{t}{2} : t-1\right]$.

Then the label vector $\hat{y}_{ij}$ processed by our label leveling operation is:

$$\hat{y}_{ij} = 0.05 * \mathbb{I}\left(j = \hat{l}_1\right) + 0.05 * \mathbb{I}\left(j = \hat{l}_2\right) + 0.9 * \mathbb{I}\left(j = \hat{l}_3\right) \tag{5}$$

We used Action Label Smoothing as described above when training our model. This allows the model to more robustly handle feature similarity around fuzzy boundaries and action boundaries.

4 Experiments

4.1 Experimental Setup

Datasets Since there is no publicly available dataset in the field of sitting detection, we performed autonomous shooting. By using a cell phone camera, we filmed different scenarios. We shot eighteen videos with the help of eighteen experimenters. The number of processed video frames is 4981. In the labeling scheme, we have a total of nine categories, namely: lying down, head forward, right body lean, left body lean, right head tilt, left head tilt, left-hand support, right-hand support, and sitting normally. In head tilt, we calculate the absolute ROLL value of Euler's angle is greater than 15 degrees, then it is considered that the left and right tilted head amplitude is too large for abnormal sitting posture. Calculating the PICH value of Euler's angle is less than -30 degrees, then it is considered that the magnitude of head bowing is too large for forward head tilt. Body lean is defined by calculating the angle between the center point of the head and the center point of the shoulder, respectively, and the left shoulder point-right shoulder point continuum. The boundary angle for body lean is set to 10 degrees. The head support is easily defined as the hand in contact with the head. The head support is easily defined as the hand in contact with the head.

Implementation Details We trained and tested the models on an NVIDIA V100 graphics card, Ubuntu 22.04 system. All networks were implemented in PyTorch [17]. The model was trained for a total of 100 epochs and the batch size was 8. We use an efficient AdamW optimizer with a weight decay of 0.02. During training, the learning rate grows linearly to 6e-4 in the first 10 epochs, and then the learning rate is tuned using cosine annealing in the remaining epochs.

4.2 Results

Comparison of Sitting Recognition To measure the performance of our models on sitting detection, we report the Acc@1, Precision, and F1-score to demonstrate the ability of each model for the sitting recognition task. Several models were also used for comparison including MobileNetV3, Conv+GRU, Conv+LSTM, ResNet18, and ResNet34. The results are shown in Table 1. We observe that our model(ML-2SN) performs better than other common models in sitting detection. Specifically, the proposed model achieves 0.8894 on Acc@1, which is 0.0481, 0.0336, 0.0288, 0.0240, and 0.0192 better than the accuracy of MobileNetV3, Conv+GRU, Conv+LSTM, ResNet18, and ResNet34, respectively. These results demonstrate the superiority of the proposed model.

Contributions of Individual Components We verify the contributions of the proposed components in Table 2. We observe that i) When we reduce the spatial information obtained by the model or enhance the temporal information obtained by the model, the Acc@1 of the model improves by 0.0098 or 0.0188. Therefore, it can be said that the information

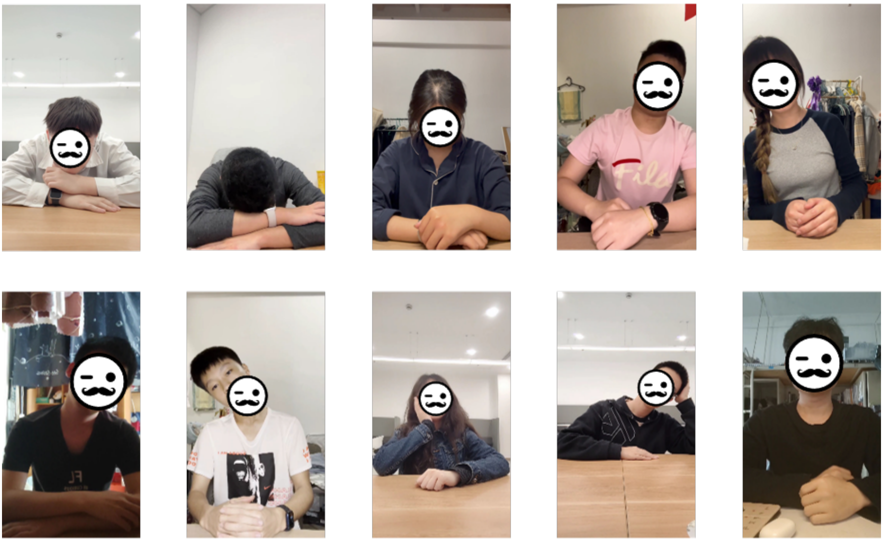

Figure 3: Presentation of datasets taken with a cell phone in different environments, with the help of different participants. From left to right, top to bottom. In order: lying down, lying down, head forward, right body lean, left body lean, right head tilt, left head tilt, left-hand support, right-hand support, and sitting normally.

