iMiGUE: An Identity-free Video Dataset for Micro-Gesture Understanding and Emotion Analysis

Xin Liu[†], Henglin Shi[‡], Haoyu Chen[‡], Zitong Yu[‡], Xiaobai Li[‡], Guoying Zhao[‡*]
[‡]Center for Machine Vision and Signal Analysis, University of Oulu, Finland
[†]School of Electrical and Information Engineering, Tianjin University, China
https://github.com/linuxsino/iMiGUE

Abstract

*We introduce a new dataset for the emotional artificial intelligence research: **i**dentity-free video dataset for **Mi**cro-**G**esture **U**nderstanding and **E**motion analysis (iMiGUE). Different from existing public datasets, iMiGUE focuses on nonverbal body gestures without using any identity information, while the predominant researches of emotion analysis concern sensitive biometric data, like face and speech. Most importantly, iMiGUE focuses on micro-gestures, i.e., unintentional behaviors driven by inner feelings, which are different from ordinary scope of gestures from other gesture datasets which are mostly intentionally performed for illustrative purposes. Furthermore, iMiGUE is designed to evaluate the ability of models to analyze the emotional states by integrating information of recognized micro-gesture, rather than just recognizing prototypes in the sequences separately (or isolatedly). This is because the real need for emotion AI is to understand the emotional states behind gestures in a holistic way. Moreover, to counter for the challenge of imbalanced sample distribution of this dataset, an unsupervised learning method is proposed to capture latent representations from the micro-gesture sequences themselves. We systematically investigate representative methods on this dataset, and comprehensive experimental results reveal several interesting insights from the iMiGUE, e.g., micro-gesture-based analysis can promote emotion understanding. We confirm that the new iMiGUE dataset could advance studies of micro-gesture and emotion AI.*

1. Introduction

Emotional artificial intelligence (emotion AI) is using machine learning methods to enable computers to understand human emotions. It plays a vital role in human-computer interaction since emotions are on all the time, presented in all kinds of human activities, thinking, and decision makings. According to psychological studies, body language is an essential part for understanding human emotions. Every day, we respond to thousands of such nonverbal behaviors including facial expressions, eye movements or gaze, tone of voices, gestures, touches, and the use of space. Body language-based emotion understanding has attracted extensive attention in the communities of computer vision and affective computing, and a considerable number of datasets have been proposed, *e.g.*, posed facial expressions [30, 56, 92, 82, 25, 96], spontaneous facial behaviors [1, 47, 4, 51, 12, 35], micro-expressions [91, 41, 11], voice/speech [65, 66, 51, 61], social signals [28, 29], and multi-modal datasets with facial expressions and physiological signals [73, 34, 51, 61]. Although computational methodologies were proposed correspondingly and consecutively to improve the performance on these datasets, there are still significant gaps between current studies and the needs of real applications. Major limitations include:

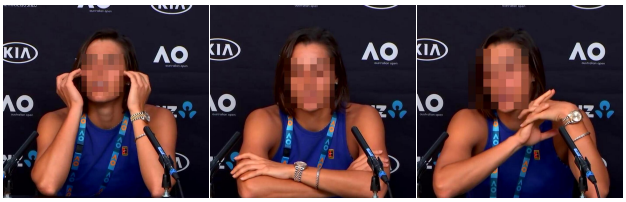

Figure 1. Three frames (cropped) from a post-match press conference video to illustrate the identity-free micro-gestures, such as "cover face", "fold arms", and "cross fingers". Could machine recognize these micro-gestures, and understand emotional states of the player in a holistic way, and further identify if the player has won or lost the match (positive or negative emotional states)?

