Department: Affective Computing and Sentiment Analysis
Editor: Erik Cambria, Nanyang Technological University

Multi-scale 3D Shift Graph Convolution Network for Emotion Recognition from Human Actions

Henglin Shi
University of Oulu

Wei Peng
University of Oulu

Haoyu Chen
University of Oulu

Xin Liu
Lappeenranta-Lahti University of Technology LUT

Guoying Zhao
University of Oulu

***Abstract*—Emotion recognition from body gestures is challenging since similar emotions can be expressed by arbitrary spatial configurations of joints, which results in relying on modelling spatial-temporal patterns from a more global level. However, most recent powerful graph convolution networks (GCNs) separate the spatial and temporal modelling into isolated processes, where GCN models spatial interactions using partially fixed adjacent matrices and 1-D convolution captures temporal dynamics, which is insufficient for emotion recognition. In this work, we propose the 3D-Shift GCN which enables interactions of joints within a spatial-temporal volume for global feature extraction. Besides, we further develop a multi-scale architecture, the MS-Shift GCN, to fuse features captured under different temporal ranges for modelling richer dynamics. After conducting evaluation on two regular action recognition benchmarks and two gesture based emotion recognition datasets, the results show that the proposed method outperforms several state-of-the-art methods.**

AFFECTIVE COMPUTING is essential for next generation artificial intelligence (AI) applications, and emotion analysis is a core piece of this puzzle [1]. However, current efforts on emotion analysis are more focusing on human facial expressions, less attention has been paid to emotion analysis from gestures. Human body gestures are also conveying emotions, and sometimes it could be more important than facial expressions for emotion analysis. Firstly, body gestures are easier to be observed than facial expressions, so gesture based emotion recognition is more easily to be implemented at the application level. Secondly, compared with facial expression, body gestures are more likely to express the real emotion of the person since gestures are more difficult to suppress if without training.

In this work, we focus on recognizing emotions from human body gestures using the skeleton data. Various methods have been developed for skeleton based human gesture analysis, such as handcrafted features based, recurrent neural networks (RNNs) based, and the recent graph convolution network (GCN) based [2] [3] which are gaining increasingly popularity because of their outstanding performances However, most of these studies were about actions or gestures recognition, and few works have been done on gesture based emotion analysis. We consider emotion recognition can benefit from the successful experiences of skeleton based action recognition methods.

However, unlike actions, emotions are insensitive to the spatial configurations of joints. For example, when people are performing the same action while holding different emotions, the spatial configurations of joints can be similar due to the same action, which cannot provide sufficient support for recognizing emotions. Thus, for emotion recognition, it is better to put more effort on modelling the spatial-temporal evolutions of joint interactions. Unfortunately, most current GCN based methods process the spatial dimension and temporal dimension separately, and the temporal modelling is performed with 1-D convolution, which are insufficient to capture spatial-temporal dynamics. Inspired by recently proposed Shift-GCN [2] which captures the global spatial feature by shifting channels among all joints within a frame, we propose the 3D-Shift GCN module which shifts channels among joints within a chunk of several frames to achieve the spatial-temporal feature extraction globally.

As far as the authors' concern, this work makes following contributions. Firstly, we propose the 3D-Shift GCN, a GCN based module for global spatial-temporal feature extraction for emotion recognition using human skeleton data. Additionally, we develop the MS-Shift GCN, a multi-scale architecture which integrates several 3D-Shift GCN modules with different temporal ranges for capturing richer dynamics from different scales. Moreover, we conduct extensive experiments on two large-scale action recognition datasets, one spontaneous emotional gesture dataset, and one posed emotional gesture dataset to evaluate the performance of the proposed 3D-Shift GCN and the multi-scale architecture. Lastly, an investigation of emotion recognition performances on different action classes is made and discussed.

