International Journal of Computer Vision (2023) 131:1346–1366
https://doi.org/10.1007/s11263-023-01761-6

SMG: A Micro-gesture Dataset Towards Spontaneous Body Gestures for Emotional Stress State Analysis

Haoyu Chen[1] · Henglin Shi[1] · Xin Liu[2] · Xiaobai Li[1] · Guoying Zhao[1]

Received: 4 May 2022 / Accepted: 3 January 2023 / Published online: 16 February 2023

Abstract
We explore using body gestures for hidden emotional state analysis. As an important non-verbal communicative fashion, human body gestures are capable of conveying emotional information during social communication. In previous works, efforts have been made mainly on facial expressions, speech, or expressive body gestures to interpret classical expressive emotions. Differently, we focus on a specific group of body gestures, called micro-gestures (MGs), used in the psychology research field to interpret inner human feelings. MGs are subtle and spontaneous body movements that are proven, together with micro-expressions, to be more reliable than normal facial expressions for conveying hidden emotional information. In this work, a comprehensive study of MGs is presented from the computer vision aspect, including a novel spontaneous micro-gesture (SMG) dataset with two emotional stress states and a comprehensive statistical analysis indicating the correlations between MGs and emotional states. Novel frameworks are further presented together with various state-of-the-art methods as benchmarks for automatic classification, online recognition of MGs, and emotional stress state recognition. The dataset and methods presented could inspire a new way of utilizing body gestures for human emotion understanding and bring a new direction to the emotion AI community. The source code and dataset are made available: https://github.com/mikecheninoulu/SMG.

Keywords Micro-gestures · Gesture recognition · Emotion recognition · Statistical modeling · Deep learning · Affective computing

1 Introduction

Human beings are innately able to express and interpret emotional expressions via various non-verbal communication (Shiffrar et al. 2011), which should also be an indispensable part of intelligent agents. As an important non-verbal communicative fashion, human body gestures are capable of conveying rich emotional information during social communication (Aviezer et al. 2012). However, as shown in Fig. 1a, when it comes to machines, analyzed emotional cues were mostly limited to human facial expressions and speech (El Ayadi et al. 2011; Li and Deng 2020).

Compared to other modalities, body gestures have several advantages in emotion recognition tasks. Firstly, the data acquisition of body gestures is more accessible, especially when high-resolution surveillance cameras or portable microphones are not available for capturing facial expressions or speech in public areas (e.g., airports, metro, or stadiums). With the recent success of deep learning on large-scale datasets, concerns about privacy protection and ethical

Communicated by Maja Pantic.

✉ Guoying Zhao
guoying.zhao@oulu.fi

Haoyu Chen
chen.haoyu@oulu.fi

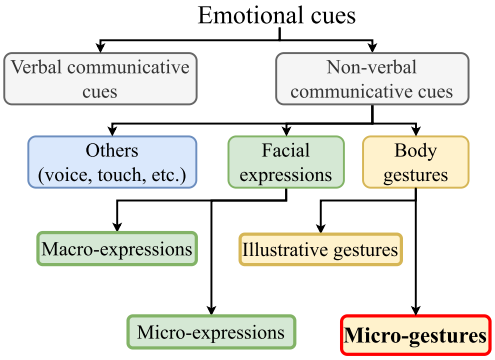
Henglin Shi
henglin.shi@oulu.fi

Xin Liu
xin.liu@lut.fi

Xiaobai Li
xiaobai.li@oulu.fi

[1] Center for Machine Vision and Signal Analysis (CMVS), University of Oulu, Oulu, Finland

[2] Computer Vision and Pattern Recognition Laboratory, School of Engineering Science, Lappeenranta-Lahti University of Technology LUT, Lappeenranta, Finland



(a) Micro-gestures as one of the emotional cues.

(b) Scenarios that micro-gestures apply.

Fig. 1 **a** Taxonomy of emotional cues. MGs serve as one of the non-verbal communicative cues for emotional understanding. **b** Example scenarios to which MG recognition can be applied. In the interview or game, the subjects tend to hide their intentions, while MGs can leak their hidden emotions

issues have started to emerge (Oh et al. 2016). Meanwhile, body gestures involve less identity information, which is promising. Lastly, studies (Ekman 2004) showed that when people were trying to hide their emotions, most of them would attempt to tune their facial expressions but could not prohibit their *micro-expressions*. Besides, only a few people referred to the need to manage their body movements. Thus, it would be encouraging to use gestures to capture people's suppressed/hidden emotions.

With the above observations, this study focuses on a specific group of gestures called Micro-Gesture (MG) for emotional understanding. However, unlike any previous research that uses expressive body gestures to interpret classical expressive emotions, we propose a brand new research topic: analyzing people's hidden emotional states with MGs. MGs are defined as subtle and involuntary body movements that reveal peoples' *suppressed/hidden* emotions. They are often used in the psychology research field to interpret inner human feelings (Serge 1995). Although MGs cover a wide group of gestures (e.g., scratching the head, touching the nose, playing with clothes), they share one important attribute which differentiates them from other gestures: MGs are not performed for any illustrative or communicational purposes at all; they are spontaneous or involuntary body responses to the onset of certain stimuli, especially negative ones. Meanwhile, ordinary gestures are usually performed to facilitate communications, e.g., to illustrate specific semantic meanings or to explicitly express one's feelings or attitudes, which are referred to as illustrative gestures or iconic gestures (Khan and Ibraheem 2012). As shown in Fig. 1b, in high-stake situations such as interviews and games, although the subjects try to conceal or suppress their true feelings for either gaining advantage (win the game) or avoiding loss (keep social image), they spontaneously initiate some body gestures responding to the stimuli. Studies (Pentland 2008) showed that these gestures are important clues in revealing people's hidden emotional status, especially negative feelings such as stress, nervousness, and fear, which can be used to detect anomalous mental status, e.g., for Alzheimer's or autism diagnosis. Expectedly, automatic MG recognition has great potential in applications, i.e., human-computer interaction, social media, public safety, and health care (Krakovsky 2018).

The study aims to answer this research question: *How to train a machine to recognize and better understand hidden human emotions via body gestures like a trained expert?* Specifically, we break down this question into several sub-problems with corresponding solutions: (1) Unlike the Action Units (AU) in facial action coding system (FACS) (Ekman 1997), a common standard is absent for body gesture-based emotion measurement. The lack of this empirical guidance leaves even psychological professionals without complete agreement on annotating bodily expressions (Luo et al. 2020). Thus, we present a novel dataset of MGs, which was collected under objective proxy tasks to stimulate two states of emotional stress. (2) The high heterogeneity in the same gesture class makes the classification of MG much more complicated than ordinary gestures. Thus, we provide various state-of-the-art models from recent top computer vision venues to demonstrate the benchmark. (3) Accurately spotting MGs from unconstrained streams is another highly challenging task, as MGs are subtle and rapid body movements that can easily be submerged in other unrelated body movements. To this end, we propose a novel online detecting method that has a parameter-free attention mechanism to differentiate MGs from non-MGs adaptively. (4) The conventional paradigm that imposes each gesture with an emotional state does not resemble real-world scenarios, we explore a new paradigm that achieves emotional understanding by holistically considering all the MGs.

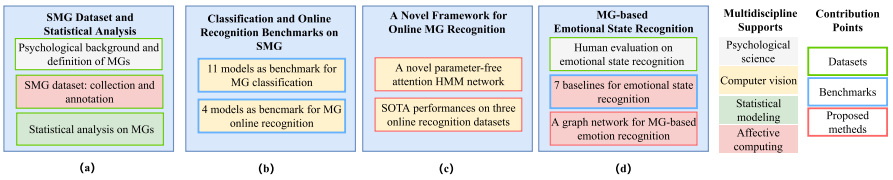


Fig. 2 The overview of the main research topics of this work. **a** A novel SMG dataset with a comprehensive statistical analysis. **b** Multiple benchmarks on the SMG dataset. **c** A novel online MG recognition framework for complicated gesture transition patterns. **d** Baselines and a newly proposed framework for emotional state recognition

As shown in Fig. 2, this work consists of four main research topics for comprehensively researching MGs from the computer vision aspect, and the contributions of each topic can be summarized as follows:

1. To the best of the authors' knowledge, this is the first work to investigate MGs with computer vision technologies for hidden emotional state analysis. A new MG dataset is built through interdisciplinary efforts, which contains rich MGs towards spontaneous body emotional stress states understanding.
2. Comprehensive statistical analysis is conducted on the relationship between body gestures and emotional stress states, investigating the various features of MGs. Various benchmark results for classifying and online recognizing MGs are reported based on multiple state-of-the-art methods.
3. A hidden Markov model (HMM) recurrent network for online MG recognition is proposed with a novel *parameter-free* attention mechanism. The method is intensively validated on three online gesture recognition datasets with competitive performances.
4. A novel paradigm is explored via a spectral graph-based model to infer the emotional states via MGs clues of the holistic videos, instead of the previously prevailing one-gesture-one-emotion paradigm.

This research is based on our previous work (Chen et al. 2019), but extended in several aspects: 1) more comprehensive dataset statistical analysis, 2) extensive benchmark experimental results with state-of-the-art methods, 3) an HMM recurrent network with a novel *parameter-free* attention mechanism validated on three datasets, and 4) a spectral graph neural network as a baseline for emotional stress state recognition.

The rest of this paper is structured as follows. Section 2 reviews related work in the literature. The SMG dataset and its analysis are presented in Sect. 3. Benchmarks of MG classification are provided in Sect. 4. Section 5 focuses on the online GM recognition task with a newly proposed method. Body gesture-based emotional stress state recognition is conducted in Sect. 6, and we conclude the work in Sect. 7.

2 Related Work

2.1 Body Gesture Recognition in Computer Vision

Accurate recognition is the foundation of all the further implementations of body gestures, such as gesture language recognition, human-robot interaction, and also emotional gesture recognition (Carreira and Zisserman 2017; Shahroudy et al. 2016; Soomro et al. 2012). Over the decades, human gesture recognition has been intensively researched in the field of computer vision. From the machine learning point of view, body gesture recognition can be sorted into two settings: 1) the classification of pre-segmented body gestures and 2) temporal body gesture detection and recognition upon the long non-stationary sequence. The former task that conducts the classification of the pre-segmented clips draws more attention from researchers, and most of the existing state-of-the-art technologies can achieve considerably promising performances. Towards video-based resources such as RGB, depth, and optical flow data, classical models for the body action and gesture classifications mainly includes 2DCNN families (Lin et al. 2019; Wang et al. 2018; Xu et al. 2019a) and 3DCNN families (Carreira and Zisserman 2017; Hara et al. 2018; Tran et al. 2015). Based on skeleton resources obtained from such as Kinect (Shotton et al. 2011) or OpenPose (Cao et al. 2019), state-of-the-art methods nowadays are mainly derived from graph-based convolutional networks (Cheng et al. 2020; Liu et al. 2020; Peng et al. 2020; Shi et al. 2019; Yan et al. 2018). Methods are also proposed to fuse different resources and modalities (Crasto et al. 2019; Sun et al. 2018; Yu et al. 2020). When it comes to the latter online recognition setting, research efforts are relatively few due to the computational complexity (Chen et al. 2020; Li et al. 2016; Liu et al. 2018; Neverova et al. 2016; Wu et al. 2016; Xu et al. 2019b). Different from other gesture recognition tasks such as gesture language recognition, the recognition of emotional body gestures and MGs has its specific challenges: (1) the duration ranges from several frames to hundreds of frames; (2) the kinetic scale varies from only subtle finger movements to overall body changes; (3) the variations of the movements associated with a gesture can be large due to the individual differences of subjects

and (4) meaningful emotional gestures are submerged within plenty of irrelevant body movements.