1) **Intentional behavioral-based gestures**. Previous gesture studies mostly focused on illustrative (or iconic) gestures [84], *e.g.*, waving hands as "hello" or "goodbye", which are intentionally performed for conveying certain

*Corresponding author. Xin Liu's work mostly done at the University of Oulu. This paper is supported by Academy of Finland, KAUTE foundation, and National Natural Science Foundation of China.

meanings or feelings during interactions. However, in many occasions people would suppress or hide their feelings (especially negative ones) rather than expressing them. Previous studies [2, 3, 6] showed that there is a special group of gestures, the micro-gestures (MGs), which are helpful to understand such suppressed or hidden emotions. The major difference between MGs and illustrative gestures is that MGs are unintentional behaviors elicited by people's inner feelings, *e.g.*, rubbing hands due to stress, and the function of MGs is for relieving or protecting oneself from negative feelings rather than presenting for others. Thus, being able to automatically recognize MGs would allow emotion understanding at a better level. To the best of our knowledge, there is no publicly available dataset for this emotional MGs research in the field of computer vision.

2) **Gap between behavior recognition and emotion understanding**. Most existing datasets only aim to evaluate approaches that can detect and recognize prototypes of behaviors (including gestures). In fact, the actual need of emotion AI is not merely to recognize certain behaviors, but to uncover the emotion underneath. Consider a post-match interview scenario, a player is interviewed by reporters over several question & answer rounds (see Fig. 1). Some MGs could be observed, *e.g.*, cross arms (defensive) and cover face (upset or ashamed), but it is hoped that the machine can understand (identify) if the player has a positive or negative feeling (*e.g.*, caused by winning or losing of the match).

3) **Sensitive biometric data**. Most of the existing datasets involve sensitive biometric data. Actually, biometric data based identity recognition plays a critical role in a variety of applications and has gained great success in the past decade. While every coin has two sides, biometric information is so sensitive that is particularly prone to be (identity) stolen, misused, and unauthorized tracked. With the concerns of privacy grows, more attention should be paid to protect biometric data of individuals.

Psychological studies [16] showed that there are over 215 behaviors associated with psychological discomfort and most of them are not in the face. MGs are subtle and (some of them) short, mostly out of our awareness or notice during live interactions [21]. It would be of great value if we can develop computer vision methods to capture and recognize these neglected clues for better emotion understanding. In this paper, we introduce a novel MGs dataset to address afore mentioned limitations. The key contributions are summed up as follows:

1) Instead of using facial or vocal-expressions for emotion understanding, the proposed dataset offers an approach where the identity-free MGs are explored for hidden emotion understanding, and privacy of the individuals could be preserved. As far as we know, iMiGUE is the first public benchmark focuses on emotional MGs. This is the first investigation of such gestures from the computer vision perspective. Moreover, to deal with the issue of imbalanced classes distribution, an unsupervised model is proposed.

2) iMiGUE is not only for MG recognition, but also provides a hierarchy that allows exploration of the relationship between MGs and emotion, *i.e*, associates the MGs holistically for emotion understanding. As such, the data in iMiGUE are annotated on two levels: the MG categories were annotated on video clip-level, and the emotion categories were labeled on video-level.

3) Comprehensive experiments are conducted on the iMiGUE to provide baseline results. In video clip-level, the experimental results show that even fully supervised learning SOTA models cannot yield satisfactory accuracy on iMiGUE, which could verify that the challenges of recognizing such hardly noticeable MGs. The proposed unsupervised method can achieve competitive performance compared with many supervised models. In video-level, we find micro-gesture is a vital factor for emotion understanding. We only employed a simple recurrent neural network (RNN) network to achieve MGs analysis in a holistic way, but its emotion understanding result can beat existing action/gesture recognition-based models. The dataset and findings will serve as a launch-pad for exploring identity-free MG-based emotion AI.

2. Related Work

A person's emotional state is often conveyed through bodily expression. As such, analyzing body based activities, including action, gesture and posture are the popular research topics [75, 36, 23, 62, 67, 71, 24, 53, 93] in the community. However, these datasets focused on recognizing human activities (*e.g.*, a man is jumping), rarely related to the emotional states. We limit our review on the related emotional gesture-based benchmarks. Then we review related work of gesture/action recognition.