Related work

Emotion Recognition from Body Behaviours

Compared with facial expression analysis, studies on body behaviour based emotion analysis are still few. Recently, two surveys have reviewed the progress of body behaviour based emotion analysis [4][5]. Gunes *et al.* [6] developed a multimodal analyser for emotion recognition such that facial expressions and body behaviours are treated as different modalities and processed simultaneously. Gunes and Piccaa [7] further explored different approaches for fusing gestures and faces for obtaining better emotion recognition performance. Zadeh *et al.* [8] analyzed emotions based on facial gestures such as head nod and head shake. Castellano *et al.* [9] solely utilized quantities of body movements, such as amplitude, speed and fluidity, for emotion recognition. Kipp and Martin [10] investigated the correlation between basic gestures and emotions. Recently, [11] involved head poses for estimating depression levels, and [12] conducted sentiment analysis based on transcriptions extracted from videos.

These works mentioned above were conducted using the appearance data of body gestures, for example images. Saha *et al.* [13] carried

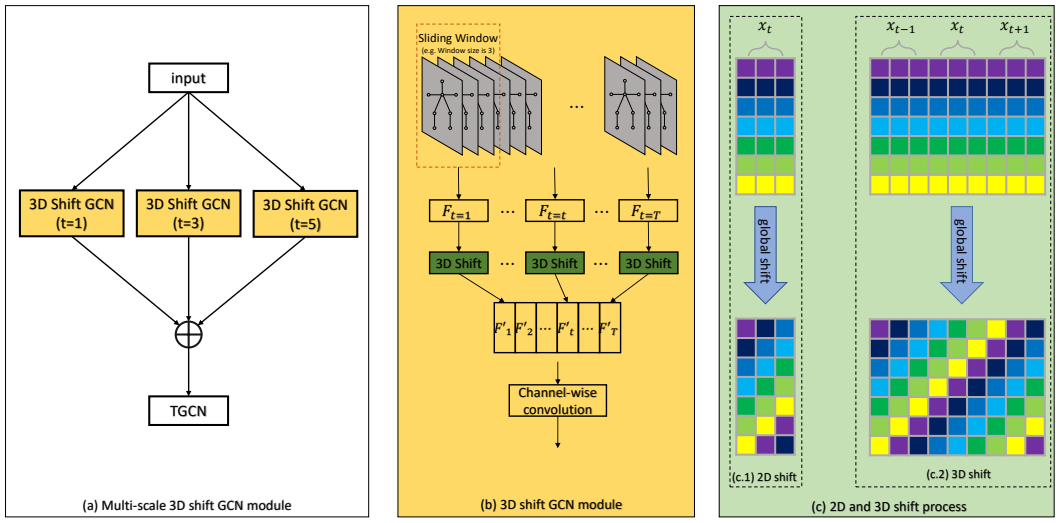

Figure 1. Diagrams of the proposed method: (a) the overall architecture of the proposed Multi-scale 3D-Shift GCN, τ denotes the temporal scale of the module; (b) the working flow of the 3D shift GCN; (c.1) 2-D feature shift used in [2]; and (c.2) the proposed 3d feature Shift.

out the recognition emotion task on skeleton data of body gestures. The data is captured using Kinect sensors. Fourati and Pelachaud [14] collected an emotion action dataset which includes high quality skeleton captured using Motion Capture system. Besides, the authors [5] collected a multi-label action emotion dataset which also provides skeleton data extracted based on RGB videos.

Action Recognition GCN

Graph convolution network has shown its outstanding performances on skeleton based action recognition. Yan *et al.* [3] firstly introduced GCN to process human skeleton by organizing human skeleton data as graphs according to the natural kinesiology connectivity of human bodies. However, researchers found that the constant-style adjacent matrix in [3] is not able to capture longer ranger features from skeleton data [15]. Specifically, the adjacent matrix defined in [3] only manifests the connectivity of 1-hop neighbour, which means joints are physically connected, so that feature between further joints will not be accumulated but there does exist patterns for example feed and hands when people are walking.