2.2 Human Emotion Recognition with Body Gestures

Recognizing emotional states through body movements has been researched for decades (Noroozi et al. 2018). Previous works are mainly based on one-gesture-one-emotion assumptions with two kinds of emotional modeling theories (Noroozi et al. 2018): the categorical and dimensional models. In the categorical model-based methods (Ginevra et al. 2008; Gunes and Piccardi 2006; Mahmoud et al. 2011), each emotion was imposed with a meaningful gesture, and participants were asked to act on those emotions with their body gestures. Recently, some researchers have explored the possibility of analyzing bodily expression with a dimensional model (Kipp and Martin 2009; Luo et al. 2020). In the work of Luo et al. (2020), the emotions of body gestures collected from movie clips are defined by the dimensions of arousal and valence. However, an essential feature of emotional gestures is neglected in all of these works: not all the body movements are highly emotion-driven (Pentland 2008) and body language could be interpreted differently by subject differences (Yu 2008). Thus, it is not convincing and accurate to interpret each isolated gesture as an emotional state and not consider subject differences. As expected, the agreement on the interpretation of one bodily expression between annotators is considerably low. For instance, during the emotion annotation in the work of Luo et al. (2020), annotators still primarily rely on facial expressions rather than gestures. This issue makes the research limited to be extended to real-world implementation.

2.3 Emotional Body Gesture Datasets

Compared to regular human gesture analysis, such as body pose, action, or sign language recognition, research efforts devoted to using gestural behaviors to interpret human emotion or affection are relatively few (Noroozi et al. 2018). The pioneering work for gesture-based emotion recognition in the computer vision field can go back more than 20 years ago (Ginevra et al. 2008; Gunes and Piccardi 2006; Schindler et al. 2008; Wallbott 1998). Wallbott (1998) collected 224 videos and, in each of their records, an actor acting a body gesture representing an emotional state through a scenario approach. In the work of Schindler et al. (2008), an image-based dataset was collected in which emotions were displayed by body language in front of a uniform background and different poses could express the same emotion. Gunes and Piccardi (2006) introduced a Bimodal face and body gesture database, called FABO, including facial and gestural modalities. Different from the above laboratory settings, Kipp and Martin (2009) proposed a Theater corpus based on two movie versions of the play Death of a Salesman trying to explore the correlations between basic gesture attributes and emotion. It also provided the emotion dimensions of pleasure, arousal, and dominance instead of emotion-specific discrete expression models. Similarly, Luo et al. (2020) collected a large-scale dataset called BoLD (Body Language Dataset) that also includes both discrete emotions and dimensional emotions. In the BoLD dataset, each short video clip has been annotated for emotional body expressions as perceived by viewers via a crowd-sourcing strategy. However, those datasets were all designed for classical expressive emotions, and none of them is specifically for hidden emotional state understanding.

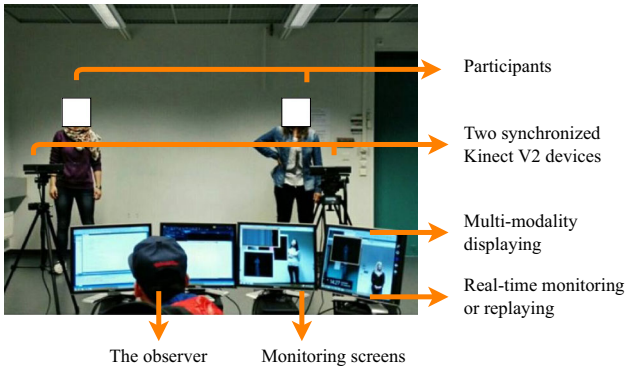


Fig. 3 Acquisition setup for the elicitation and recording of micro-gestures

3 The SMG Dataset

This section introduces the whole collecting procedure and details of the SMG dataset, from the psychological background, the elicitation design and the annotation to the final collected dataset and its statistics.

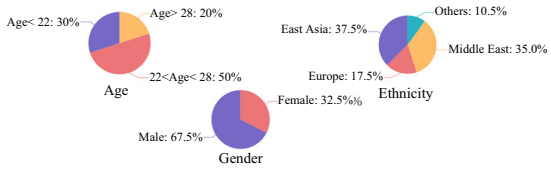
3.1 Psychological Background for Micro-gestures

The term "micro-gesture" was first used in the psychological work (Serge 1995) for assisting doctors in diagnosing patients' mental conditions via body gestures and later could also be found in popular science works (Kuhnke 2009; Navarro and Karlins 2008). The first work formally studying spontaneous gestures for hidden emotion understanding can trace back to Ekman (2004) where they found that spontaneous body gestures (e.g., "*a fragment of a shrugging gesture*"), together with micro-expressions, are more reliable clues to interpret hidden human emotions than intentionally performed facial expressions. Furthermore, the *fight*, *flight*, and *freeze* system proposed by Gray (1982) mediates reactions to aversive stimuli and threats, which reasons those spontaneous body movements from the aspect of brain sci-

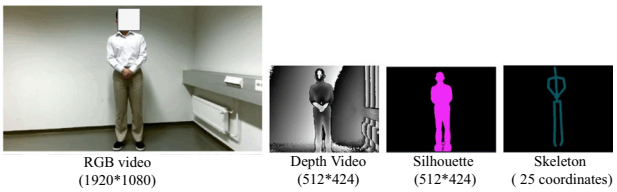
ence. The three factors, *fight*, *flight*, and *freeze*, can cause specific human behaviors at the onset of certain stimuli, including the freezing body (e.g., holding the breath), distancing behaviors (e.g., putting hands or objects to block faces or bodies) and guarding behaviors (e.g., puffing out the chest). Besides, to transfer from discomfort to comfort states, human beings develop a natural reaction, so-called *pacifying* actions, that tries to suppress the negative feelings induced by the above three factors (Panksepp 1998). Other psychological research related to MG can also be found in early work (de Becker 1997; Burgoon et al. 1994) and the most recent work (Kita et al. 2017; Pouw et al. 2016).

In total, based on the above psychology theoretical supports, we try to define the MG categories for computer vision study with criteria as (1) covering all MGs that could possibly occur on the SMG dataset, (2) corresponding to psychological theories and functions, and (3) being "properly specific" (e.g., "touching" would be too general, "scratching the left cheek" would be too specific) for a computational model to recognize. Finally, we summarized 16 types of MGs for our SMG dataset including *fight* patterns (e.g., "folding arms"), *flight* patterns (e.g., "moving legs"), *freeze* patterns (e.g., "turtling neck and shoulder"), and *pacifying* patterns (e.g., "scratching head and hand rubbing"). Non-micro gestures were also labeled as an independent category for illustrative gestures or sign gestures. The entire list of MGs and non-MGs that we collected and their psychological attributes are provided in Fig. 4a and Table 1. The 16 categories could cover the most common MGs on the SMG dataset but there might be some rare cases that were not observed in the current experimental scenario of the SMG dataset. We will further enrich the MG collection and make more comprehensive lists in future work.

(a) A list of the micro-gestures in the SMG dataset



(b) The distribution of participants' demographic



(c) Sample frames of the four different modalities

Fig. 4 Overview of MGs labeled in our SMG dataset. **a** Examples of annotated MGs and non-micro-gestures. For privacy concerns, we mask the faces of the participants here. **b** The distribution of participants' demographic. **c** The four modalities collected in our SMG dataset

3.2 Elicitation of Micro-gestures

Referring to the above supporting psychological theories, we design the procedure for the elicitation of MGs to create our SMG dataset as follows.

Eliciting Tasks. We designed two proxy tasks for stimulating the corresponding emotional stress states and eliciting micro-gestures. Precisely, the two proxy tasks are (i) given a true story with a title and detailed content, repeating the content of the story, as the "baseline stimuli", and (ii) given an empty story with only a title and no content, making up a fake story off-the-cuff, as the "deviation stimuli". The stories are short newscasts or reports with an average of 141 words and with rich details (see more detailed design principles in "Appendix B"). Participants have to repeat (baseline stimuli) or make up (deviation stimuli) the content of the story, and they need to prove that they knew the story content, respectively, no matter which task they were assigned. Participants were told that there would be a punishment for them if they got caught, so they had to conceal their emotions, especially for the "deviation stimuli" ones. Compared to repeating a true story (baseline stimuli, which can be regarded as the counterpart of the placebo group in the psychological field), creating a fake story off-the-cuff (deviation stimuli) needs higher mental-load requirements and more inner activities with mental presence and emotional involvement (Palena et al. 2018). In this way, the two emotional stress states are obtained for our SMG dataset as the hidden emotional states. For ease of the reader, we denote the two states as NES (non-stressed emotional state) and SES (stressed emotional state) for short.

Participants. In total, 40 participants were recruited for our dataset collection (age: M = 25.8, SD = 4.87). They are 27 men and 13 women from multicultural backgrounds (16 countries). The distribution of participants' demographic is given in Fig. 4b. They were recruited via advertisements, and no specific educational major was restricted. Although some of the participants were familiar with machine learning and computer vision, none of the participants were privy to the

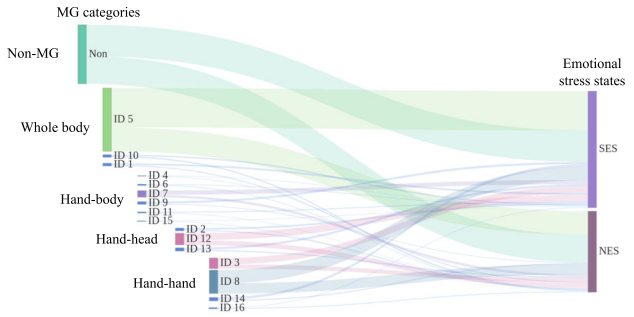


Fig. 5 Visualized distribution of MGs among different emotional stress states. The size of the blocks stands for the amount of the MGs. We can observe that, MGs are rare and fine-grained compared to ordinary gestures. There are multiple types of fine-grained MGs under each coarse category. MGs can be easily submerged by non-MGs

workings of the machine learning algorithm of the study we were conducting.

Apparatus. Two Kinect V2 sensors were placed two meters away in front of the participants to capture their whole body movements, with the RGB resolution of 1920 × 1080 pixels at 28 frames per second. The resulting modalities are RGB, silhouette, depth videos, and skeleton coordinates, as shown in Fig. 4c. The smoothing function of Kinect V2 was disabled to obtain detailed and subtle body movements as much as possible due to the particularity of MGs.

Procedure. The data collection was carried out in a normal office room of a college, as shown in Fig. 3. Two participants took turns telling stories, and an observer monitored them behind the scenes to ensure that participants felt the need to conceal their true emotions. For one round experiment, two participants were assigned two different (SES/NES) stimuli, respectively, and they needed to persuade the observer to believe that they knew the content. We ensure that the round numbers of NES and SES collected from each participant is the same. In other words, the numbers of NES and SES instances are evenly distributed in the SMG dataset. The time duration of one complete round is controlled in six minutes.

3.3 Data Annotation and Quality Control

Our SMG dataset's annotation contains two levels: (1) the temporal allocation and MG categories and (2) the emotional state categories.

MG Labeling. Four human annotators were assigned to go through long video sequences to spot and annotate all the 16 categories of MGs (as well as all the non-MGs). To guarantee the quality of annotation, we arranged two rounds of labeling. In the first round, the four annotators were trained on how to spot and classify MGs based on the MG list (see Table 1) and related psychological theories. After confirming the labeling criteria, they annotated the MGs separately based on their judgments of the collected video sequences. The MG category labels of the four annotators were summarized and cross-checked, and majority voting decided inconsistent cases. In the second round, the temporal labeling of MG clips was cross-checked to ensure that the labeling style of the start and endpoints of the MGs are unified at the frame level. Finally, we have all the MGs clips with start-, end-points, and their categories among the collected long video sequences.

