Component attention network for multimodal dance improvisation recognition

Jia Fu
RISE Research Institutes of Sweden
KTH Royal Institute of Technology
Stockholm, Sweden

Jiarui Tan
KTH Royal Institute of Technology
Stockholm, Sweden

Wenjie Yin
KTH Royal Institute of Technology
Stockholm, Sweden

Sepideh Pashami
RISE Research Institutes of Sweden
Stockholm, Sweden

Mårten Björkman
KTH Royal Institute of Technology
Stockholm, Sweden

ABSTRACT

Dance improvisation is an active research topic in the arts. Motion analysis of improvised dance can be challenging due to its unique dynamics. Data-driven dance motion analysis, including recognition and generation, is often limited to skeletal data. However, data of other modalities, such as audio, can be recorded and benefit downstream tasks. This paper explores the application and performance of multimodal fusion methods for human motion recognition in the context of dance improvisation. We propose an attention-based model, component attention network (CANet), for multimodal fusion on three levels: 1) feature fusion with CANet, 2) model fusion with CANet and graph convolutional network (GCN), and 3) late fusion with a voting strategy. We conduct thorough experiments to analyze the impact of each modality in different fusion methods and distinguish critical temporal or component features. We show that our proposed model outperforms the two baseline methods, demonstrating its potential for analyzing improvisation in dance.

CCS CONCEPTS

• **Computing methodologies → Activity recognition and understanding**; • **Human-centered computing**;

KEYWORDS

Dance Recognition; Multimodal Fusion; Attention Network

1 INTRODUCTION

Dance improvisation is an integral part of contemporary dance, allowing dancers to create spontaneous movements in response to external stimuli or internal impulses. Automated systems for recognizing dance improvisation can aid in the analysis of a dancer's originality, skill, and individuality. Such systems can also provide dance educators with a valuable tool for understanding the cognitive and physiological mechanisms underlying choreography and assessing the effectiveness of improvisation training programs. However, despite its importance in the dance world, improvisation has received relatively little attention in the field of computer vision and machine learning.

Most existing research on dance recognition focuses on the discrimination of dance types [11, 20, 27] or the classification of movements in a specific type of dance [2, 9, 16]. There is a lack of research highlighting the identification of different expressive qualities in improvisational dance. Although frameworks that perform well on general human action prediction tasks [4, 26, 28] can also be employed in improvisational dance recognition, they are limited in unimodal skeletal data. Besides skeleton dynamics, data of other modalities, e.g., inertial measurement unit (IMU) signals and respiratory cadence, can change idiosyncratically with dance moves. In order to investigate how to make better use of multimodal information for improvisational dance classification, we carry out this work and make the following contributions:

Firstly, we propose multimodal fusion methods in three levels: 1) feature fusion by a component attention network (CANet) adapted from BodyAttentionNet [24]; 2) model fusion by fusing a graph convolutional network (GCN) [25] with CANet; 3) and late fusion by a simple voting. Our proposed fusion strategies exceed the two State-of-the-Art frameworks in multimodal human motion prediction. Furthermore, we analyze the temporal/component attention scores and visualize them with heat maps, which leads to a quantifying of creative expression. The source code is available for verification. [1]

2 RELATED WORK

Multimodal fusion is among the most critical topics in multimodal learning. It aims to aggregate information from multiple modalities to infer discrete labels for classification tasks, such as audio-visual speech recognition [8], or continuous values for regression tasks, such as emotion prediction [1]. Three paradigms are characterized based on the stages when fusion is conducted: feature-level fusion, decision-level fusion, and model-level fusion. Feature-level fusion (a.k.a. early fusion) concatenates feature vectors right after they are extracted from various single modalities. [3] successfully achieves

[1] https://github.com/JasonFu1998/ComponentAttentionNetwork

a feature-level fusion of facial expressions and speech. However, feature-level fusion can be hindered by the high dimensionality of the feature space [29]. Decision-level fusion (a.k.a. late fusion) gives the final decision by incorporating the predictions inferred from different modalities in a voting process. It can be well applied to some multimodal inference tasks, such as affective computing [21], by ignoring low correlation modalities. Model-level fusion integrates intermediate representations of different modalities, which balances the benefits of the two approaches mentioned above and demonstrates effectiveness in [18].