Method	Acc@1	Precision	F1-score
MobileNetV3 [8]	0.8413	0.8745	0.8378
Conv+GRU [20]	0.8558	0.8683	0.8490
Conv+LSTM [21]	0.8606	0.8843	0.8573
ResNet18 [7]	0.8654	0.8782	0.8643
ResNet34 [7]	0.8702	0.8802	0.8694
ML-2SN	**0.8894**	**0.8912**	**0.8886**

Table 1: Comparison of the results of each model on our test dataset

on the movement changes is more useful for recognizing sitting postures. ii) The model that used the Action Label Smoothing method improved Acc@1 by 0.0182. Therefore, we believe that attenuating the effect of the action boundaries on the model facilitates the model's performance on the classification task.

5 Conclusion

In this paper, we research the detection of abnormal sitting postures and propose ML-2SN to solve this problem. ML-2SN mixes the spatial features obtained from the MobileNetV3 model and the temporal features obtained from the LSTM model to classify sitting postures. Meanwhile, we propose Mapping Block and ResLSTM Block to enhance the model to capture the action information. Action Label Smoothing is used to attenuate the effect of action boundaries on the model. The experiment results show that compared with the best existing methods, ML-2SN can significantly improve the recognition accuracy, and Acc@1 achieves

Method	Mapping Block	ResLSTM Block	Action Label Smoothing	Acc@1	Precision	F1-score
ML-2SN				0.8462	0.8565	0.8394
	✓			0.8560	0.8633	0.8520
		✓		0.8650	0.8901	0.8583
	✓	✓		0.8712	0.8735	0.8689
	✓	✓	✓	**0.8894**	**0.8912**	**0.8886**

Table 2: Ablation study on Mapping Block for spatial streaming, ResLSTM Block for temporal streaming, and Action Label Smoothing. We report Acc@1, Precision, and F1-score

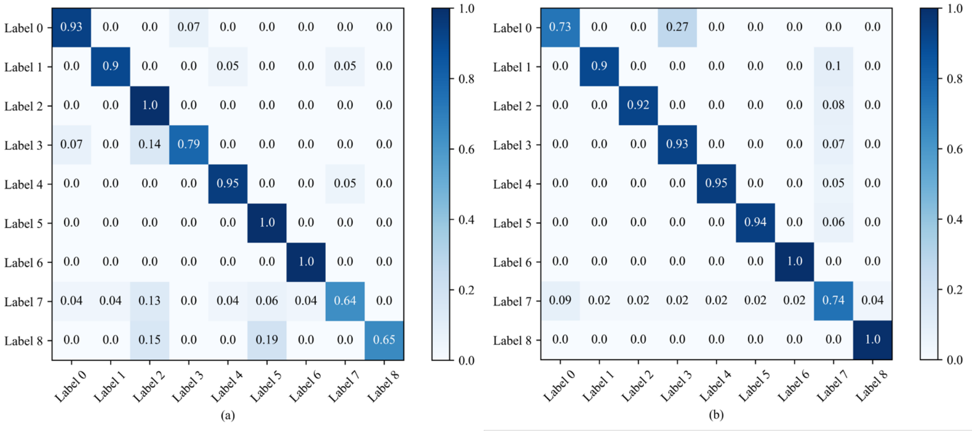

Figure 4: (a) Confusion matrix for MobileNetV3, (b) Confusion matrix for ML-2SN. Label 0 to Label 8 in the confusion matrix are, in order, lying down, right body lean, left body lean, head forward, right head tilt, left head tilt, right-hand support, sitting normally, left-hand support.

an improvement of 1.92%.

In order to have a more detailed look at the performance of the model on each category. We plotted the confusion matrices of MobileNetV3 and ML-2SN for comparison. As shown in Fig. 4. Although compared to MobileNetV3, our model though performs poorly on Label 0 recognition. However, the accuracies on Label 3, Label 7, and Label 8 are greatly improved.

6 Acknowledgements

This work was supported by the China Postdoctoral Science Foundation under grant no. 2023M742573.

References

[1] Kehan Chen. Sitting posture recognition based on openpose. In *IOP Conference Series: Materials Science and Engineering*, volume 677, page 032057. IOP Publishing, 2019.