Emotional Stress State Annotation. The emotional stress states in our SMG dataset are straightforward and objective based on the two proxy tasks, i.e., NES and SES are naturally assigned based on the corresponding task.

3.4 SMG Dataset Statistics

Dataset Structure. The final SMG dataset comprises 414 long video instances (around one minute for each instance) from 40 participants, resulting in 821,056 frames in total (more than 488 min). Each long video instance has one of the two emotional states (NES or SES). The video instances are evenly distributed in the two emotional states (207 v.s. 207). Among those 414 long-video instances, 3712 MG clips were labeled out, and the average length of those MGs is 51.3 frames (with the shortest MG as 8 frames), which is significantly shorter than the length of common gestures collected in other datasets like 100–300 frames (Escalera et al. 2015; Li et al. 2016). The distribution of MGs in the two emotional stress states can be seen in Fig. 5.

Correlations of MGs and Emotional States. We validate the MG distributions in the two emotional stress states using the t-test after a placebo-controlled study. The detailed statistical results are given in Table 2 as a quantitative report. Specifically, we deploy the paired sample t-test, using T-distribution (two-tailed) to compare MG distributions over the two emotional stress states among the 40 participants. From the last line of Table 2, we can see that, there was a significant increase in the volume of MGs performed under SES (M = 38.58, SD = 32.5095) compared to NES (M = 24.50, SD = 20.8671), $t(39) = 4.6300$, $p < 0.0001$. The result rejected the null hypothesis, thus a significant correlation between MGs and emotional stress states is found. When it comes to non-MGs, it shows that no significant changes are found with $t(39) = 0.9198$, $p < 0.3633$.

Visualized MG Distributions. We present a visualized distribution of MGs on the two emotional stress states as shown in Fig. 5. We observe certain features of MG patterns. The first and most prominent feature is that non-MGs and whole-body MGs occupy the majority of the body movements, demonstrating it is challenging to efficiently distinguish rare MGs from unconstrained upcoming streams as they can be easily submerged among the dominating amount of non-MGs. Secondly, although MGs cover a large range of body gestures, the major categories are extremely fine-grained: six kinds of MGs in "hand-body" interactions and four in

Table 1 The list of MGs collected in the SMG dataset, and MG IDs correspond to the indexes in Fig. 4

MG ID	Kinematic description	Psychological attribute	Number in SES/NES	MG ID	Kinematic description	Psychological attribute	Number in SES/NES
1	Turtling neck and shoulder	Freezing	35/19	10	Crossing legs	Pacifying	28/23
2	Rubbing eyes and forehead	Freezing	32/19	11	Scratching some part of body	Pacifying	15/8
3	Folding arms	Fighting	122/77	12	Scratching or touching facial parts other than eyes	Pacifying	131/88
4	Touching or covering suprasternal notch	Pacifying	6/5	13	Playing or adjusting hair	Pacifying	41/21
5	Moving legs	Fleeing	742/447	14	Holding arms	Fighting	46/22
6	Touching or scratching neck	Pacifying	13/11	15	Pulling shirt collar	Pacifying	6/5
7	Folding arms behind body	Freezing	81/44	16	Playing with jewelry, and manipulating other objects	Pacifying	10/13
8	Rubbing hands and crossing finger	Pacifying	248/192	Non-MGs	Illustrative hand gestures	–	593/518
9	Arms akimbo	Fighting	38/13	Total	–	–	2187/1525

Psychological attribute stands for the psychological basis that drives the corresponding micro-gesture

Table 2 The statistical distribution of gestures over the two states of emotional stress

MG type	Emotional stress state						*t* value	*p* value
	NES			SES				
	S	M	SD	S	M	SD		
Non-MGs	518	12.95	12.61	589	14.72	13.81	0.9198	0.3633
MGs	980	24.50	20.87	1543	38.58	32.51	4.6300	<**0.0001**

S, M, and SD stand for "sum", "mean" and "standard deviation". The significance level equals to 0.01 and the significant terms are marked in bold. Note that the total number of gestures is not equal to 3712, because some gestures are in the transitions between two emotional stress states, which are not counted

"hand-hand" interactions. Thus, compared to other body gesture/action recognition tasks, MGs require a more fine-grained and accurate recognizing ability from the machine learning aspect.

Relationship Between MGs and Subjects. As mentioned, the use of body gestures to interpret emotions could be heavily affected by individual differences. Here, we conduct a qualitative analysis of MG patterns of different subjects. Specifically, *Pearson*'s correlation coefficient is used to measure the correlation of different MG performing patterns from 40 subjects in our SMG dataset. The MG performing pattern is presented by the frequency distribution of 17 MGs of a given subject. *Pearson*'s correlation coefficient varies from -1 to 1, and the higher it is, the stronger the evaluated correlation is. According to the statistic calculation, the average *Pearson*'s correlation coefficient of these 40 subjects is 0.456, with the highest one of 0.966 and the lowest one of -0.240. It indicates a trend that subjects share MGs patterns, especially in the exposing frequency of MGs, while individual inconsistency of the MG patterns is still not negligible. As a result, although the above *t*-test proves the effectiveness of SES for eliciting MGs, it is necessary to emphasize *the inconsistency of MG performing patterns brought by different subjects*.

4 Micro-gesture Classification

In this section, we focus on the task of classification of pre-segmented MG clips from our SMG dataset. Analogous to the classical action/gesture recognition task, algorithms need to classify a given sequential clip into the correct MG category, from a certain data modality, such as RGB, depth, optical flows, or body skeletons. In order to set up the benchmark of MG classification, we select over ten state-of-the-art models for the classical action recognition task from recent top venues like AAAI, ECCV, ICCV, CVPR, and TPAMI, and evaluate them on the SMG dataset, including two representative modalities as RGB and skeleton. We first report the evaluation protocols and introduce the models used for MG classification on the two modalities. At last, we present the experimental results and related analysis.

Table 3 MG classification performance on the test set of the SMG dataset

Method	Modality	Accuracy	
		Top-1	Top-5
ST-GCN (Yan et al. 2018)	Skeleton	41.48	86.07
2 S-GCN (Shi et al. 2019)		43.11	86.90
Shift-GCN (Cheng et al. 2020)		55.31	87.34
GCN-NAS (Peng et al. 2020)		58.85	85.08
MS-G3D (Liu et al. 2020)		**64.75**	**91.48**
R3D (Hara et al. 2018)	RGB	29.84	67.87
I3D (Carreira and Zisserman 2017)		35.08	85.90
C3D (Tran et al. 2015)		45.90	79.18
TSN (Wang et al. 2018)		50.49	82.13
TSM (Lin et al. 2019)		58.69	83.93
TRN (Xu et al. 2019a)		59.51	88.53
TSN* (Wang et al. 2018)		53.61	81.98
TRN* (Xu et al. 2019a)		60.00	91.97
TSM* (Lin et al. 2019)		**65.41**	**91.48**

In total 11 state-of-the-art models for RGB and skeleton modalities are reported. Methods with stars are pretrained on large-scale datasets Methods with the best performance are marked in bold

4.1 MG Classification Benchmark Setup

We propose the benchmark of classification MGs on SMG dataset with two modalities. Given 3712 pre-segmented MGs clips with their labels, the task is to achieve accurate classification among 16 MG classes and non-MG classes. We implement a cross-subject protocol that the 2470+632 MG clips from 30+5 subjects are used for training+validating, and 610 clips from the remaining five subjects are used for testing. The overall accuracy on the testing set is reported as results. Eleven state-of-the-art models are provided for this task, including RGB and skeleton modalities.

RBG-based MG Classification. For RGB modality-based gesture classification, we adopt six state-of-the-art models which are well known in the action recognition research field. Those models can be sorted into two groups. The first group is 2D CNNs based models that capture the temporal information from features learned via 2D CNNs, including Temporal segment networks (TSN) (Wang et al. 2018), Temporal shift

module (TSM) (Lin et al. 2019) and Temporal Relation Networks (TRN) (Xu et al. 2019a). The second group is the 3DCNN family that directly learns the temporal information from features learned through 3D CNNs, including 3DCNNS (C3D) (Tran et al. 2015), 3D ResNets (R3D) (Hara et al. 2018), Inflated 3D ConvNet (I3D) (Carreira and Zisserman 2017).

Skeleton-based MG Classification. For MG classification with skeleton modality, Graph Convolutional Networks (GCNs) are the main stream architectures to deal with skeleton joint data. Here, we implement five recent graph convolutional-based methods that all achieved state-of-the-art performance on large-scale action datasets, like NTU (Shahroudy et al. 2016) and Kinetic (Carreira and Zisserman 2017). The models include Spatial Temporal GCN (STGCN) (Yan et al. 2018), Two-Stream Adaptive GCN (2 S-AGCN) (Shi et al. 2019), Shift-GCN (Shift-GCN) (Cheng et al. 2020), GCNs with Neural Architecture Search (GCN-NAS) (Peng et al. 2020) and Multi-scale Unified Spatial-temporal GCN (MS-G3D) (Liu et al. 2020).

4.2 Evaluation Results

The experimental results are given in Table 3. As shown in Table 3, we can observe that MS-G3D (Liu et al. 2020) achieves the best performance (top-1 64.75%, top-5 91.48%) than RGB modality based models (with best model TRN (Xu et al. 2019a) of top-1 59.51%, top-5 88.53%) and generally skeleton-modality based methods outperform RGB-modality based methods. Possible reasons include (1) compared to the RGB modality, skeleton data collected from Kinect contains more detailed and accurate depth information. This is critical for distinguishing subtle differences of MGs such as "touching or covering suprasternal notch" and "illustrative hand gestures", (2) GCN-based models with a compact network structure and efficient skeleton-based representations can prevent overfitting issue thus does not severely reply on a large number of training samples as 3DCNN based models. This overfitting problem can be spotted also on R3D (Hara et al. 2018) (top-1 29.84%, top-5 67.87%) and I3D (Carreira and Zisserman 2017) (top-1 35.08%, top-5 85.90%) which might need pre-training on large-scale datasets. Thus, we further conducted extra experiments that explore the impact of pretrained training strategy on the performances of those models by selecting several representative models, including TSN, TRN, and TSM. The results are presented with methods marked with stars in Table 4.1. From the results we can observe that, after initializing the model with weight trained on an action recognition dataset, the performances indeed can increase to some extent. Lastly, we can see that even though the Top-5 accuracy of the MG classification can reach 90%, the Top-1 accuracy of all methods is still below 66%. As shown, our SMG dataset is challenging, especially for inter-class and long-tail issue handling.

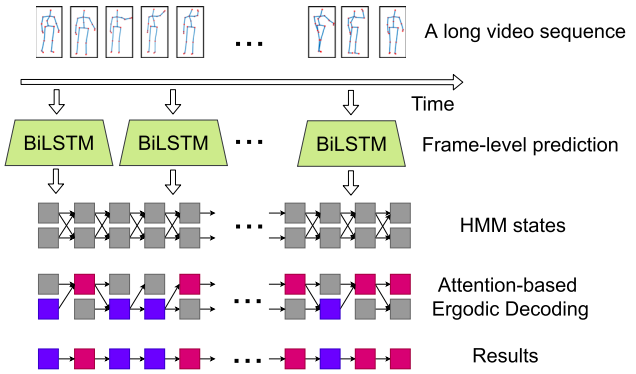


Fig. 6 The HMM model for online recognizing MGs. Our method can adaptively conduct the HMM decoding with a parameter-free attention mechanism

5 Online Micro-gesture Recognition

In this section, we take one step further by providing the benchmark of online MG recognition, i.e., processing raw, unsegmented sequences containing multiple body gestures, including MGs and non-MGs, on the SMG dataset. First, we discuss the specific challenges of online MG recognition. Then, we propose a novel HMM-DNN network for the task with parameter-free attention mechanism. At last, the evaluation metrics, together with the evaluating results of various methods on three online gesture recognition datasets, are presented.