Dance recognition. Dance motion data typically involves both temporal and spatial information. Classic approaches, such as long short-term memory (LSTM) [10], gated recurrent unit (GRU) [5], and GCN, are common solutions for handling such data. Some studies have attempted to address multimodal data of human movements. Fusion-GCN [6] incorporates other modalities, e.g., RGB data and IMU signals, into the GCN that represents the human skeleton. Gimme Signals [17] provides another novel approach for capturing motion features from multimodal data: signals of different modalities but the same length can be plotted in the same image. Afterward, such images can be used to train convolutional neural networks (CNNs). In particular, there has been extensive research exploring the applications of machine learning (ML) and deep learning (DL) in the classification of dance genres or dance figures. [20] compares the performance of DL and traditional ML methods on dance classification with Kinect sensor data. In [16], a neural network is trained to classify 3D pose data from wearable sensors into different ballroom dances. [27] designs a late fusion network for multimodal dance classification, involving four different modalities: RGB frames, optical flows, skeletal data, and audio.

3 METHODOLOGY

Problem Formulation. We focus on N-class classification on multimodal data. Each sample is a sequence of length T. At a time step t, each sample consists of C *components*, each of which is a feature vector $f_t^{(c)}$. Such vectors of different components may belong to the same or different modalities and therefore are not necessarily of the same size. Our goal is to train a neural network ϕ that serves as a mapping from input features $\{f_{1:T}^{(1)}, f_{1:T}^{(2)}, \ldots, f_{1:T}^{(C)}\}$ to probabilities of categories $\boldsymbol{p} = [p_1, p_2, \ldots, p_N]$. The predicted category is given by $n = \arg\max_i p_i$. In our scenario, a binary classification task for dance movement data is addressed, where $N = 2$.

CANet. We first present CANet, as depicted in Fig. 1, which can be seen as an upgraded version of BodyAttentionNet [24]. It targets multimodal data rather than just body parts. In addition, although CANet is designed for feature fusion, it can be further used for model fusion. To begin with, an input feature matrix is constructed by aligning components by frame and concatenating them. Next, CANet extracts features through a two-stage attention mechanism. Temporal attention modules extract information separately for different components. A subsequent component attention module performs further information extraction at the global level.

We first analyze each branch of CANet and ignore batch operations. For example, in the branch for component c, LSTM outputs a matrix $H_{1:T}^{(c)} \in \mathbb{R}^{T \times K}$, where K is the output dimension. The parameters of LSTM are shared by all branches. Temporal attention
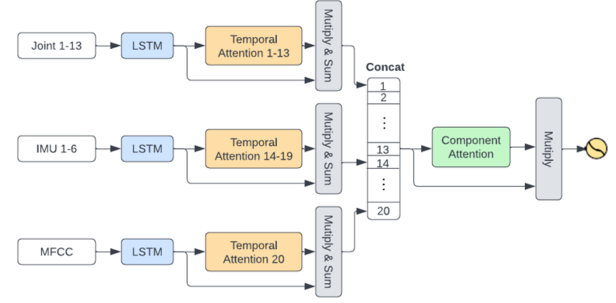
Figure 1: The Network Architecture of CANet. Different input features are fed into their respective branches, each containing a shared LSTM and a temporal attention module. Extracted information from different branches is concatenated and then fused by a component attention module.

scores $a_{1:T}^{(c)} \in \mathbb{R}^T$ are calculated and then normalized by the softmax function: $a_{1:T}^{(c)} = \text{Softmax}\left(H_{1:T}^{(c)} w^{(c)}\right)$, where $w^{(c)} \in \mathbb{R}^K$ are trainable parameters. The output of the temporal attention layer $\theta^{(c)} \in \mathbb{R}^K$ is a weighted sum of information from different time steps: $\theta^{(c)} = H_{1:T}^{(c)\top} a_{1:T}^{(c)}$.

The outputs from temporal attention layers are then concatenated into a matrix $\Theta = [\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(C)}] \in \mathbb{R}^{K \times C}$. By this point, the independent branches of the network are merged. Then the component attention module calculates a component attention map $B \in \mathbb{R}^{K \times C}$ as follows: $B = \text{Softmax}\left(\tanh\left(\Theta W_1 + b_1\right) W_2 + b_2\right)$, where W_i, b_i denote the weight and bias of a fully connected layer, the same as below. This attention map reflects the different importance of the components in different modalities and is subsequently used to weight the outputs. Weighted output matrix $O \in \mathbb{R}^{K \times C}$ embeds an entire input sample and is given by $O = B \odot \Theta$, where $\odot$ denotes element-wise multiplication. In the end, the predicted probabilities $\boldsymbol{p} \in \mathbb{R}^{1 \times N}$ are computed as follows: $\boldsymbol{p} = \text{Softmax}\left(\text{vec}(O)W_3 + b_3\right)$, where vec denotes flattening a matrix into a row vector.