2.1. Emotional Gesture-based Datasets

Gesture is one of the key cues of social communication which includes movements of hands, head and other parts of human body that express various feelings, thoughts and emotions [55]. Table 1 summarizes the attributes of widely used databases of emotional gestures. In this field, early studies were mostly built on acted or posed gestures. The Tilburg University Stimulus set [64] collected photographic still images of 50 actors enacting different emotions. FABO database [26] is one of the pioneer work which proposed using video clips of posed prototype gestures to recognize emotions. These videos were labelled with six basic emotions, as well as four more states, namely, neutral, anxiety, boredom, and uncertainty. Following that posed behavior which was captured in controlled recording conditions, researchers extended emotional gesture analysis into many directions. In HUMAINE [14, 9], the researchers elicited

Datasets	# Gestures	# Emotions	#Subjects (F/M)	# Samples	#Videos	Duration	Context	Expression	Resolution	Modalities	Recognition
FABO [26]	-	10	23 (12/11)	206	23	6 Min	C	Posed	1024×768	F + G	Isolated
HUMAINE [14]	8	8	10 (4/6)	240	240	5-180 Sec	C	Posed	-	F + G	Isolated
GEMEP [22]	-	18	10 (5/5)	7 000+	1260	-	C	Posed	720×576	F + G	Isolated
THEATER [32]	-	8	-	258	-	-	U	SP	-	G	Isolated
EMILYA [19]	7	8	11 (6/5)	7 084	23	5.5 Sec	C	Posed	1280×720	G	Isolated
LIRIS-ACCEDE [20]	6	6	64 (32/32)	-	-	1 Min	C	Posed	-	F + G	Isolated
emoFBVP [60]	23	23	10 (-)	1 380	-	20-66 Sec	C	Posed	640×480	F + G + V	Isolated
BoLD [48]	-	26	-	13 239	9 876	-	U	SP	-	G	Isolated
iMiGUE (Ours)	32	2	72 (36/36)	18 499	359	0.5-25.8 Min	U	SP	1280×720	IMG	Holistic

Table 1. The attributes comparison of iMiGUE with other widely used datasets for recognizing gesture-based expression of emotion. F/M: Female/Male, C: Controlled (in-the-lab), U: Uncontrolled (in-the-wild), SP: Spontaneous, F: Face, G: Gesture, V: Voice, IMG: Identity-free Micro-Gesture.

emotions via interaction with computer avatar of its operator. The Geneva multi-modal emotion portrayals (GEMEP) database [22] contains more than 7 000 audio-video portrayals of 18 emotions portrayed by 10 actors. Also in a controlled setting, the subset [20] of LIRIS-ACCEDE database [5] collected upper body emotional gestures from 64 subjects. Using a Kinect sensor, Saha *et al*. [63] collected 3D skeleton gesture data of 10 subjects, which included five induced emotions, *i.e*., anger, fear, happiness, sadness, and relaxation. Psaltis *et al*. [59] collected skeletal gestural expressions that frequently appear in a game-play scenario. Similarly, the emoFBVP [60] dataset has a multi-modal recordings of actors including body gestures with skeletal tracking. Emilya [19] dataset captured 3D body movements of posed emotions via a motion capture system.

Later studies focused more toward spontaneous emotional gestures, which are more challenging than posed ones. In the Theater [32] dataset, the emotional gesture video clips were extracted from two movies, which are close to real world scenes. In [33], gesture movements are recorded while subjects were playing body movement based video games. Luo *et al*. collected a large-scale bodily expression dataset, the BoLD [48], in which the in-the-wild perceived emotion data were segmented from movies and reality TV shows. To date, few efforts were made on the fine-grained micro visual of the body, *i.e*., the MG, which is important clue for understanding suppressed or concealed emotions.