5.1 Challenges of Online MG Recognition

The online recognition of MG has two parallel sub-tasks: detecting the potential body gestures from upcoming frames and classifying the ongoing body gestures into corresponding MG categories. However, some challenges make online recognition of MG different from other ordinary gestures. First, although some existing methods (Liu et al. 2018; Wu et al. 2016; Xu et al. 2019b) can achieve the detection and classification of actions/gestures, they all need various redundant post-processing procedures to optimize the predictions, which is not practical for online detection task. Meanwhile, it is proven that sequential aligning models such as HMM and Connectionist Temporal Classification (CTC) can provide transition priors to reason and enhance predictions from neural networks (Kuehne et al. 2019; Richard et al. 2018), which enable the online recognition of gesture/action to be more robust and accurate. However, we argue that in the dataset with spontaneous MGs, like SMG, the prior learned by sequential aligning models from training sets could be biased, and lead to inferior recognizing results to some extent. For instance, there could be a lot of "rubbing hands"

after "touching nose" in training subjects, while the testing subjects could perform no "rubbing hands" at all. Second, the "non-movement" interval, the so-called Ergodic state, was introduced in most of the previous works (Neverova et al. 2016; Wu et al. 2016) to achieve accurate allocation and segmentation of gestures. Meanwhile, MGs usually occur continuously without any "non-movement" intervals and sometimes can be incompletely performed. Therefore, a more flexible and efficient transition scheme is needed. Lastly and most importantly, MGs are *rare* and subtle. How to boost the HMM decoding escaping from local optimal brought by the dominating amount of the Ergodic states (irrelevant/noisy body movements) and non-MGs, is exceptionally challenging.

5.2 A Parameter-Free Ergodic-Attention HMM Network for Online MG Recognition

Mathematical Framework. We chose the sequences of the 3D skeletal stream as inputs because its lower dimensionality is suitable for online processing tasks and the reliable performance shown in the MG classification task. Similarly to the work of Chen et al. (2020), we model the local temporal dynamics with an attention-based Long Short-Term Memory (BiLSTM) network (giving an initial prediction of the current frame) and use an HMM model to enhance inference reasoning (finalizing the prediction of the current frame with priors in the past frames), shown in Fig. 6. The full probability of the is specified as follows:

$$\begin{aligned} &p(x_1, x_2, \ldots, x_T, h_1, h_2, \ldots, h_T) \\ &\quad = p(h_1)p(x_1|h_1)\prod_{t=2}^{T} p(x_t|h_t)p(h_t|h_{t-1}), \end{aligned} \tag{1}$$

where T is the total length of the sequence, $p(h)$ and $p(x)$ stand for the probabilities of hidden state and observed states, respectively. $p(h_t|h_{t-1})$ is the transition matrix to reason for the alignment on the long sequence. The emission probability $p(x_t|h_t)$ can be expended as:

$$p(x_t|h_t) = w(h_t|x_t)p(x_t)/p(h_t), \tag{2}$$

where $p(h_t)$ is the prior probability of hidden states that corrects the prediction when the classes are imbalanced (we argue this raw prior is biased and insufficient, see next section). $p(x_t)$ is a constant value that does not depend on the class. At last, $w(h_t|x_t)$ is the posterior probability which is estimated by a trained BiLSTM network:

$$w(h_t|x_t) = \begin{bmatrix} p_1 \\ p_2 \\ \cdots \\ p_{M+1} \end{bmatrix} = \begin{bmatrix} W_{1:M} \\ W_{M+1} \end{bmatrix}, M = N \times C, \tag{3}$$

where C is the total number of gesture classes (as 17 in practice, including MGs and non-MGs), N is the HMM state number used to present one gesture (set as 5 for the best performance) and M is the resulting total HMM state number (85 in practice). We take an additional HMM state $M+1$ (86 in practice) as the "non-movement" state. Then, $W_{1:M}$ and W_{M+1} stands for the probability distribution of the HMM states of all the gestures (MGs and non-MGs) and the "non-movement" state, respectively.

A Novel Parameter-free Attention Mechanism for Ergodic HMM Decoding. Based on the above HMM full probability, we find that although the prior $p(h_t)$ is to correct the data imbalance, this prior is not strong enough or even harmful. The MGs are still submerged in the dominating noisy/irrelevant body movements. Thus, we propose a novel method to address this issue, called Attention-Based Ergodic Decoding (AED), that has a parameter-free self-attention mechanism to modeling the HMM alignment. It has two folds of improvements based on the conventional HMM framework (Wu et al. 2016): an attention mechanism on $W_{1:M}$ to lift the probability of meaningful gestures, and an inhibition on W_{M+1} for the probability of noisy body movements. Specifically, we exploit the AED by replacing $w(h_t|x_t)/p(h_t)$ with a new form of posterior probability $w'(h_t|x_t)$ that have a more effective prior ability:

$$\begin{aligned} w'(h_t|x_t) &= \begin{bmatrix} W'_{1:M} \\ W'_{M+1} \end{bmatrix} \\ &= \begin{bmatrix} \mu \cdot softmax(W_{1:M} \odot W_{1:M}) \odot W_{1:M} + W_{1:M} \\ W^{\lambda}_{M+1} \end{bmatrix}. \end{aligned} \tag{4}$$

For the top part $W'_{1:M}$ in the formula, we obtain it by calculating the self-attention map from the Hadamard product $\odot$ of $W_{1:M}$ itself, weighing the softmax result of this attention map with a scale parameter μ, and then performing an element-wise sum operation with the original distribution $W_{1:M}$ to obtain the updated distribution $W'_{1:M}$. For the bottom part W'_{M+1}, we suppress it by adding W_{M+1} to the λth power. We do not use the dot product in the original attention version (Vaswani et al. 2017) because the Hadamard product has both the calculating efficiency and better performances under a non-parameter setting, while the dot product will lead to inferior results according to our experiments. In this way, we exploit the attention mechanism to the posterior probability and the problem of subject-dependent MG patterns were made possible.

Inference. After BiLSTM is trained to give an estimation of each upcoming frame with a SoftMax probability $w(h_t|x_t)$ of the HMM state, we can conduct the inference together with the learnt transition probability $p(h_t|h_{t-1})$. During the testing phase, we want to solve the decoding of hidden state sequence $\hat{g}$ to obtain the most likely explanation (namely, the gesture alignment), which is determined as:

$$\begin{aligned}\hat{g} &= arg\ \max_{h}\ p(x_1, x_2 \dots x_T, h_1, h_2 \dots h_T) \\ &\cong arg\ \max_{h} \pi_0 \prod_{t=2}^{T} w'(x_t|h_t) p(h_t|h_{t-1}),\end{aligned} \tag{5}$$

where π_0 stands for the constant value. By using Eq. 5, we can break down the problem of solving the utmost probability of a long non-stationary sequence into continuously solving HMM states probability with hidden states $h_{1:T}$. While the HMM states are aligned in real-time, the testing sequence can be inferred for both segmentation (non-movement) and recognition (MGs and non-MGs). Finally, we improved the method proposed by (Wu et al. 2016) by treating not only the "non-movement" state but also the middle HMM states of every gesture as ergodic states. In this way, the segmentation of several continuous incomplete gestures becomes possible.

The complete network structures and technical implementation details of our AED method, such as the value of μ and λ are given in "Appendix F". Note that $w'(h_t|x_t)$ is calculated based on $w(h_t|x_t)$ which is given by the BiLSTM output without any fine-tuning. Thus, our proposed attention scheme can be used directly in the testing phase without *extra-training* and *extra-parameters*, which is *parameter-free* and can be plugged into other existing models for online gesture recognition.

5.3 Evaluation on SMG Dataset

Evaluation Metrics. Following the protocols used in online action detection from the work of Li et al. (2016), we jointly evaluate the detection and classification performances of algorithms by using the F1 score measurement defined below:

$$F1_{score} = \frac{2Precision * Recall}{Precision + Recall}, \tag{6}$$

given a long video sequence that needs to be evaluated, $Precision$ is the fraction of correctly classified MGs among all gestures retrieved in the sequence by algorithms, while $Recall$ (or sensitivity) is the fraction of MGs that have been correctly retrieved over the total amount of annotated MGs.

Also, we define a criterion to determine a correct detection with the overlapping ratio α_{th} between the predicted gesture intervals and ground truth intervals. The overlapping ratio

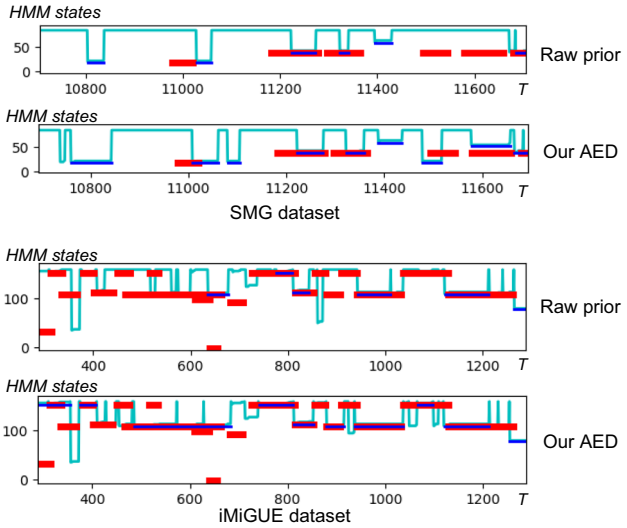


Fig. 7 Visualized HMM decoding of **failure cases**. We present the HMM decoding of sample sequence #36 in the SMG dataset and #72 in the iMiGUE dataset, using raw prior (top) and our AED (bottom). The x-axis represents time, and the y-axis represents the hidden states of all classes. The cyan lines represent the highest probability given by networks, while red lines denote the ground truth labels, and the blue lines are the predictions

α_{th} is defined as follows,

$$\alpha_{th} = \frac{|I_{gt} \cap I_{pred}|}{|I_{gt} \cup I_{pred}|}, \tag{7}$$

where I_{pred} and I_{gt} denote the predicted gesture and ground truth intervals, respectively. If α_{th} is greater than a threshold, we say that it is a correct detection. In practice, we set α_{th} to 0.3 as default (see ablation studies of different α_{th} values in "Appendix F").

Performances on the SMG Dataset. As a comparison to the MG online recognition performance of our HMM BiLSTM-AED, we also implemented four related methods as baselines: FCN-sliding window (Chen et al. 2019), DBN-HMM (Wu et al. 2016) and STABNet-MES (Chen et al. 2020). The results of online recognition of both our method and the baselines compared are shown in Table 4. Our method is considerably effective in recognizing continuous gestures in unconstrained long sequences (accuracy of 0.173, recall of 0.245, and F1 score of 0.203). Technical implementation details of all the compared methods are available in "Appendix E".

5.4 AED-BiLSTM on Other Datasets

We also evaluate our proposed AED-BiLSTM framework on two other existing online detection datasets, iMiGUE (Liu et al. 2021) and OAD (Li et al. 2016) to verify its generalizability.

Table 4 MG online recognition performances on the test sets of SMG and iMiGUE datasets

Online recognition method	SMG dataset			iMiGUE dataset		
	Accuracy	Recall	F1-score	Accuracy	Recall	F1-score
FCN-sliding window (Chen et al. 2019)	0.082	0.110	0.094	0.059	0.067	0.063
DBN-HMM (Wu et al. 2016)	0.128	0.167	0.145	–	–	–
STABNet (Chen et al. 2020)	**0.206**	0.164	0.182	0.137	0.082	0.103
AED-BiLSTM (Ours)	0.173	**0.245**	**0.203**	**0.166**	**0.177**	**0.171**

Methods with the best performance are marked in bold

iMiGUE Dataset is a newly published dataset that also focuses on involuntary micro-gestures occurring during the post-match interview of tennis players. There are 359 videos of post-match press conferences. The videos' duration varies with an average length of 350 s, and the total length is 2 092 min. A total of 18 499 MG samples were labeled out with the multi-label annotation, which means there could be multiple MGs labeled for one frame. It has more than 70 subjects that contain 32 categories of MGs with 25 joints estimated by OpenPose (Cao et al. 2019). We follow the same cross-subject protocol provided by Liu et al. (2021) that uses 255 long video sequences (with 13,936 MG samples) from 37 selected subjects for training and 104 sequences (with 4,563 MG samples) from the remaining 35 subjects for testing. We removed all the samples with null skeleton joints for the robust training for both compared methods and ours for a fair comparison.