GCN-CANet, as exhibited in Fig. 2, upgrades CANet by constructing an undirected graph of keypoints extracted from the human body. One special branch of GCN-CANet, termed GCN-LSTM in this paper, contains several cascading GCN layers and LSTM layers. It outputs intermediate features $H_{1:T}^{(g)} \in \mathbb{R}^{T \times K}$, the dimensionality of which remains unchanged compared to CANet. This branch captures the spatio-temporal dynamics of human skeletons, and the intermediate features can be fused at the component attention module without modifying the main architecture of CANet.

Decision voting. Given predictions from multiple models, the final decision is made by voting, which is regarded as late fusion.

4 EXPERIMENTS

Data preparation. We use the Unige-Maastricht Dance dataset [19, 22]. It contains 152 improvisational dance segments of two expressive qualities known as lightness or fragility. The primary criterion for distinguishing them is whether the fluidity is presented without interruption. Each segment is recorded in four modalities:
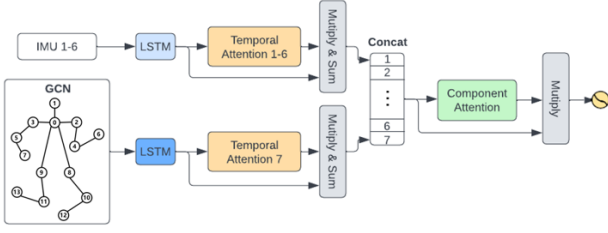
Figure 2: The Structure of GCN-CANet. Its modification over CANet is that the features of different joints are first integrated into one component using a GCN. In addition, the LSTM in this branch is not shared.

Figure 3: Body Joints Ordering and Extracted Skeletons.

Table 1: Results of Three Fusion Strategies. For late fusion, the best models (GCN-CANet, CANet, GRU) for corresponding modalities (J: Joints, I: IMU, M: MFCC) are used for voting.

Fusion Level	Modalities	Model	Accuracy	F1
-	J	CANet	74.26%	0.74
-	**J**	**GCN-CANet**	**76.79%**	**0.77**
-	**I**	**CANet**	**76.37%**	**0.76**
-	M	CANet	56.54%	0.54
-	**M**	**GRU**	**59.49%**	**0.59**
Feature	**J + I**	**CANet**	**82.28%**	**0.82**
Feature	J + M	CANet	68.78%	0.68
Feature	J + I + M	CANet	77.64%	0.78
Model	**J + I**	**GCN-CANet**	**83.12%**	**0.83**
Model	J + I + M	GCN-CANet	81.86%	0.82
Late	**J + I + M**	**Vote**	**78.90%**	**0.79**

Table 2: Comparison of Methods on Classification Accuracy.

Modalities	Ours	Fusion-GCN [6]	Gimme [17]
Joints	76.79%	76.79%	71.39%
Joints + IMU	**83.12%**	73.52%	77.67%
Joints + IMU + MFCC	**81.86%**	72.25%	-

dance video, IMU data, respiration audio, and electromyography (EMG) signals. They are synchronized at 50fps with an average length of 10.3s. 130 segments are randomly chosen to construct the training set, and the remaining 22 segments are used for testing. Based on [23], we run a sliding window to produce data instances from each segment, with a length of 3s and an overlap ratio of 80%.

The pose of a dancer in the video is estimated with AlphaPose [7, 13], which has been pretrained on COCO 2017 [14]. AlphaPose estimates the coordinates and visibility (x, y, v) of 17 keypoints for each detected person. We adopt the nose keypoint to mark the position of the dancer's head, with other facial keypoints deprecated for simplicity. Fig. 3 displays the body graph and the correspondence between ordered keypoints and body joints. The dancer wears IMUs on the left and right hands. Each IMU outputs (x, y, z) of the accelerometer, gyroscope, and magnetometer, respectively. As for audio, we compute the Mel-scale frequency cepstral coefficients (MFCC) [15] with 13 dimensions, which is the most common feature in speech recognition. Notice we abandon EMG in experiments due to some segments losing one of two EMG signals. In summary, there are 13 Joints components, 6 IMU components (2 accelerometers, 2 gyroscopes, and 2 magnetometers), and 1 MFCC component.