[2] Stephen J Edmondston, Hon Yan Chan, Gorman Chi Wing Ngai, M Linda R Warren, Jonathan M Williams, Susan Glennon, and Kevin Netto. Postural neck pain: an investigation of habitual sitting posture, perception of 'good'posture and cervicothoracic kinaesthesia. *Manual therapy*, 12(4):363–371, 2007.

[3] Deborah Falla, Gwendolen Jull, Trevor Russell, Bill Vicenzino, and Paul Hodges. Effect of neck exercise on sitting posture in patients with chronic neck pain. *Physical therapy*, 87(4):408–417, 2007.

[4] Zhe Fan, Xing Hu, Wen-Ming Chen, Da-Wei Zhang, and Xin Ma. A deep learning based 2-dimensional hip pressure signals analysis method for sitting posture recognition. *Biomedical Signal Processing and Control*, 73:103432, 2022.

[5] Yi Fang, Shoudong Shi, Jingsen Fang, and Wenting Yin. Sprnet: sitting posture recognition using improved vision transformer. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6. IEEE, 2022.

[6] Emmanouil Fragkiadakis, Kalliopi V Dalakleidi, and Konstantina S Nikita. Design and development of a sitting posture recognition system. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3364–3367. IEEE, 2019.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[8] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019.

[9] Audrius Kulikajevas, Rytis Maskeliunas, and Robertas Damaševičius. Detection of sitting posture using hierarchical image composition and deep learning. *PeerJ computer science*, 7:e442, 2021.

[10] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Skeleton-based action recognition with convolutional neural networks. In *2017 IEEE international conference on multimedia & expo workshops (ICMEW)*, pages 597–600. IEEE, 2017.

[11] Linhan Li, Guanci Yang, Yang Li, Dongying Zhu, and Ling He. Abnormal sitting posture recognition based on multi-scale spatiotemporal features of skeleton graph. *Engineering Applications of Artificial Intelligence*, 123:106374, 2023.

[12] Wenjun Liu, Yunfei Guo, Jun Yang, Yun Hu, and Dapeng Wei. Sitting posture recognition based on human body pressure and cnn. In *AIP Conference Proceedings*, volume 2073. AIP Publishing, 2019.

[13] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, et al. Mediapipe: A framework for perceiving and processing reality. In *Third workshop on computer vision for AR/VR at IEEE computer vision and pattern recognition (CVPR)*, volume 2019, 2019.

[14] Slavomir Matuska, Martin Paralic, and Robert Hudec. A smart system for sitting posture detection based on force sensors and mobile application. *Mobile Information Systems*, 2020:1–13, 2020.

[15] Isaac Morales-Nolasco, Sandra Arias-Guzman, and Laura Garay-Jiménez. A method for complex posture recognition during long-term sitting using neural networks and pressure mapping systems. *Biomedical Signal Processing and Control*, 95:106306, 2024.

[16] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019.

[17] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[18] Ming-Chih Tsai, Edward T-H Chu, and Chia-Rong Lee. An automated sitting posture recognition system utilizing pressure sensors. *Sensors*, 23(13):5894, 2023.

[19] Hanbo Wu, Xin Ma, and Yibin Li. Spatiotemporal multimodal learning with 3d cnns for video action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1250–1261, 2021.

[20] Santosh Kumar Yadav, Achleshwar Luthra, Kamlesh Tiwari, Hari Mohan Pandey, and Shaik Ali Akbar. Arfdnet: An efficient activity recognition & fall detection system using latent feature pooling. *Knowledge-Based Systems*, 239:107948, 2022.

[21] Santosh Kumar Yadav, Kamlesh Tiwari, Hari Mohan Pandey, and Shaik Ali Akbar. Skeleton-based human activity recognition using convlstm and guided feature learning. *Soft Computing*, 26(2):877–890, 2022.

[22] Tao Yang, Qing Tao, Bin Wu, and Zirui Zhao. Research on sitting posture recognition based on deep fusion neural network. In *2023 4th International Conference on Computer Engineering and Application (ICCEA)*, pages 639–645. IEEE, 2023.

[23] Yongfang Ye, Shoudong Shi, Tianxiang Zhao, Kedi Qiu, and Ting Lan. Patches channel attention for human sitting posture recognition. In *International Conference on Artificial Neural Networks*, pages 358–370. Springer, 2023.

[24] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. View adaptive neural networks for high performance skeleton-based human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1963–1978, 2019.

[25] Liang Zhao, Jingyu Yan, and Aiguo Wang. A comparative study on real-time sitting posture monitoring systems using pressure sensors. *Journal of Electrical Engineering*, 74(6):474–484, 2023.