OAD Dataset includes 10 daily human action categories. It was captured as long skeleton sequences with Kinect v2. The annotation of start and end frames are provided within peak duration (not a from-none-to-action pattern), similar to the work of Chen et al. (2020), we compensate 12 frames to the beginning of actions to learn pre-action information for better online recognition. "MovingPose" (Zanfir et al. 2013) is also adopted to generate features for each frame. There are more than 50 long sequences in total, and 30 of them are used for training, 20 for testing, and the remaining sequences are for processing speed validation. In the OAD dataset, we use the same protocol as Liu et al. (2018) that sets different observation ratios to validate the algorithm. Thus the accuracy is reported for this dataset.

Performance Discussion. The experimental results are presented in Tables 4 and 5. As shown, our AED-BiLSTM outperforms all other methods with significant margins (2.1% on SMG and 6.8% on iMiGUE) on the MG online recognition task. Our AED-BiLSTM brings a huge improvement, especially in the iMiGUE dataset, because the skeleton joints in this dataset are extracted from OpenPose, which are relatively noisy. By using our enhanced prior to suppressing those noisy body movements, the results are effectively improved. From Table 5, we can see that our AED-BiLSTM framework can also efficiently improve the performance of online recognition for regular gestures (88.1%, 93.4%, 94.2% in an observational ratio of 10%, 50%, 90%). As we can see, our method achieves superior results to StabNet on all metrics for SMG and iMiGUE datasets, except for accuracy on SMG where it is considerably lower than StabNet. Essentially, our AED module works as a regulation to suppress the non-MG and putting attention to MGs. It behaves as a tendency to weight MGs while neglect non-MGs, resulting in a higher recall of those MGs. The high recall will naturally lead to the situation that many non-MGs are suppressed and misclassified as MGs, resulting in a relatively low accuracy.

Table 5 The early online detection performances on the OAD dataset

Observational Ratio	10%	50%	90%
ST-LSTM (Liu et al. 2016)	60.0%	75.3%	77.5%
Attention Net (Liu et al. 2017)	59.0%	75.8%	78.3%
JCR-RNN (Li et al. 2016)	62.0%	77.3%	78.8%
SSNet (Liu et al. 2018)	65.6%	79.2%	81.6%
STABNet (Chen et al. 2020)	87.2%	92.0%	93.1%
AED-BiLSTM (ours)	**88.1%**	**93.4%**	**94.2%**

Results of accuracy are reported
Methods with the best performance are marked in bold

Failure Case Analysis. From Fig. 7, we visualize the HMM decoding path to analyze the failure cases. As we can see, online recognition of in-the-wild body gestures is challenging due to the complicated transition patterns between gestures and the high requirements for accurate temporal allocation. Even though, our AED method, with its attention mechanism, has a better correcting performance than the raw prior. For instance, around frames 11,500-11,600 of the SMG case, AED can help to escape from the false positive prediction of "non-movement" intervals and give potential MG predictions, while around frames 100–600 in the iMiGUE case, the AED can help to emphasize the true positive prediction of the MGs with self-attention. At last, the visualization of the attention maps is presented in Fig. 8, which also shows that our AED can effectively suppress the biased priors brought by certain classes (yellow color means lower probability) thus can better handle long-tail class distribution.

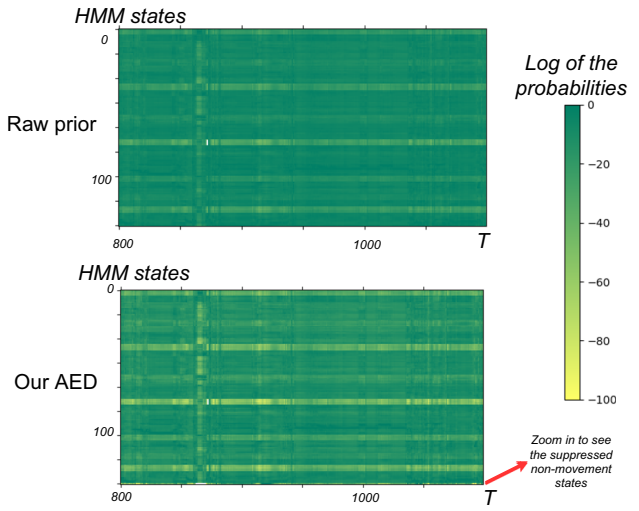


Fig. 8 Visualized **attention map**. We present the attention map of sample #72 in the iMiGUE dataset, using raw prior (top) and our AED (bottom). The x-axis represents time, and the y-axis represents the hidden states of all classes. The value of matrix is the probability given by networks, we take the log value for a better visualization and computational convenience in Viterbi decoding. The last line is the probability of non-movement state. The white spot on the matrix stands for the NaN value when taking the logarithm operation

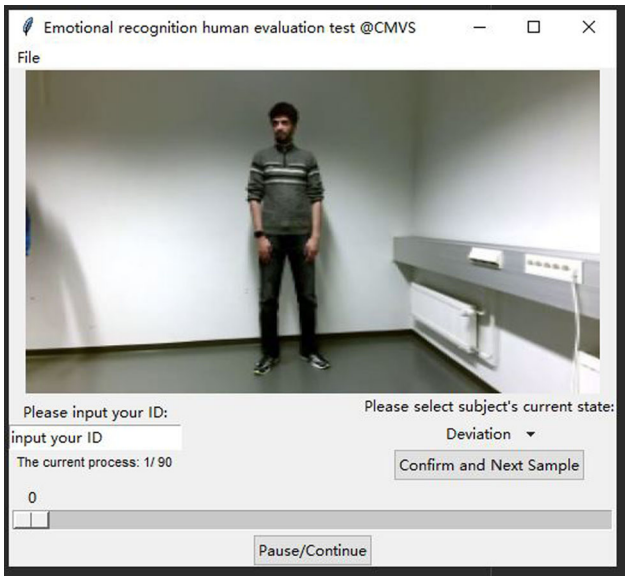


Fig. 9 The GUI of the human evaluation test for emotional state recognition. A screenshot of one sample is shown. For each video clip, evaluators are asked to go through the video and annotate the emotional state as a comparison of our methods

6 Body Gesture-Based Emotional Stress State Recognition

In this section, we conduct experiments on body gesture-based recognition of the emotional stress state. The task is defined as predicting the emotional stress state (i.e., SES or NES) within the context of the body gestures with a given video sequence. We first introduce the benchmark for evaluations by implementing several state-of-the-art models. Then, we present a new graph-based network for this task with a better performance compared to others.

6.1 Evaluation Protocols and Human Evaluation

Two Evaluation Protocols. As discussed in Sect. 2 and Sect. 3, subject differences could bring considerable influence to gesture-based emotion recognition. Thus, we define two types of evaluation protocols: subject-independent (SI) and semi-subject-independent (semi-SI). In SI evaluation, we use the same protocol as the classification and online recognition tasks that split the 40 subjects into 30+5 for training+validating (with 294 emotional state instances) and the remaining five for testing (with 90 instances). In semi-SI evaluation, we select 294 emotional state instances from all the 40 subjects for training+validating and the remaining 90 instances for testing. Each instance belongs to a specific emotional stress state (NES/SES). The emotional states of the instances are evenly distributed in the testing set, i.e., 45-45 for SES-NES. We report the emotional state recognition accuracy in percentage for each of these two protocols.

Human Evaluation. We assess the difficulty of the emotional state recognition task by enrolling human evaluators to observe the emotional instances and give their predictions. Sixteen ordinary college students with different academic majors were recruited as normal human evaluators. Another three university staff were trained to recognize MGs with related psychological backgrounds as expert evaluators. These evaluators were offered both skeleton/RGB videos to conduct the task (skeleton modality was always presented first and then the RGB modality to avoid any significant learning effect). The GUI of the human test is shown in Fig. 9 and the results are shown in Table 6.

The human evaluators were also interviewed after the evaluation test. Most of the testers claimed that it was tough only to use gestures (the skeleton modality) to infer, and it was a random guess. Meanwhile, for RGB videos, people tend to use multiple cues such as facial expressions and even overall impressions (e.g., if the subject looks confident) to determine the emotional stress states. We can also observe that trained evaluators perform better than ordinary people (accuracy of 0.75 for emotional stress states) as they know how to utilize MGs as clues to infer emotional states. As discussed above, MGs are often neglected by humans in interactions. Thus using body gestures for emotional state recognition, especially hidden ones, is a significantly challenging task.

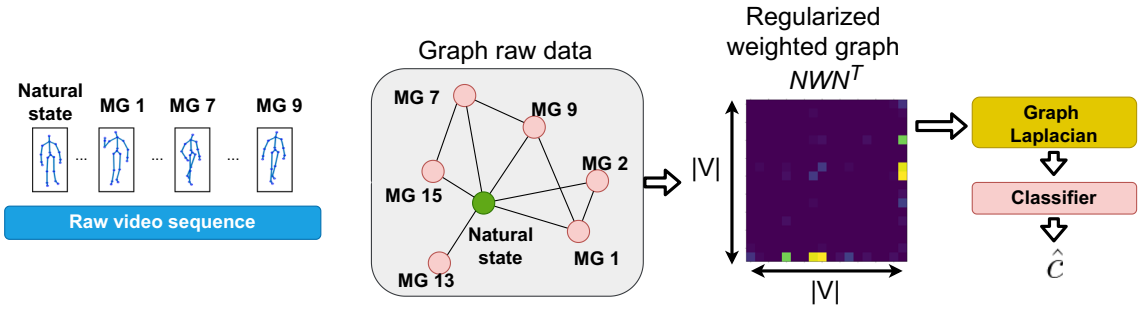


Fig. 10 Spectral decomposition of the graph network for emotional state recognition

Table 6 Body gesture-based emotional state recognition results of human evaluators

Human evaluator	Modality	Emotion state recognition accuracy
Random guess	–	0.50
Common people	Skeleton	0.48
	RGB	0.53
Trained evaluators	Skeleton	0.66
	RGB	0.75

6.2 Emotional State Recognition with State-of-the-Art Methods

As introduced in Sect. 3, instead of using a conventional paradigm that maps one gesture into one emotional status (Gu et al. 2013; Gunes and Piccardi 2006), we use two proxy tasks to present the emotional states. Thus, the task of emotional state recognition in SMG dataset is to predict the corresponding emotional state on a given long video sequence (the state of proxy task, NES/SES). Intuitively, there are two directions to approach this problem, one is raw context-based recognition that directly conducts the inference on the whole sequence and the other one is MG context-based recognition that predicts the emotional states based on the MGs on the sequences. Here we provide six machine learning-based methods for emotional state recognition, including both of these two kinds of methods.