In our implementation of CANet and GCN-CANet, each LSTM has 3 layers with 8 hidden units, and each GCN consists of 3 layers with 16 hidden units. Models were trained with Adam optimizer [12]. We first conducted experiments on the individual modalities, followed by incorporating ancillary modalities, IMU and MFCC, into Joints modality through three fusion methods. Table 1 summarizes the performance comparison.

Feature fusion. We can conclude that: 1) For unimodal in CANet, IMU works best, and MFCC performs only marginally better than the random classification. 2) Joints and IMU are fused effectively. 3) Fusing MFCC will degrade the results compared to excluding it.

Model fusion. We arrive at the following conclusions from our experiments: 1) GCN-CANet beats plain CANet when they have

the same training data, i.e., model-level strategy is preferable to feature-level for the accordant fused components; 2) and linking discrete body joints into an integrated graph component by GCN mitigates the detrimental effect of bringing MFCC to the fusion.

Decision fusion. We choose the best-performing model for each modality respectively. Late fusion demands at least triple predictions for the same sample. Unlike before, MFCC is necessary and beneficial for late fusion. With all three modalities, late fusion performs better than feature fusion but worse than model fusion.

Comparison to baselines. In order to compare our approach with recent works, we conducted three experiments (unimodal, bimodal, and trimodal classification) to compare our methods with two baselines, Fusion-GCN [6] and Gimme Signals [17]. In Fusion-GCN, the GCN was implemented the same as ours for a fair comparison, but IMU and MFCC information were fused according to the original approach. The results presented in Table 2 illustrate that our approach achieves higher classification accuracy compared to the baselines in all experiments. The fusion scheme of Fusion-GCN resulted in lower accuracy with more modalities, demonstrating that it is not a good choice for this dataset. For Gimme, it is hard to encode data with different dimensions into a single image, especially when some modalities have high dimensions. Therefore, it is not suitable for trimodal data.

5 DISCUSSION

The deficiency of MFCC. From the video recordings, it can be observed that fragility movements generally have a narrower spatial range than lightness ones. Some typical fragility situations are performed at a fixed site on the stage throughout the whole segment.
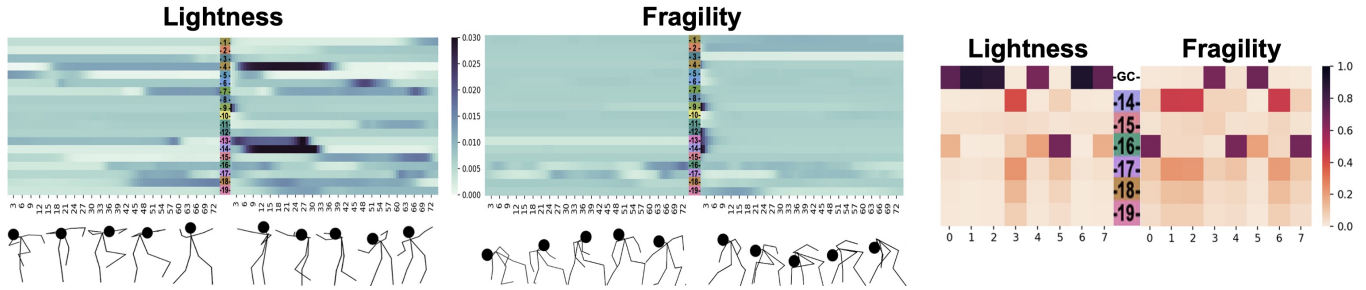
Figure 4: Heat Maps of Temporal (left) and Component (right) Attention Score for Selected Instances. The x-axis are the frame index and the numerical order of the entries in attention score vector respectively, and the y-axis are the index of components which has the same order and color of components listed in Fig. 5. GC represents graph convolution of the Joints 1 - 13.
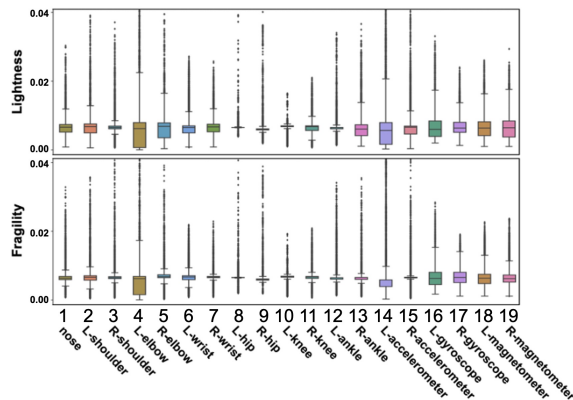
Figure 5: Temporal Attention Scores of All Test Instances. The x-axis lists a total of 19 Joints and IMU components.