To solve this problem, [15] traded the performance with the complexity, which captured longer range spatial features using several higher order adjacent matrices which can capture relationships from $L-$hop neighbors of joints. Besides, [16] tried to solve the problem from another direction that proposed to use learnable adaptive adjacent matrices to augment the handcrafted one so that the model is able to learn to extract long range features. Furthermore, Liu *et al.* [17] proposed MS-G3D which enlarges the receptive field in the temporal domain and achieved dominant performances on several datasets. To this end, GCN based methods are getting extremely large, and some researchers start to seek light-weight solutions.

Zhang *et al.* [18] developed a light-weight GCN with shallow architecture which achieves state-of-the-art performances. Cheng *et al.* [2] presented the Shift GCN to reduce the computation cost by eliminating the dependence to the adjacent matrix, which reduced the computation cost by $5 \sim 10$ times while achieving the state-of-the-art performance.

Proposed Method

Figure 1(a) and 1(b) illustrate the general architecture of the proposed MS-Shift GCN and detailed working flow of the 3D-Shift GCN, respectively. In this section, we firstly brief some preliminary of GCN. Moreover, we introduce the proposed 3D-Shift GCN based on Shift GCN proposed in [2]. Lastly, we explain the multi-scale mechanism of the proposed method.

Preliminaries

In recent years, GCN based methods have dominated skeleton based action recognition tasks, in which an action sequence (or feature map) is organized as $X \in \mathbb{R}^{T \times V \times C_l}$, where T, V and C_l denote the temporal length, joint number, and channel number of the l^{th} layer of the action sample, respectively. Given one frame x_t from X where $x_t \in \mathbb{R}^{V \times C_l}$, the feature map of GCN at layer l is calculated according to:

$$x_t^{l+1} = \sum_k^{K_v} A_k (x_t^l W_k^l) \odot M_k^l, \tag{1}$$

where $W_k^l \in \mathbb{R}^{C_l \times C_{l+1}}$ denotes the trainable parameter at layer l which is applied on each channel. $A_k \in \mathbb{R}^{V \times V}$ is the adjacent matrix which manifests the spatial relationship between different joints such that its entry $A_k^{(i,j)} = 1$ if and only if joint j and i are connected, and belongs to the partition k. Specifically, in the skeleton based action recognition, most GCNs chose that $K_v = \{root, centripetal, centrifugal\}$. For example, $A_{centripetal}^{(i,j)} = 1$ means joint j connects to i and joint j is closer to the gravity center of the body than joint i. Lastly, M_k^l is the importance for each joint, which is applied element-wisely.

Shift GCN

In ordinary GCN related methods, the capability of spatial feature extraction is endowed by adjacent matrices, which has two shortcomings. Firstly, the spatial receptive is limited by the adjacent matrices. Even with augmented adaptive adjacent matrices, the optimization of adjacent matrices is not guaranteed. Moreover, the multiplication of the adjacent matrices introduces extract computations. Shift GCN proposed to shift the channel of feature maps so that the receptive field of the model is enlarged to the global level without using adjacent matrices. By removing the dependency of the adjacent matrix, the computational cost is reduced by $\sim 75\%$ compared with the vanilla ST-GCN [3].

The Shift process in [2] is described in Figure 1(c.1). Here we denote the shifting process by a function $Shift(\cdot)$. For a feature map of a skeleton is $x_t \in \mathbb{R}^{V \times C_l}$, the shifting process attempts to circularly shift each channel in the channel direction with different distances, and the shift distance is corresponding channel index $mode$ V. Replacing the adjacent matrix and the partition summation with the shifting process so that we can obtain the formulation of Shift GCN:

$$x_t^{l+1} = Shift(x_t^l) W^l \odot M^l. \tag{2}$$

3D shift GCN Operator

In this work, we extend the Shift-GCN in the temporal dimension. In order to enable the GCN module to extract the interactions between nodes from different temporal locations, e.g. increasing the receptive field in the temporal dimension, we introduce auxiliary frames from the temporal neighbours.