Raw Context Recognition. Three state-of-the-art models for the skeleton-based action recognition task, ST-GCN (Yan et al. 2018), NAS-GCN (Peng et al. 2020) and MS-G3D (Liu et al. 2020) are provided as baselines that infer the emotional state based on the raw long instances. The input of the models is the full sequence of the body skeleton streams, which is to validate if the emotional patterns can be captured via body movements straightforwardly. The network structure is end-to-end whose hyper-parameters are the same for the task of MG classification mentioned in Sect. 4.1 aside from the output head dimensions (as NES/SES). The performances of the three baseline methods are presented in the "Sequence+NN" group of Table 7. As shown in the table, the three baseline methods (46%, 46%, and 50%) cannot even exceed the random selecting rate (50%). As expected, the inference based on raw video sequences involves many redundant, irrelevant body movements and easily fails to capture desired body movements (such as MGs) for emotional stress state recognition. Thus, conducting the recognition on long video sequences performs poorly (near random guessing) with existing state-of-the-art models.

MG-based Recognition. Unlike the above raw context recognition methods, we also present several MG-based methods for emotion understanding. A baseline strategy that uses the Bayesian network to encode the distribution vectors of MGs (with dimensions of $1 \times N$, N is the MG number) was provided in our previous work (Chen et al. 2019). It experimentally validated the contribution that MGs can bring to the emotion understanding context. In Table 7 bottom part ("MG+classifier" group), we can observe that micro-gesture is beneficial for the emotional state inference from the BayesianNet (0.59&0.66). Besides, we go one step further by encoding the MG relationships on a long sequence into graph representation (with dimensions of $N \times N$) so that the transitions of MGs are also involved with node relationships. Intuitively, this should bring more gains as the information of the feature increases, and we selected two state-of-the-art high-dimensional graph convolutional networks L2GCN, BGCN (You et al. 2020; Zhang et al. 2019) to verify it. However, as shown in Table 7, we find that for these two high-dimensional models, the emotional state performances (0.44&0.47 and 0.54&0.53) are not as competitive as the simple BayesianNet. Thus, in the next section, we try to tackle the issue and propose a customized graph network for better mining the potential of the graph-based representations.

6.3 A Weighted Spectral Graph Network for Emotional State Recognition

We find that existing graph representation learning methods all rely on high-dimensional weight parameters. Limited sample amount easily leads to over-fitting on these models (Scarselli et al. 2008) (e.g., in our cases, a graph with

Table 7 Body gesture-based emotional state recognition results of the proposed method and compared baselines

Methods	Framework	Emotion state recognition accuracy	
		Subject-independent	Semi subject-independent
Random guess	–	0.50	0.50
ST-GCN (Yan et al. 2018)	Sequence +NN	0.46	0.42
MS-G3D (Liu et al. 2020)		0.46	0.49
NAS-GCN (Peng et al. 2020)		0.50	0.52
MG+L2GCN (You et al. 2020)	MG +classifier	0.44	0.47
MG+BGCN (Zhang et al. 2019)		0.54	0.53
MG+BayesianNet (Chen et al. 2019)		0.59	0.66
MG+WSGN (ours)		**0.65**	**0.68**

Note that "MG" includes both MG and non-MG instances as the input feature
Methods with the best performance are marked in bold

only 17 nodes of MGs) as shown in Table 7. Meanwhile, classical spectral graph handling methods like the Laplacian operator (de Lara and Pineau 2018) are suitable for insufficient samples to get node "gradients" without the need for high-dimensional weights. Thus, we utilize the strength of classical Laplacian operator to obtain the measurements of the "gradients" of each node and extend it to the directed, weighted graph case to better fit the task. The whole framework is presented in Fig. 10.

We give the mathematical definition of a graph as $G = (V, E, W)$ to represent the relationship of MGs. With the MGs of number N as graph nodes $V = \{v_p | p = 1, \ldots, n\}$ and the transitions between MGs as graph edges $E = \{e_q | q = 1, \ldots, m\}$, the input is therefore the transition frequency vectors as the weights on the graph edges $W = \{w_{i,j} | i, j = 1, \ldots, n\}$, where $w_{i,j}$ is obtained by counting the transition number between MG i and j. In this way, we map the distribution of MGs into raw graph data with the dynamic transition patterns between MGs maintained by W. Specifically, to tackle the directed graph issue, consider the vertex space $\mathbb{R}^V$ with standard basis $\{e_1, \ldots, e_n\}$ and, a $n \times n$ matrix N can be defined as for $N = \{n_i = e_j - e_k | i = 1, \ldots, m$ and $j, k = 1, \ldots, n\}$. This matrix N is called the signed vertex edge incidence matrix of the original G (with respect to the fixed orientation). The key fact is that the Laplacian $\mathcal{L}$ of the G is the (transpose of the) Gram matrix of N, that is, $\mathcal{L}$=NN^T with which the directed graph can be deployed. Now recall that W is the weight matrix of G. Then we can define the Laplacian of G as the matrix product NWN^T where N is the signed vertex-edge incidence matrix of the underlying unweighted graph of G. In this way, the Laplacian operator can be exploited to extract "gradient" features from the MG graph representation. The resulting feature vectors from Laplacian operator are fed into the classifiers to predict the final emotional state $\hat{c}$. Eventually, the whole formulation of our proposed weighted spectral graph network (WSGN) is given as follows:

$$\hat{c} = f_{classifier}(\mathcal{L}(NWN^T)), \tag{8}$$

where for $f_{classifier}$, we experimented with different standard classifiers combined to our spectral embedding. That is, Multi-layer Perceptron with Relu non-linearity (MLP) (Rumelhart et al. 1986), k-nearest neighbors (kNN) (Fix and Hodges 1989), Random Forest (RF) (Ho 1995), and Adaptive Boosting (AdaBoost) (Schapire 2013).

6.4 Discussion and Limitations

The experimental results for emotional state recognition are shown in Table 7. In practice, **MLP** outperforms other classifiers, which is reported in Table 7 as a result of our proposed WSGN. The detailed experimental settings can be found in the "Appendix G". Besides, extra experimental results (see "Appendix H") show that taking natural states (the non-movement snippets) into account as an extra MG in the transition representation will bring an improvement to the results, as well as the Laplacian operation. In the last line of Table 7, we can observe that our proposed WSGNN model outperforms all the compared methods, which further verifies that the MG-based analysis is beneficial to the final emotion understanding. By comparing the performances of MG+classifier frameworks and Sequence+NN frameworks, we can observe that the MG-based feature vectors are more beneficial to the present emotional states. This proves that MG-based analysis, with its effective representation capability of emotional state, can be a better option for emotional understanding. We believe that this can bring inspiration and new paradigms to the community over bodily emotion understanding.

The limitation of this experiment could be that the stakes of the subjects' emotional states were relatively low. Thus, this might decrease the distinction between baselines and deviations. Additionally, the sample size was relatively limited. Therefore, more research should explore how the similarity scoring system performs when more extensive samples are used.

7 Conclusions and Future Work

We proposed a novel, psychology-based and reliable paradigm for body gesture-based emotion understanding with computer vision methods. To our knowledge, our effort is the first to interpret hidden emotion states via MGs, with both quantitative investigations of human body behaviors and machine vision technologies. A related spontaneous micro-gesture dataset towards hidden emotion understanding is collected. A comprehensive static analysis is performed with significant findings for MGs and emotional body gestures. Benchmarks for MG classification, MG online recognition, and body gesture-based emotional stress state recognition are provided with state-of-the-art models. Our proposed AED-BiLSTM framework can efficiently provide a more robust correction to the prior with a parameter-free mechanism. Experiments show that AED-BiLSTM can efficiently improve online recognition performance in a practice closer to a real-world setting. Moreover, a graph-based network is proposed for the MG pattern representations to better analyze the emotional states.

This work involves and bridges the interdisciplinary efforts of psychology, affective computing, computer vision, machine learning, etc. We wish to break the fixed research paradigm of emotional body gestures which is limited to classical expressive emotions and argue for more diverse research angles for emotional understanding. Thus, we propose our spontaneous micro-gestures for hidden emotion understanding. We believe that the SMG dataset and proposed methods could inspire new algorithms for the MG recognition tasks from the machine learning aspect, such as combining more non-verbal cues such as facial expressions with MGs using the RGB modality in the SMG dataset to improve emotional recognition performance. The work can also facilitate new advances in the emotion AI field and inspire new paradigms for analyzing human emotions with computer vision methods. The community can be benefited from MGs with significant application potential in many fields, e.g., using machines to automatically detect MGs to enhance people's communicative skills, or assist experts in conducting Alzheimer's and autism disease diagnoses.

Acknowledgements This work was supported by the Academy of Finland for Academy Professor project EmotionAI (Grants 336116, 345122), project MiGA (grant 316765), the University of Oulu & The Academy of Finland Profi 7 (grant 352788), Postdoc project 6+E (Grant 323287) and ICT 2023 project (grant 328115), and by Ministry of Education and Culture of Finland for AI forum project. As well, the authors wish to acknowledge CSC - IT Center for Science, Finland, for computational resources.

Funding Open Access funding provided by University of Oulu including Oulu University Hospital.

Appendix A SMG Evaluation Protocols

In the proposed SMG dataset, the criteria of the three benchmarks (MG classification, MG online recognition, and emotional state recognition) are provided. Specifically, for the MG classification and online recognition tasks, we utilized the subject-independent evaluation protocol, while for the emotional state recognition task, both subject-independent and -dependent evaluation protocols are used.

Subject-independent protocol. In this protocol, we divide the 40 subjects into a training group of 30 subjects, a validating group of 5 subjects, and a testing group of 5 subjects.

The subject IDs of training and testing are:

Training set: {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30};

Validating set: {31, 32, 33, 34, 35};

Testing set: {36, 37, 38, 39, 40}.

Under this protocol, MG classification task has 2417 MG clip samples for training, 632 for validating and 593 for testing (each for around 50 frames); MG online recognition task has 30 long MG sequences for training, five for validating and five for testing (each for around 25,000 frames); and emotional state recognition task has 294 videos (i.e., emotional state instances) for training and 60 for validating and 60 for testing (each for around 8000 frames), respectively.

Semi-subject-independent Protocol. In this protocol, we selected 294 + 60 videos (147 + 30 SES and 147 + 30 NES instances) from all the 40 subjects as the training + validating sets, and the remaining 60 videos (30 SES and 30 SES instances) as the testing set. The participants' emotional states (SES/NES) are recognized via analysis of micro-gestures.

The video IDs of training and testing are:

Training set: {0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 96, 97, 98, 99, 100, 101, 108, 109, 110, 111, 112, 113, 120, 121, 122, 123, 124, 125, 132, 133, 134, 135, 136, 137, 156, 157, 158, 159, 160, 161, 168, 169, 170, 171, 172, 173, 180, 181, 182, 183, 184, 185, 192, 193, 194,

195, 196, 197, 204, 205, 206, 207, 208, 209, 216, 217, 218, 219, 220, 221, 228, 229, 230, 231, 232, 233, 237, 238, 239, 243, 244, 245, 249, 250, 251, 255, 256, 257, 258, 259, 260, 261, 262, 263, 264, 265, 266, 267, 268, 269, 270, 271, 272, 273, 274, 275, 276, 277, 278, 279, 280, 281, 282, 283, 284, 285, 286, 287, 288, 289, 290, 291, 292, 293, 294, 295, 296, 297, 298, 299, 300, 301, 302, 303, 304, 305, 306, 307, 308, 309, 310, 311, 312, 313, 314, 315, 316, 317, 318, 319, 320, 321, 322, 323, 324, 325, 326, 327, 328, 329, 330, 331, 332, 333, 334, 335, 336, 337, 338, 339, 340, 341, 342, 343, 344, 345, 346, 347, 348, 349, 350, 351, 352, 353, 354, 355, 356, 357, 358, 359, 360, 361, 362, 363, 364, 365, 366, 367, 368, 369, 370, 371, 372, 373, 374, 375, 376, 377, 378, 379, 380, 381, 382, 383, 384, 385, 386, 387, 388, 389, 390, 391, 392, 393, 394, 395, 396, 397, 398, 399, 400, 401, 402, 403, 404, 405, 406, 407, 408, 409, 410, 411, 412, 413, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 162, 163, 164, 165, 166, 167, 174, 175, 176, 177, 178, 179, 186, 187, 188, 189, 190, 191, 198, 199, 200, 201, 202, 203, 210, 211, 212, 213, 214, 215, 222, 223, 224, 225, 226, 227, 234, 235, 236, 240, 241, 242, 246, 247, 248, 252, 253, 254};

Validating set: {144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 162, 163, 164, 165, 166, 167, 174, 175, 176, 177, 178, 179, 186, 187, 188, 189, 190, 191, 198, 199, 200, 201, 202, 203, 210, 211, 212, 213, 214, 215, 222, 223, 224, 225, 226, 227, 234, 235, 236, 240, 241, 242, 246, 247, 248, 252, 253, 254};

Testing set: {84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 102, 103, 104, 105, 106, 107, 114, 115, 116, 117, 118, 119, 126, 127, 128, 129, 130, 131, 138, 139, 140, 141, 142, 143, 150, 151, 152, 153, 154, 155, 162, 163, 164, 165, 166, 167, 174, 175, 176, 177, 178, 179, 186, 187, 188, 189, 190, 191}.