2.2. Methods for Gesture/Action Recognition

Early work of automatic modeling of emotional gestures depends largely on hand-crafted features [26, 64, 33]. Recently, numerous neural networks have been introduced for gesture/action recognition. Among them, supervised learning is the predominate technique for which labeled data are utilized to train the models. The earliest attempts utilized a 2D convolutional neural network (CNN) [72, 18, 86, 98, 43, 94] to extract spatial features from the selected frames, and the temporal aggregation is considered by an additional stream of optical flow or the temporal pooling layers. The 3D CNNs [79, 8, 87, 81, 27, 88, 80] can jointly capture spatial-temporal semantics, where the filters are designed in a 3D manner. Compared to the 2D CNNs, 3D ones can process the temporal information hierarchically throughout the whole network. Also, some models like the Slow-fast [17] considered a joint implementation of both two streams (fast and slow) and 3D CNN. The RNN is also commonly used for temporal integration. Specifically, the long short-term memories (LSTMs) [13, 15, 83, 99, 50, 42, 44, 45] have demonstrated their strength on learning sequential data. Recently, the skeleton data is gaining increasingly popularity because of their invariance to background dynamics. Current work on skeleton based action recognition can mainly be categorized into two types: one is RNN based methods [15, 44, 74, 45, 95, 40] which directly process gesture skeletons as time series, and the other one is graph convolutional network (GCN) [90, 38, 70, 69, 58, 10, 46] based methods which reorganize the skeleton data as a graph.

Compared with the supervised methods, the task of behavior recognition with unsupervised approaches is much more challenging, and only a few attempts have been reported. Some methods focused on leveraging temporal information of videos to learn visual representation, such as Shuffle & Learn [52], OPN [37], and [89, 78, 85, 31]. Other methods utilized the encoder-decoder-based video sequence/frame reconstruction, *e.g*., RGB-based [76, 49, 39], and skeleton-based LongT GAN [97] and Predict & Cluster (P&C) [77]. The problem is that P&C just used the reconstruction loss in an element-wise manner without considering any informative constraint or prior. In addition, P&C employed a fixed-length input scheme which is hard to encode the long-term motion dependencies since the downsampled sequences may lose essential information.

3. The iMiGUE dataset

3.1. Key Challenges

We hope to draw more attention on analyzing micro-gestures starting with building and sharing a new MG database. We need to solve several unprecedent challenges to build the iMiGUE as it is different than previous gesture dataset. 1) ***How to define and organize the categories of MGs related to emotions?*** We take reference from psycho-