As Figure 1.b shows, firstly, considering current frame as the anchor frame, a sliding window samples previous frames and later frames of the anchor frames according to a given temporal range τ so that the sampling result $x \in \mathbb{R}^{\tau \times V \times C_l}$. Secondly, x is reshaped to $\mathbb{R}^{V \times (C_l \tau)}$ such that channels of every joint from every neighbouring frames are appended to the corresponding joints of the anchor frame. Therefore, features from other frames can be viewed as auxiliary channels of the anchor frame. Lastly, the global shift mechanism [2] is applied to the reshaped input for feature extraction:

$$x_t^{l+1} = Shift(x_{t-\tau:t+\tau}^l) W^l \odot M^l, \tag{3}$$

where $W \in \mathbb{R}^{C_l \tau \times C_{l+1}}$ is the weight of the proposed 3D-Shift GCN, and M^l is also the importance mask like other GCNs.

Multi-scale 3D-Shift GCN

The multi-scale mechanism has been explored in [15] and [17]. In this work, we propose the multi-scale mechanism based on the proposed 3D-Shift GCN. Define set $S = \{\tau_1, \tau_2, \cdots, \tau_m\}$, the feature of the scale $\tau_s \in S$ is equivalent to making $\tau = \frac{\tau_s - 1}{2}$ in Eq. (3). At last, the multi-scale feature is obtained by summing all feature from different scales:

$$X_t^{l+1} = \sum_{\tau_s \in S} Shift(X_{t-\tau:t+\tau}^l) W_s^l \odot M^l, \tag{4}$$

where $\tau = \frac{\tau_s - 1}{2}$.

Like other GCN models such as ST-GCN [3], 2s-AGCN [16], the proposed model is designed as the basic module to construct a GCN layer. In this work, following [2], a deep multi-scale 3D-Shift GCN network is constructed by stacking several of the proposed modules followed by a temporal Shift GCN module.

Experiment

Datasets

NTU RGB+D is a large-scale action recognition dataset which is selected in this work to evaluate the effectiveness of the proposed method on regular action recognition [19]. This dataset collects 56,880 action samples of 60 categories from 40 subjects and 3 view angles. This dataset offers four modalities, RGB frames, depth maps, body masks, and skeleton joints, and in this work only the skeleton modality is used. This dataset provides two evaluation protocols: (1) the cross-subject (Xsub) protocol which assigns samples collected from half of the subjects (20 subjects) for training, and the rest of the subjects for testing; and (2) the cross-view (Xview) protocol which assigns samples collected from view 2 and view 3 for training, and samples from view 1 for testing.

NTU RGB+D 120 is an extended version of the NTU RGB+D dataset. The numbers of its samples, classes, and participated subjects are increased to 111,480, 120, and 106, respectively [20]. Instead of collecting data from multi-views, samples of this data are recorded under 32 setups, and each setup specifies different backgrounds and locations. This dataset provides two default evaluation protocols: (1) the cross-subject (Xsub) protocol which assigns samples collected from half of the subjects (56 subjects) for training, and the rest of the subjects for testing; and (2) the cross-setup (Xset) protocol which assigns samples with even setup IDs for training, and the left setups for testing.

Emilya [14] is an emotion action dataset contains 8 emotions including **Ax (Anxiety)**, **Pr (Pride)**, **Jy (Joy)**, **Sd (Sad)**, **PF (Panic Fear)**, **Sh (Shame)**, **Ag (Anger)**, and **Nt (Neutral)**. Each emotion is expressed under 8 types of actions, including *SW (Simple Walk)*, *MB (Move Books)*, *WH (Walk with an object in Hand)*, *KD (Knock at the Door)*, *BS (Being Seated)*, *SD (Sit Down)*, *Lf (Lift an object)*, and *Th (Throw an object)*. In this part we apply the proposed method to recognize emotions using the skeleton data of this dataset. The evaluation protocol is 3-fold cross-validation adopted from the original publication [14], where 1/3 of the data is selected for testing and the rest is selected for training. This process is rotated for three times to make sure all data has been tested We report the averaged recognition accuracy of the testing result from all three rotations.