Appendix B Materials for Stress Emotional States

We set up two proxy tasks for stimulating the emotional stress states and eliciting micro-gestures based on the findings in related research: (1) only a comparable truth condition ("baseline stimuli", and "deviation stimuli") rather than casual small talk can induce emotional difference between truth tellers and story makers (Palena et al. 2018) and (2) complications of the truth (more details of the story) can help to differ the truth tellers and story makers (Vrij et al. 2018, 2020). Thus, we selected five short reports and newscasts with full details as the materials. The five stories are "world's largest swimming pool" (119 words), "world's longest hair" (170 words), "world's biggest dog" (155 words), "world's hottest chilli" (133 words) and "world's largest pizza" (129 words). Excerpts from the story (world's longest hair) are as follows: "…She washes the hair once a week, using up to six bottles of shampoo at a time. Then it takes two days for the hair to dry – and they weigh 25 pounds when wet. She says that the extra weight of her hair makes her doctors very concerned. They seem to think that she has a curvature of her spine due to the length and weight of her hair." Before the experiment, participants were told that if they got caught they would have a punishment, i.e., to fill in a long questionnaire which contains more than 500 questions, so they had to try their best when making up a story (deviation stimuli), pretend to be telling/reading a given one (baseline stimuli). The long questionnaire works as the 'punishment' or a 'pressure', aiming to stimuli and elicit the emotional states and micro-gestures, and there was no actual punishment conducted after the data collection.

Appendix C Relationship Between MGs and Subjects

We visualize the Pearson's correlation coefficient of the MG performing patterns from 40 subjects in our SMG dataset as shown in Fig. 11.

Appendix D Experimental Settings for MG Classification on SMG

In the practical implementation of RGB modality based baselines, we trained all the models on SMG dataset with the same protocol as: 120 epochs are trained on the TSN, TRN, and TSM models; the batch size is set to 64 for TSN and TRN, and set to 32 for TSM; the base learning rate is set to 0.001 for all three models, and the learning rate is scaled with a factor

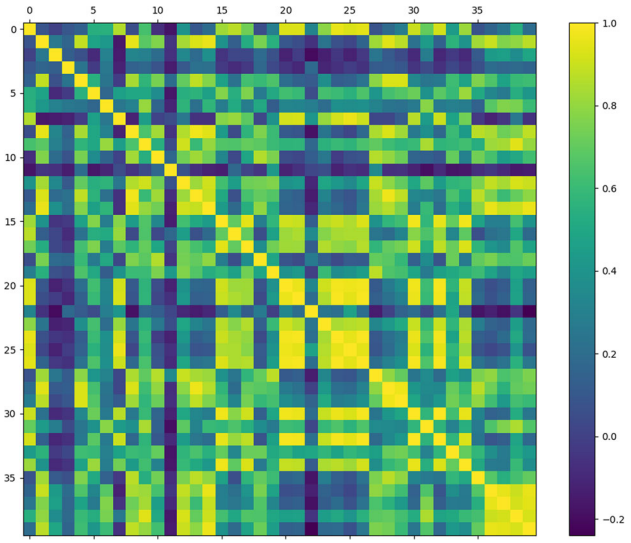


Fig. 11 The correlation distribution of MG patterns between subject pairs in our SMG dataset. The correlation factor is calculated by Pearson's correlation coefficient based on MG distribution of 40 subject pairs. We can see that, the MG performing patterns can vary a lot over some subject pairs

Table 8 The ablation study of AED-BiLSTM under different values of λ on SMG dataset. μ is fixed as 0.0

λ	Raw prior	0.7	1.0	2.0	2.1	2.2	3.0	4.0
F1-score	0.1825	0.1461	0.1958	0.1986	**0.2020**	0.2006	0.1823	0.1749

⋆We show the most representative values that cause F1-score to change considerably
Methods with the best performance are marked in bold

Table 9 The ablation study of AED-BiLSTM under different values of μ on SMG dataset. λ is set as 2.1 to obtain the best performance

μ	Raw prior	−5.0	−4.0	−2.5	−0.5	0.0	1.0
F1-score	0.1825	0.2006	0.2026	**0.2030**	0.2023	0.2020	0.2017

⋆We show the most representative values that cause F1-score to change considerably
Methods with the best performance are marked in bold

of 0.1 at epoch 50 and 100, respectively. For C3D, R3D, and I3D, 60 epochs are trained for each model; the batch size is set to 128 for C3D and R3D, 48 for I3D; and the learning rate is set to 0.0002 for all models. The optimizer is SGD which is consistent to the settings of all the models. The loss function is Categorical Cross Entropy. The training platform was Pytorch (Paszke et al. 2019) with a single GPU: NVidia Titan (24 GB).

In the practical implementation of skeleton modality based baselines, we trained all the models on the SMG dataset with the same protocol: 30 training epochs (all fully converged), batch sizes of 32, the base learning rate of 0.05, and weight decay of 0.0005. Prepossessing was conducted for all the baselines: null-frame padding, translating to the center joint, and paralleling the joints to the corresponding axis. Input length is set as 60 frames. For all the remaining network hyperparameters, we kept their original settings (e.g., for MSG3D, the numbers of GCN scales and G3D scales are kept as 13 and 6). The optimizer is SGD which is consistent to the settings of all the models. The loss function is Categorical Cross Entropy. The training platform was with a single GPU: NVidia Titan (24 GB).

For pretraining the models for RGB modality, we used Resnet50 (He et al. 2016) as the backbone pretrained on something-something v2 (Goyal et al. 2017) dataset as it has been commonly used for all the three methods and the trained weighs are available. The hyperparameters are set as the same as the original work.

Appendix E Experimental Details for Online Recognition

We first conduct the same pre-processing of the skeleton streams on the three validating datasets. Since pre-segmented clips and their global temporal information of ongoing gestures, are not available in online recognition tasks, it's demanding to have an efficient local temporal feature extraction. For skeleton joint feature extraction, we followed the work of (Zanfir et al. 2013). "MovingPose" are features that utilize 3D position difference characters of joints to generate spatio-temporal information with efficient dimensional requirements.

Table 10 The online recognition performances of AED-BiLSTM under different threshold values of α_{th} on SMG dataset

Threshold of α_{th}	10%	30%	50%	70%	90%
F1-score	0.312	0.203	0.087	0.030	0.004

For the training phase, our AED-BiLSTM network was trained with a batch size of 32, the learning rate as 0.01 (reducing LR factor as 0.5, patience as 3 epochs) for 20 epochs on the SMG dataset, with a batch size of 64, the learning rate as 0.01 (reducing LR factor as 0.5, patience as 3 epochs) for 80 epochs on the iMiGUE dataset and 40 epochs on the OAD dataset. The optimizer is RMSprop, following the setting of Chen et al. (2020). he structure of AED-BiLSTM is consistent with the STABNet (Chen et al. 2020). Specifically, RMSprop is used as optimizer. The structure of STABNet is given as: a two-layer BiLSTM with 2000 GRUs and 1000 GRUs separately. A spatial attention layer and a temporal attention layer are attached before and between the two BiLSTM layers, respectively. The dense layer of 1000 units is stacked with the sigmoid activation to the BiLSTM layers, followed by a output layer with units as the total hidden state number (MG class number* HMM state number used for representing each MG, 16*5 in practice).Since the skeleton joints of the iMiGUE dataset are extracted from OpenPose (Cao et al. 2019) which contains noises, we filtered out all the training samples with null skeleton joints. For the compared methods, we use the same training scheme and ensure the models are converged. The training time of a single BiLSTM is around three hours with over 18,000/8,000 (training/validating) frame-level samples in our SMG dataset and around six hours with over 47,000/5,000 (training/validating) frame-level samples in the iMiGUE dataset. The loss function is Categorical Cross Entropy. The training platform was Tensorflow with a single GPU: NVidia Titan (24 GB).

Table 11 Ablation study of WSGN

Methods	Emotion state recognition accuracy			
	k-NN	Random Forest	AdaBoost	MLP
Basic graph	0.50	0.45	0.48	0.45
+ LP	0.50	0.50	0.48	0.45
+ SFFS	0.50	0.50	0.50	0.50
+ TS	0.56	0.51	0.38	0.51
+ LP & SFFS	0.50	0.50	0.50	0.50
+ LP & TS	**0.57**	0.42	0.48	0.50
+ SFFS & TS	0.47	0.62	0.62	0.60
+ SFFS & TS & LP (full model)	0.53	**0.65**	**0.65**	**0.65**

LP: Laplacian operation, TS: transition state embedding, and SFFS: sequential forward floating selection
Methods with the best performance are marked in bold

For testing and post-processing, the threshold of the minimal frames to filter out noisy gestures is set as 14 frames for all the methods. The values of λ and μ are set as $-2.5/2.1$, $-1.0/3.0$ and $-0.2/2.0$ for SMG, iMiGUE and OAD datasets.

Appendix F Ablation Study of Online Recognition

Although our AED is parameter-free and can be directly exploited to the inference, the correct value setting of λ and μ will affect the performance of the AED. Thus, we present the ablation study of AED-BiLSTM under different values of λ and μ, as shown in Tables 8 and 9.

Note that the λ will affect the inhibition of the "non-movement"s, which determines the segmentation results. Meanwhile, μ for the attention of the MGs will affect the classification results. Thus, we fix μ as 0.0 to conduct the ablation study to obtain the best value of λ, then get the best value of μ with obtained λ.

The online recognition performances of AED-BiLSTM under different threshold values of overlapping ratio α_{th} is shown in Table 10. The higher the α_{th} is, the more challenging the task is, as it requires more accurate temporal allocation of the frame boundaries. When it comes to 90%, it means the temporal allocation of the MGs should be extremely accurate. This is especially challenging due to the subtle and swift nature of MGs.

Appendix G Experimental Settings for Stress Emotional State Recognition

RGB modality was not used as it might bring unnecessary texture patterns like facial information into neural networks and makes the analysis contested. We focus on the skeleton modality in order to specifically explore the relationship between gestures and stress states. All the settings used Cross Entropy as the loss function.

Full Context Recognition. In practical implementation, we trained the baseline methods with the same protocols as the classification task (e.g., training epoch number, batch size, etc.). Besides, the input is the long skeleton sequence of an emotional state instance with a frame number of 90 via linear down-sampling. The dimension of the output layers of networks are modified into two in relation to the two emotional states.

MG-based Context Recognition. We construct the for hidden emotional recognition. The transition of the middle state (non-movements) is enabled, and the transition direction is enabled. Bayesian prior is added. Sequential Forward Floating Selection (SFFS) strategy was used for selecting MGs with the most contributions. From SFFS, "Turtling neck and shoulder", "Rubbing eyes and forehead", "Folding arms behind body" and "Arms akimbo" are the most contributed features for the emotional state recognition in the subject-independent protocol; meanwhile, "Rubbing eyes and forehead", "Moving legs", "Arms akimbo" and "Scratching or touching facial parts other than eyes" are the most contributed features for the semi-subject-independent protocol.