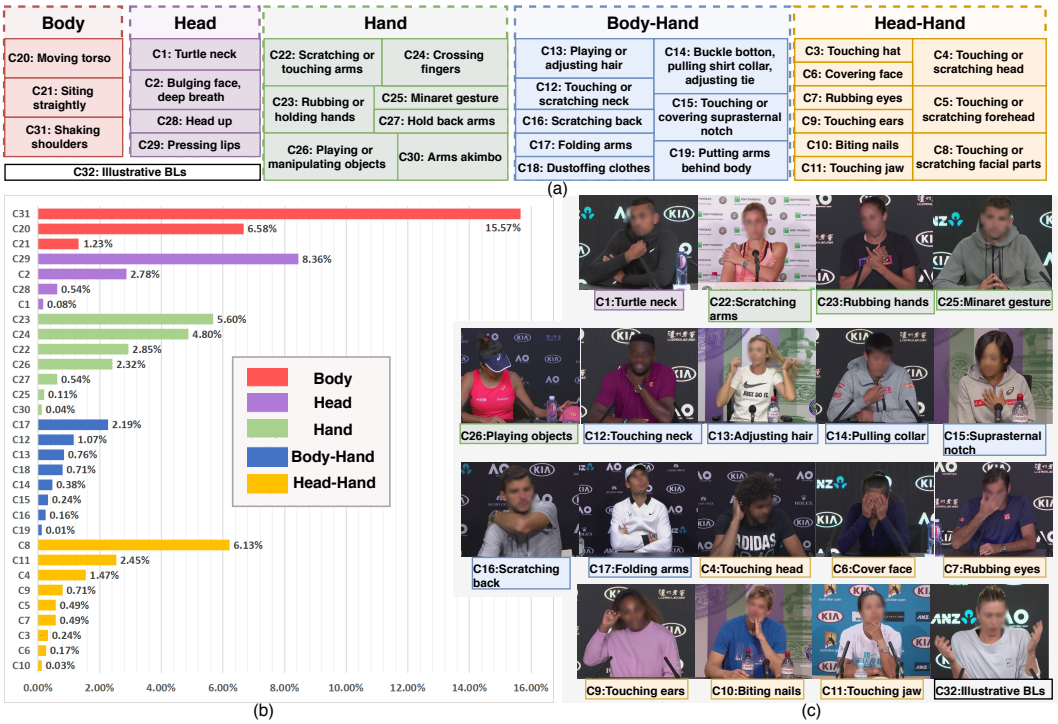

Figure 2. (a) Categories of MGs in iMiGUE dateset which refers to psychological studies [16, 57, 54]. (b) Sample percentages of each category in iMiGUE dataset. (c) Examples (face masked) of micro-gesture categories in iMiGUE dataset.

logical studies [16, 57, 54] when considering the categories of MGs, the focus is to clarify the boundary to differentiate MGs from illustrative behaviors. 2) ***How to elicit or collect the spontaneous MGs?*** The new dataset will definitely contain genuine expressions but not posed ones, and we think it is good to start with selecting and collecting real world videos that contain MG occasion, *e.g.*, from the online video-sharing platforms. Furthermore, as an important goal of iMiGUE is to study the emotional states behind MGs, we need to find and provide the root cause of these (MGs) emotional outbursts. 3) ***How to annotate the data?*** The quality of labeling will greatly influence the final results of model training and testing. We need to annotate the data on two levels, first is the labels of all MG occasions (clip-level), and second is the labels of corresponding emotion (video-level). The MG labels are based on criteria from related psychological studies. We consider "positive" and "negative" as two emotional categories to start with, and the labels are based on objective facts, *i.e.*, winning (positive emotion) or losing (negative emotion) of a match. Considering the complexity and diversity of the MGs, a team trained specifically for the annotation job has been hired. To further make sure the high-quality annotation, an efficient mechanism for quality control has been designed (see Sec. 3.2).

3.2. Dataset Construction

Data collection. Based on the above factors, we search and collect videos containing scenarios of "post-match press conferences" (see Fig. 1), in which a professional athlete was interviewed by journalists and reporters over several question & answer rounds after a tough match. The players had no (or a little) time to prepare as the press conference will be held right after the match, and he or she needs to respond to the questions rapidly. Even though an experienced person could respond with "witty" statements, the emotion-related MGs may leak out unintentionally and unconsciously, since "the subconscious mind acts automatically and independently of our verbal lie" [57]. The result of the match, winning or losing is a natural emotion inducer leads to positive or negative emotion states of the interviewed player. Accordingly, for the research purpose of emotion AI, these identity-free MGs should be recognized and understood by machines, and a real AI could use them holistically to understand emotional state of the player.

We choose the post-match press conferences of "Grand Slam" (tennis) tournaments as the first data-source for iMiGUE, because they have several merits: 1) There are large numbers of openly accessed videos with high recording quality, *e.g.*, resolutions of at least 720P, so that even subtle differences between MG instances are well preserved. 2) MG-centric instances without background interference. Unlike several datasets where the background is either disturbing or distinguishable for different categories, all post-match press conference videos were recorded with the same static advertising wall background. 3) Diversity of cultures and nationalities. The players of tournaments

come from almost every corner of the Earth (see statistics of iMiGUE in Sec.3.3). 4) Good gender balance. Each Grand Slam tournament has 128 male and 128 female players, so it is easy to setup a gender-balanced dataset.