Spontaneous Micro-Gesture (SMG) is collected for analysing subtle human body movements called 'micro-gestures' which convey human true hidden emotions [21]. In this dataset, 3,692 micro-gesture of 17 classes are collected from 40 subjects. All micro-gestures are performed spontaneously. In the experiment, we conduct micro-gesture recognition on each individual micro-gesture clip. The evaluation protocol adopted is cross-subject evaluation where 35 subjects are used for training, and the rest 5 subjects are used for testing.

Experimental settings

We implement the MS-Shift GCN network by stacking 10 GCN layers where each layer is consist of one MS-Shift GCN module and one Shift TCG module proposed by [2]. In this experiment, the implemented MS-Shift GCN involves the scales of $S = \{1, 3, 5\}$. For comparison purposes, we also implement single-scale 3D-Shift GCN networks with temporal range of 1, 3, and 5.

To train the network, a cross-entropy loss function and the stochastic gradient descent (SGD) optimizer are used. For all datasets, the initial learning rate is set to 0.1, and the weight decay is set to 0.0001. The learning rate is reduced by 10 times at epoch 60, 80, and 100, respectively. Like other GCN based methods, all input sequences are padded by themselves circularly to reach a specific temporal length. Specifically, samples of the NTU and NTU 120 datasets are padded to 300 frames, samples of the Emilya dataset are padded to 600 frames, and samples of SMG are padded to 90 frames.

Table 1. Top-1 accuracy % of different 3D-Shift GCN networks and MS-Shift GCN network.

Methods	S={1}	S={3}	S={5}	S={1,3,5}
NTU Xsub	87.8	88.0	88.3	**89.0**
NTU Xview	95.1	95.2	95.1	**95.3**
NTU 120 Xsub	80.9	**82.4**	82.1	82.2
NTU 120 Xset	83.2	83.9	83.3	**84.7**

Table 2. Top-1 accuracy % compared with state-of-the-art methods on NTU RGB+D and NTU RGB+D 120 datasets.

Methods	NTU		NTU-120	
Methods	Xsub	Xview	Xsub	Xset
PA-LSTM [19]	62.9	70.3	25.6	26.3
ST-LSTM [22]	69.2	77.7	55.7	57.9
GCA-LSTM [23]	76.1	84.0	61.2	63.3
ST-GCN [3]	81.5	88.3	-	-
AS-GCN [15]	86.8	94.2	-	-
AGC-LSTM [24]	89.2	95.0	-	-
AGCN 2s [16]	88.5	95.1	82.9	84.9
SGN [18]	89.0	94.5	79.2	81.5
Shift-GCN Js [2]	87.8	95.1	80.9	83.2
Shift-GCN 2s [2]	89.7	96.0	85.3	86.6
3D-Shift-3 Js	88.0	95.2	**82.4**	83.9
3D-Shift-3 Bs	88.7	95.09	83.9	85.7
3D-Shift-3 2s	90.0	96.27	85.9	87.4
MS-Shift Js	**89.0**	**95.3**	82.2	**84.7**
MS-Shift Bs	**89.1**	**95.11**	**84.7**	**86.3**
MS-Shift 2s	**90.6**	**96.33**	**86.2**	**88.1**

Ablation study

To evaluate the effectiveness of the proposed 3D Shift strategy and multi-scale architecture, we compare the single-scale Shift GCN ($S = \{1\}$) which is equivalent to the Shift GCN in [2]; single-scale 3D-Shift GCNs with temporal range of 3 ($S = \{3\}$) and 5 ($S = \{5\}$); and MS-Shift GCN ($S = \{1, 3, 5\}$) at each column of Table 1.