Appendix H Extra Experimental Results of WSGN

Ablation study. We present the contribution of each component in the WSGN with an ablation study as shown in Table 11 in the subject-independent protocol on the SMG dataset. As we can see, the three components (Laplacian operation, SFFS and transition state embedding) can jointly contribute to the performance.

References

Aviezer, H., Trope, Y., & Todorov, A. (2012). Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science, 338*(6111), 1225–1229.

Burgoon, J., Buller, D., & WG, W. (1994). *Nonverbal communication: The unspoken dialogue*. Greyden Press.

Cao, Z., Hidalgo, G., Simon, T., Wei, S. E., & Sheikh, Y. (2019). Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(1), 172–186.

Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6299–6308).

Chen, H., Liu, X., Li, X., Shi, H., & Zhao, G. (2019). Analyze spontaneous gestures for emotional stress state recognition: A micro-gesture dataset and analysis with deep learning. In *Proceedings of the IEEE international conference on automatic face & gesture recognition* (pp. 1–8).

Chen, H., Liu, X., Shi, J., & Zhao, G. (2020). Temporal hierarchical dictionary guided decoding for online gesture segmentation and recognition. *IEEE Transactions on Image Processing, 29*, 9689–9702.

Cheng, K., Zhang, Y., He, X., Chen, W., Cheng, J., & Lu, H. (2020). Skeleton-based action recognition with shift graph convolutional network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 183–192).

Crasto, N., Weinzaepfel, P., Alahari, K., & Schmid, C. (2019). MARS: Motion-augmented RGB stream for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

de Becker, G. (1997). *The gift of fear*. Dell Publishing.

de Lara, N., & Pineau, E. (2018). A simple baseline algorithm for graph classification. In *Relational representation learning workshop, the conference on neural information processing systems*.

Ekman, P. (2004). Darwin, deception, and facial expression. *Annals of the New York Academy of Sciences, 1000*, 205–21.

Ekman, R. (1997). *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press.

El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition, 44*(3), 572–587.

Escalera, S., Baró, X., Gonzàlez, J., Bautista, M.A., Madadi, M., Reyes, M., Ponce-López, V., Escalante, H.J., Shotton, J., & Guyon, I. (2015). Chalearn looking at people challenge 2014: Dataset and results. In *Proceedings of the European conference on computer vision* (pp. 459–473).

Fix, E., & Hodges, J. L. (1989). Discriminatory analysis. nonparametric discrimination: Consistency properties. *International Statistical Review/Revue Internationale de Statistique, 57*(3), 238–247.

Ginevra, C., Loic, K., & George, C. (2008). Emotion recognition through multiple modalities: Face, body gesture, speech (pp. 92–103). Springer.

Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., & Mueller-Freitag, M., et al. (2017). The" something something" video database for learning and evaluating visual common sense. In: *Proceedings of the IEEE international conference on computer vision* (pp. 5842–5850).

Gray, J. A. (1982). Précis of the neuropsychology of anxiety: An enquiry into the functions of the septo-hippocampal system. *Behavioral and Brain Sciences, 5*(3), 469–484.

Gu, Y., Mai, X., & Luo, Y. (2013). Do bodily expressions compete with facial expressions? Time course of integration of emotional signals from the face and the body. *PLOS ONE, 8*(7), 1–9.

Gunes, H., & Piccardi, M. (2006). A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In *18th international conference on pattern recognition* (vol. 1, pp. 1148–1153).

Hara, K., Kataoka, H., & Satoh, Y. (2018). Can spatiotemporal 3d cnns retrace the history of 2D CNNs and imagenet? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6546–6555).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Ho, T. K. (1995). Random decision forests. In *Proceedings of international conference on document analysis and recognition* (vol. 1, pp. 278–282).

Khan, R. Z., & Ibraheem, N. A. (2012). Hand gesture recognition: A literature review. *International Journal of Artificial Intelligence & Applications, 3*(4), 161.

Kipp, M., & Martin, J. C. (2009). Gesture and emotion: Can basic gestural form features discriminate emotions? In *International conference on affective computing and intelligent interaction and workshops* (pp. 1–8).

Kita, S., Alibali, M., & Chu, M. (2017). How do gestures influence thinking and speaking? the gesture-for-conceptualization hypothesis. *Psychological Review, 124*, 245–266.

Krakovsky, M. (2018). Artificial (emotional) intelligence. *Communications of the ACM, 61*(4), 18–19.

Kuehne, H., Richard, A., & Gall, J. (2019). A hybrid RNN-HMM approach for weakly supervised temporal action segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Kuhnke, E. (2009). *Body language for dummies*. Wiley.

Li, S., & Deng, W. (2020). Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*.

Li, Y., Lan, C., Xing, J., Zeng, W., Yuan, C., & Liu, J. (2016). Online human action detection using joint classification-regression recurrent neural networks. In *Proceedings of the European conference on computer vision*.

Lin, J., Gan, C., & Han, S. (2019). TSM: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 7083–7093).

Liu, J., Shahroudy, A., Xu, D., & Wang, G. (2016). Spatio-temporal LSTM with trust gates for 3D human action recognition. In: *Proceedings of the European conference on computer vision*.

Liu, J., Shahroudy, A., Wang, G., Duan, L.Y., & Kot, A.C. (2018). Ssnet: Scale selection network for online 3D action prediction. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Liu, J., Wang, G., Hu, P., Duan, L.Y., & Kot, A. C. (2017). Global context-aware attention LSTM networks for 3d action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Liu, X., Shi, H., Chen, H., Yu, Z., Li, X., & Zhao, G. (2021). imigue: An identity-free video dataset for micro-gesture understanding and emotion analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10631–10642).

Liu, Z., Zhang, H., Chen, Z., Wang, Z., & Ouyang, W. (2020). Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 143–152).

Luo, Y., Ye, J., Adams, R. B., Li, J., Newman, M. G., & Wang, J. Z. (2020). Arbee: Towards automated recognition of bodily expression of emotion in the wild. *International Journal of Computer Vision, 128*(1), 1–25.

Mahmoud, M., Baltrušaitis, T., Robinson, P., & Riek, L.D. (2011). 3D corpus of spontaneous complex mental states. In *International conference on affective computing and intelligent interaction* (pp. 205–214).

Navarro, J., & Karlins, M. (2008). *What every BODY is saying: An ex-FBI agent's guide to speed reading people*. Collins.

Neverova, N., Wolf, C., Taylor, G., & Nebout, F. (2016). Moddrop: Adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence 38*(8).

Noroozi, F., Kaminska, D., Corneanu, C., Sapinski, T., Escalera, S., & Anbarjafari, G. (2018). Survey on emotional body gesture recognition. *IEEE Transactions on Affective Computing*.

Oh, S. J., Benenson, R., Fritz, M., & Schiele, B. (2016). Faceless person recognition: Privacy implications in social media. In *Proceedings of the European conference on computer vision* (pp. 19–35).

Palena, N., Caso, L., Vrij, A., & Orthey, R. (2018). Detecting deception through small talk and comparable truth baselines. *Journal of Investigative Psychology and Offender Profiling* 15.

Panksepp, J. (1998). *Affective neuroscience: The foundations of human and animal emotions*. Oxford University Press.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019) Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*.

Peng, W., Hong, X., Chen, H., & Zhao, G. (2020). Learning graph convolutional network for skeleton-based human action recognition by neural searching. In *Proceedings of the AAAI conference on artificial intelligence*.

Pentland, A. (2008). *Honest signals: How they shape our world*. MIT Press.

Pouw, W. T., Mavilidi, M. F., Van Gog, T., & Paas, F. (2016). Gesturing during mental problem solving reduces eye movements, especially for individuals with lower visual working memory capacity. *Cognitive Processing, 17*(3), 269–277.

Richard, A., Kuehne, H., Iqbal, A., & Gall, J. (2018). Neuralnetwork-viterbi: A framework for weakly supervised video learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*(6088), 533–536.

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2008). The graph neural network model. *IEEE Transactions on Neural Networks, 20*(1), 61–80.

Schapire, R. E. (2013). Explaining adaboost. In *Empirical inference* (pp. 37–52). Springer.

Schindler, K., Van Gool, L., & De Gelder, B. (2008). Recognizing emotions expressed by body pose: A biologically inspired neural model. *Neural Networks, 21*(9), 1238–1246.

Serge, G. (1995). International Glossary of Gestalt Psychotherapy. FORGE.

Shahroudy, A., Liu, J., Ng, T. T., & Wang, G. (2016). Ntu rgb+d: A large scale dataset for 3D human activity analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Shi, L., Zhang, Y., Cheng, J., & Lu, H. (2019). Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12026–12035).

Shiffrar, M., Kaiser, M., & Chouchourelou, A. (2011). Seeing human movement as inherently social. The Science of Social Vision.

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., & Blake, A. (2011). Real-time human pose recognition in parts from single depth images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1297–1304).

Soomro, K., Zamir, A. R., & Shah, M. (2012). Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402.

Sun, S., Kuang, Z., Sheng, L., Ouyang, W., & Zhang, W. (2018). Optical flow guided feature: A fast and robust motion representation for video action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3D convolutional networks. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4489–4497).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems* 30.

Vrij, A., Leal, S., Jupe, L., & Harvey, A. (2018). Within-subjects verbal lie detection measures: A comparison between total detail and proportion of complications. *Legal and Criminological Psychology, 23*(2), 265–279.

Vrij, A., Mann, S., Leal, S., & Fisher, R. P. (2020). Combining verbal veracity assessment techniques to distinguish truth tellers from lie tellers. *European Journal of Psychology Applied to Legal Context, 13*(1), 9–19.

Wallbott, H. G. (1998). Bodily expression of emotion. *European Journal of Social Psychology, 28*(6), 879–896.

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Van Gool, L. (2018). Temporal segment networks for action recognition in videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(11), 2740–2755.

Wu, D., Pigou, L., Kindermans, P.J., Le, N.D.H., Shao, L., Dambre, J., & Odobez, J.M. (2016). Deep dynamic neural networks for multimodal gesture segmentation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence 38*(8).

Xu, M., Gao, M., Chen, Y. T., Davis, L. S., & Crandall, D. J. (2019a). Temporal recurrent networks for online action detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 5532–5541).

Xu, M., Gao, M., Chen, Y.T., Davis, L. S., & Crandall, D. J. (2019b). Temporal recurrent networks for online action detection. In *Proceedings of the IEEE/CVF international conference on computer vision*.

Yan, S., Xiong, Y., & Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence* (vol. 32).

You, Y., Chen, T., Wang, Z., & Shen, Y. (2020). L2-gcn: Layer-wise and learned efficient training of graph convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2127–2135).

Yu, N. (2008). Metaphor from body and culture. The Cambridge handbook of metaphor and thought (pp. 247–261).

Yu, Z., Zhou, B., Wan, J., Wang, P., Chen, H., Liu, X., Li, S. Z., & Zhao, G. (2020). Searching multi-rate and multi-modal temporal enhanced networks for gesture recognition. *IEEE Transactions on Image Processing*.

Zanfir, M., Leordeanu, M., & Sminchisescu, C. (2013). The moving pose: An efficient 3D kinematics descriptor for low-latency action recognition and detection. In *Proceedings of the IEEE/CVF international conference on computer vision*.

Zhang, Y., Pal, S., Coates, M., & Ustebay, D. (2019). Bayesian graph convolutional neural networks for semi-supervised classification. In *Proceedings of the AAAI conference on artificial intelligence* (vol. 33, pp. 5829–5836).