Intuitively, smaller motion amplitude and velocity will lead to a smoother breath rhythm. MFCC is initially envisioned to reflect such tempo fluctuations. However, because the interval between two consecutive deep breaths is usually longer than the length of a sliding window, clear breaths captured by the microphone do not appear in all audio instances. Stability and complementarity of the information from each modality are prerequisites for successful fusion, which MFCC falls short of delivering.

Analysis of temporal attention scores. Based on the best results on CANet, when 13 Joints components and 6 IMU components are fused, we assess the temporal attention scores. Fig. 5 shows boxplots of attention statistics for all test instances categorized by the predicted motion labels. Boxplots associated with lightness are usually wider than those with fragility for the identical component, especially for left shoulder, right elbow, right ankle, and two accelerometers. This can be traced to two roots.

Dancers show more various choreography to convey lightness than fragility. This may rely on the diverse considerations made by dancers over which body parts should be highlighted as light and suspending. We choose two test lightness instances that are correctly classified with high confidence and then visualize their temporal attention scores in Fig. 4. For conciseness, only 75 successive frames are clipped from the selected instances. The corresponding dancer sketches below are exhibited around every 15 frames. An

attention switch from right elbow to left elbow is evident in the left instance as the dancer's left arm expands gradually from inside to outside of the trunk. More attention is paid on right ankle in the first 60 frames, meanwhile, the dancer's right foot moves from hovering to landing. In the first half of the right instance, the most decisive cues are discovered on left elbow and accelerometer of the left hand. The simultaneous dancer's left arm is stuck still, conforming to the characteristics of lightness. Moreover, the concern for right ankle is also reflected on the attention heat map, where the dancer elevates right foot to execute a horizontal side turn.

Conversely, the limited width of fragility boxplots illuminates the pattern that each joint preserves equally consistent relevance throughout the dance, emphasizing the collaboration of different body parts in maintaining balance on the verge of falling. Fig. 4 contains two representative fragility instances. Almost all components receive even attention throughout, except for gyroscope pairs.

Analysis on component attention scores. Fig. 4 includes two component attention heat maps in GCN-CANet without MFCC, matching two test instances with high prediction scores on their ground truth. The attention vector for each component rests with the hidden units of LSTM. Graph convolution is given the most attention among the involved components. Additionally, information is held in a complementary manner by different units of the same component. Take the graph convolution as an example, the nature of fragility is coupled with unit 3 and 5, whilst other units are more oriented toward lightness. As for IMU pairs, especially the gyroscope, more effect is carried by the left sensor than the right sensor for both motion qualities, confirming that the dancer's control of the left and the right arms is asymmetrical.

6 CONCLUSION

We employ three multimodal schemes, namely feature fusion, model fusion, and decision fusion, to classify the qualities of dance improvisation. We present CANet for feature fusion, which leverages the attention mechanism to devote more focus to the essential moments and components. By connecting all discrete joints into a human body topology, GCN-CANet for model fusion surpasses the naive CANet and also alleviates the negative effects of incorporating MFCC in the fusion. The experimental and visualized results demonstrate the effectiveness and scientificity of our models, which outperform the baselines in improvisational dance recognition.

REFERENCES

[1] Tadas Baltrušaitis, Ntombikayise Banda, and Peter Robinson. 2013. Dimensional affect recognition using continuous conditional random fields. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 1–8.

[2] Himadri Bhuyan, Jagadeesh Killi, Jatindra Kumar Dash, Partha Pratim Das, and Soumen Paul. 2022. Motion Recognition in Bharatanatyam Dance. *IEEE Access* 10 (2022), 67128–67139.

[3] Carlos Busso, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, and Shrikanth Narayanan. 2004. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on Multimodal interfaces*. 205–211.

[4] Judith Butepage, Michael J Black, Danica Kragic, and Hedvig Kjellstrom. 2017. Deep representation learning for human motion prediction and classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6158–6166.

[5] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[6] Michael Duhme, Raphael Memmesheimer, and Dietrich Paulus. 2022. Fusion-gcn: Multimodal action recognition using graph convolutional networks. In *Pattern Recognition: 43rd DAGM German Conference, DAGM GCPR 2021, Bonn, Germany, September 28–October 1, 2021, Proceedings*. Springer, 265–281.

[7] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. 2017. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE international conference on computer vision*. 2334–2343.