All results are obtained based on the data of raw joints, and we refer $S = \{1\}$ as the baseline model. According to the table, the two single-scale 3D-Shift GCN networks outperform the baseline mode, which shows the 3D-Shift GCN is more effective than the 2D-Shift GCN. Moreover, the multi-scale architecture outperforms other three single-scale networks on most dataset protocols except the $S = 3$ on NTU 120 Xsub protocol.

Action Recognition Performances

Table 2 presents the experimental results of the single-scale 3D-Shift GCN network with $S = \{3\}$ and the MS-Shift GCN with $S = \{1, 3, 5\}$ achieved on the NTU RGB+D (columns NTU) and the NTU RGB+D 120 (columns NTU-120) datasets. The two methods are denoted as **3D-Shift-3** and **MS-Shift**, respectively. Each method is evaluated under three data modalities: (1) Joint stream (Js) which only uses the joint stream; (2) Bone stream (Bs) which only uses the bone stream; and (3) two-stream fusion (2s) which uses both of joint and bone streams. We select several state-of-the-art methods for comparison, including RNN based methods [19][22][23] and most related GCN based methods [3][15][16][2][24][18].

As the table shows, the proposed multi-scale and single-scale 3D-Shift GCN outperform selected comparison methods on both NTU RGB+D and NTU RGB+D 120 datasets. Besides, on most modalities the proposed MS-Shift GCN surpasses the single-scale version, except using the Joint data only on the NTU 120 dataset Xsub protocol. This may due to the multi-scale fusion method used is not optimal. In our future work, we will explore different fusion methods.

Table 3. Top-1 accuracy % of micro-gesture recognition on the SMG dataset

Methods	Accuracy
STGCN [3]	41.5
AGCN Js [16]	43.1
Shift-GCN Js [2]	55.3
3D-Shift-3 Js	60.5
3D-Shift-5 Js	61.3
MS-Shift Js	**61.5**

Table 4. Top-1 accuracy % of emotion recognition on the Emilya dataset.

Methods	Accuracy
Body Cues Rating [14]	32.0
Random Forest [25]	84.8
AGCN Js [16]	84.4
Shift-GCN Js [2]	91.7
3D Shift-3 Js	91.5
3D-Shift-5 Js	91.3
MS-Shift Js	**92.0**

Emotion Recognition Performances

For the performances of the proposed methods on body gesture based emotion recognition, Table 3 presents the results on the SMG dataset which show that the proposed MS-Shift GCN outperforms the single-scale 3D-Shift GCN and other comparison methods. Table 4 presents the performances of the proposed methods on the Emilya dataset. Except GCN related methods, we also compare with the methods of Body Cues Rating [14] and Random Forest [25]. The results show

Table 5. Emotion recognition accuracy on each action achieved by the MS-3D Shift GCN on the Emilya dataset.

	SW	MB	WH	KD	BS	SD	Lf	Th
Ax	77.8	86.8	90.9	89.6	91.1	95.3	85.7	92.5
Pr	100.0	95.1	96.3	97.2	92.1	81.4	91.2	92.3
Jy	88.1	89.5	90.0	85.7	89.7	87.8	94.7	82.0
Sd	97.6	92.3	100.0	97.6	89.1	86.8	100.0	93.0
PF	95.3	89.8	94.4	94.0	93.0	93.9	93.5	98.0
Sh	96.3	91.7	97.5	98.0	89.7	91.1	96.4	98.0
Ag	95.8	92.7	100.0	97.6	95.5	91.4	85.5	95.5
Nt	96.4	100.0	97.0	96.3	96.3	83.3	100.0	95.8
Avg	93.1	92.0	95.5	94.2	91.9	89.1	92.8	93.3

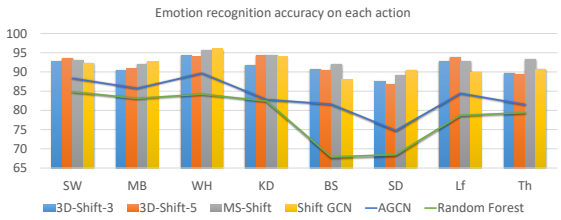

Figure 2. Average emotion recognition accuracy of each action on the Emilya dataset.

that MS-Shift GCN outperforms these selected methods.

We also conduct action specific analysis for emotion recognition on Emilya dataset, which is presented in the Table 5. The proposed method achieves the accuracy higher than 85% for most emotion and action combinations, except the emotion **Ax** on action *SW*, and emotions **Pr** as well as **Nt** on action *SD*.

Besides, we summarize the average emotion recognition accuracies on each action of six methods, which are presented in the Figure 2. Based on the figure, we can find that the MS-Shift GCN performs the best on recognizing emotions performed under several actions, namely *KD*, *BS*, and *Th*. Especially, the improvements under the action of *BS* and *Th* are promising.

Conclusion

In this work, we propose the 3D-Shift GCN which is effective for modeling global spatial-temporal dynamic for recognizing emotions using human body skeleton data. Moreover, we further develop the MS-Shift GCN which integrates several 3D-Shift GCN modules with different temporal ranges for capturing richer temporal dynamics.

Based on the experimental results, we obtain following conclusions. Firstly, the ablation study shows that the 3D-Shift GCN outperforms the baseline 2D-Shift GCN, and the MS-Shift GCN can further improve the performance of the single-scale model. Additionally, the experiments show that the proposed methods outperform several state-of-the-art methods on selected four datasets, which demonstrates that the proposed methods are not only effective for regular action recognition, but also for spontaneous and posed emotion recognition. Lastly, we investigate the effect of different actions on emotion recognition performances, and find that recognizing emotions from few actions could be relative harder.

There are three directions for our future work. For one thing, we will develop a new data dependent shift strategy that can adapt to the input features, which is expected to provide better performances. Moreover, we will also investigate other multi-scale fusion methods to improve the performance. Besides, we will develop in-depth analysis on how different actions can influence the recognition of emotions, and this result is expected to provide us qualitative instruction for developing body gesture based emotion recognition methods.

ACKNOWLEDGMENT

This work is supported by the Academy of Finland for ICT 2023 project (grant 328115) and project MiGA (grant 316765) and Infotech Oulu. As well, the authors wish to acknowledge CSC-IT Center for Science, Finland, for computational resources.

REFERENCES

1. E. Cambria, “Affective Computing and Sentiment Analysis,” *IEEE Intelligent Systems*, vol. 31, no. 2, 2016, pp. 102–107.
2. K. Cheng et al., “Skeleton-Based Action Recognition With Shift Graph Convolutional Network,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 183–192.
3. S. Yan, Y. Xiong, and D. Lin, “Spatial temporal graph convolutional networks for skeleton-based action recognition,” *Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32*, 2018.
4. F. Noroozi et al., “Survey on emotional body gesture recognition,” *IEEE Transactions on Affective Computing*, 2018.
5. Y. Luo et al., “Arbee: Towards automated recognition of bodily expression of emotion in the wild,” *International*

Journal of Computer Vision, vol. 128, no. 1, 2020, pp. 1–25.