[8] Mihai Gurban, Jean-Philippe Thiran, Thomas Drugman, and Thierry Dutoit. 2008. Dynamic modality weighting for multi-stream hmms inaudio-visual speech recognition. In *Proceedings of the 10th international conference on Multimodal interfaces*. 237–240.

[9] Danica Hendry, Kevin Chai, Amity Campbell, Luke Hopper, Peter O'Sullivan, and Leon Straker. 2020. Development of a human activity recognition system for ballet tasks. *Sports medicine-open* 6, 1 (2020), 1–10.

[10] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[11] Xiaodan Hu and Narendra Ahuja. 2021. Unsupervised 3D pose estimation for hierarchical dance video Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11015–11024.

[12] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[13] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. 2019. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10863–10872.

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[15] Beth Logan et al. 2000. Mel frequency cepstral coefficients for music modeling. In *Ismir*, Vol. 270. Plymouth, MA, 11.

[16] Hitoshi Matsuyama, Shunsuke Aoki, Takuro Yonezawa, Kei Hiroi, Katsuhiko Kaji, and Nobuo Kawaguchi. 2021. Deep Learning for Ballroom Dance Recognition: A Temporal and Trajectory-Aware Classification Model With Three-Dimensional Pose Estimation and Wearable Sensing. *IEEE Sensors Journal* 21, 22 (2021), 25437–25448.

[17] Raphael Memmesheimer, Nick Theisen, and Dietrich Paulus. 2020. Gimme signals: Discriminative signal encoding for multimodal activity recognition. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 10394–10401.

[18] Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic. 2011. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing* 2, 2 (2011), 92–105.

[19] Radoslaw Niewiadomski, Maurizio Mancini, Andrea Cera, Stefano Piana, Corrado Canepa, and Antonio Camurri. 2019. Does embodied training improve the recognition of mid-level expressive movement qualities sonification? *Journal on Multimodal User Interfaces* 13, 3 (2019), 191–203.

[20] Eftychios Protopapadakis, Athina Grammatikopoulou, Anastasios Doulamis, and Nikos Grammalidis. 2017. Folk dance pattern recognition over depth images acquired via kinect sensor. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences* 42 (2017), 587.

[21] Fabien Ringeval, Björn Schuller, Michel Valstar, Shashank Jaiswal, Erik Marchi, Denis Lalanne, Roddy Cowie, and Maja Pantic. 2015. Av+ ec 2015: The first affect recognition challenge bridging across audio, video, and physiological data. In *Proceedings of the 5th international workshop on audio/visual emotion challenge*. 3–8.

[22] Maarten J Vaessen, Etienne Abassi, Maurizio Mancini, Antonio Camurri, and Beatrice de Gelder. 2019. Computational feature analysis of body movements reveals hierarchical brain organization. *Cerebral Cortex* 29, 8 (2019), 3551–3560.

[23] Chongyang Wang, Temitayo A Olugbade, Akhil Mathur, Amanda C De C. Williams, Nicholas D Lane, and Nadia Bianchi-Berthouze. 2019. Recurrent network based automatic detection of chronic pain protective behavior using mocap and semg data. In *Proceedings of the 23rd international symposium on wearable computers*. 225–230.

[24] Chongyang Wang, Min Peng, Temitayo A Olugbade, Nicholas D Lane, Amanda C De C Williams, and Nadia Bianchi-Berthouze. 2019. Learning temporal and bodily attention in protective movement behavior detection. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, 324–330.

[25] Max Welling and Thomas N Kipf. 2016. Semi-supervised classification with graph convolutional networks. In *J. International Conference on Learning Representations (ICLR 2017)*.

[26] Sandar Win and Thin Lai Lai Thein. 2020. Real-time human motion detection, tracking and activity recognition with skeletal model. In *2020 IEEE Conference on Computer Applications (ICCA)*. IEEE, 1–5.

[27] Monika Wysoczanska and Tomasz Trzcinski. 2020. Multimodal Dance Recognition. In *VISIGRAPP (5: VISAPP)*. 558–565.

[28] Fangkai Yang, Wenjie Yin, Tetsunari Inamura, Mårten Björkman, and Christopher Peters. 2020. Group behavior recognition using attention-and graph-based neural networks. In *ECAI 2020*. IOS Press, 1626–1633.

[29] Jinming Zhao, Ruichen Li, Shizhe Chen, and Qin Jin. 2018. Multi-modal multi-cultural dimensional continues emotion recognition in dyadic interactions. In *Proceedings of the 2018 on audio/visual emotion challenge and workshop*. 65–72.