6. H. Gunes, M. Piccardi, and T. Jan, "Face and Body gesture recognition for a vision-based multimodal analyser," *Pan-Sydney Area Workshop on Visual Information Processing*, 2004.
7. H. Gunes and M. Piccardi, "Affect recognition from face and body: early fusion vs. late fusion," *2005 IEEE International Conference on Systems, Man and Cybernetics, vol. 4*, 2005, pp. 3437–3443.
8. A. Zadeh et al., "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intelligent Systems*, vol. 31, no. 6, 2016, pp. 82–88.
9. G. Castellano, S. D. Villalba, and A. Camurri, "Recognising human emotions from body movement and gesture dynamics," *International Conference on Affective Computing and Intelligent Interaction*, 2007, pp. 71–82.
10. M. Kipp and J.-C. Martin, "Gesture and emotion: Can basic gestural form features discriminate emotions?" *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, 2009, pp. 1–8.
11. S. A. Qureshi et al., "Multitask representation learning for multimodal estimation of depression level," *IEEE Intelligent Systems*, vol. 34, no. 5, 2019, pp. 45–52.
12. L. Stappen et al., "Sentiment analysis and topic recognition in video transcriptions," *IEEE Intelligent Systems*, vol. 36, no. 2, 2021, pp. 88–95.
13. S. Saha et al., "A study on emotion recognition from body gestures using Kinect sensor," *2014 International Conference on Communication and Signal Processing*, 2014, pp. 056–060.
14. N. Fourati and C. Pelachaud, "Perception of emotions and body movement in the emilya database," *IEEE Transactions on Affective Computing*, vol. 9, no. 1, 2016, pp. 90–101.
15. M. Li et al., "Actional-structural graph convolutional networks for skeleton-based action recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3595–3603.
16. L. Shi et al., "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12026–12035.
17. Z. Liu et al., "Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 143–152.
18. P. Zhang et al., "Semantics-Guided Neural Networks for Efficient Skeleton-Based Human Action Recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1112–1121.
19. A. Shahroudy et al., "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1010–1019.
20. J. Liu et al., "Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.
21. H. Chen et al., "Analyze spontaneous gestures for emotional stress state recognition: A micro-gesture dataset and analysis with deep learning," *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, 2019, pp. 1–8.
22. J. Liu et al., "Spatio-temporal lstm with trust gates for 3d human action recognition," *European Conference on Computer Vision*, 2016, pp. 816–833.
23. J. Liu et al., "Skeleton-based human action recognition with global context-aware attention LSTM networks," *IEEE Transactions on Image Processing*, vol. 27, no. 4, 2017, pp. 1586–1599.
24. C. Si et al., "An attention enhanced graph convolutional lstm network for skeleton-based action recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1227–1236.
25. N. Fourati and C. Pelachaud, "Multi-level classification of emotional body expression," *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), vol. 1*, 2015, pp. 1–8.

Henglin Shi is currently a Ph.D. candidate with the Center for Machine Vision and Signal Analysis, University of Oulu, Finland. He received his B.S. and M.S. degrees in Computer Science and Information Processing Science in 2012 and 2016, respectively. His research interests include machine learning and computer vision based human behavior analysis.

Wei Peng is currently a Ph.D. candidate with the Center for Machine Vision and Signal Analysis, University of Oulu, Oulu, Finland. He received the M.S. degree in computer science from the Xiamen University, Xiamen, China, in 2016. His current research interests include machine learning, affective computing, medical imaging, and human action analysis.

Haoyu Chen is currently a Ph.D. candidate with the Center for Machine Vision and Signal Analysis, under the supervision of Prof. G. Zhao. He received the B.Sc. degree from the China University of Geosciences, Wuhan, China, in 2015, and the

M.Sc. degree in computer sciences and engineering from the University of Oulu, Finland, in 2017. His research interests include action, gesture recognition, and emotional AI.

Xin Liu is currently an Associate Professor with Computer Vision and Pattern Recognition Laboratory, School of Engineering Science, Lappeenranta-Lahti University of Technology LUT, Finland. He received his Ph.D. degree in Computer Science and Engineering in 2019. His research interests include human behavior analysis, affective computing, image restoration, and object detection.

Guoying Zhao is the corresponding author and is currently a Professor with the Center for Machine Vision and Signal Analysis, University of Oulu, Finland. She received the Ph.D. degree in computer science from the Chinese Academy of Sciences, Beijing, China, in 2005. Her current research interests include affective computing, facial-expression and micro-expression recognition, emotional gesture analysis, and human computer interaction.