Data Annotation. The collected videos are annotated on two levels: the MG categories, and the emotion categories. The emotions are annotated on video level, *i.e.*, one emotion label for each press conference video. We consider two emotion categories: positive, *i.e.*, corresponding to the winning case, and negative, *i.e.*, corresponding to the losing cases. Then we search through the videos to spot and exert all MG instances (clips) and assign MG category labels for them. The work of MG annotation was very difficult and time consuming, and we took three measures as follows to ensure the quality of annotation. 1) ***Clarify the scope and categories of MGs.*** According to reference psychological studies [16, 57, 54], MGs could be clusters as five major groups according to the motions' locations and functions, *i.e.*, "Head", "Body", "Hand", "Body-Hand", and "Head-Hand", and each major group contains multiple fine-categories of MGs. The iMiGUE covers altogether 31 categories of MGs plus one extra category of non-MGs, *i.e.*, illustrative gestures (see Fig. 2 (a) for details). 2) ***Multiple labelers and training for labeling.*** We have five persons working together on MG annotation for two reasons: the first is to speed up the process, and the second is to reduce personal bias for more reliable annotation. Before the actual annotation, all five were trained to unify their criteria for MG annotation. First, they went through the descriptions and sample figure or video of the 32 categories of iMiGUE to get understanding of the characteristics of MGs, and primarily rules of instance durations (the starting and ending points) were also discussed. Then, they went through three rounds of labeling exercises, *i.e.*, in each exercise, every labeler first labeled two sample clips separately and then compared their labels together, different opinions were carefully discussed until agreements were reached by all annotators. 3) ***Cross check for reliable annotations.*** The task of annotating all video clips was divided for five persons to ensure that every clip has two labelers. After all five labelers finished labeling, a cross check of their annotations were carried out following the Eq. 1

$$\mathcal{R} = \frac{2 \times MG(L_i, L_j)}{\#All\_MG}, \tag{1}$$

where $MG(L_i, L_j)$ is the number of MGs on which Labeler i and Labeler j agreed, and $\#All\_MG$ is the total number of MGs annotated by the two labelers. The average inter-labeler reliability $\mathcal{R}_{avg}$ of iMiGUE is 0.81 which indicates reliable annotations. For the inconsistent annotation cases, the five labelers discussed them through and kept those with unified opinions while the rest (still with diverse opinions) were left out of the final label list.

3.3. Dataset Statistics and Properties

iMiGUE collected 359 videos (258 wins and 101 losses) of post match press conferences of Grand Slam tournaments from online video sharing platforms, *e.g.*, YouTube, of the total length of 2 092 minutes. The videos' duration varies with an average length of 350 seconds. The videos' resolution is 1280×720, and their frame rate is 25 fps. A total of 18 499 MG samples were labeled out and assigned with 32 category labels, *i.e.*, about 51 MG samples each video on average. The length of MG instances also varies, from 0.18s (second) to 80.92s with an average duration of 2.55s. Table 1 shows the key characteristics numbers of the iMiGUE compares with other gesture datasets. Notice that the sample numbers of the 32 MG categories vary a lot (see Fig. 2 (b)), which is a common situation in many spontaneous emotion datasets [91, 41, 11] as it is not control-recorded data and the occurrence of different behaviors naturally varies. The sample unbalance is one challenge for MG recognition, which we will elaborate later in Sec. 4.

This iMiGUE dataset has some attracting properties that distinguishes it from existing work. 1) **Micro-gesture-based dataset.** To the best of our knowledge, this is the first public dataset of micro-gestures, which is built to analyse these very fine clues with computer vision methods for recognizing and understanding suppressed or concealed emotions. 2) **Identity-free.** The sensitive biometric data, such as the face and voice have been masked and removed. 3) **Ethnic diversity.** iMiGUE contains 72 players from 28 countries and regions (*e.g.*, Argentina, Australia, Spain, Canada, China, United States, and South Africa) covering every continent which enables MGs analysis from diverse cultures. 4) **Gender-balanced.** iMiGUE comprises 36 female and 36 male players whose ages are between 17 and 38. 5) **Winning and losing as the natural and objective reference for emotion categories.** The iMiGUE is built not only for MG recognition but more importantly for exploring the relationship of MGs and the emotional states. As a new dataset with many unestablished factors, instead of arbitrarily assigned emotion labels which could be biased by subjective judgments, the results of matches could serve as a more objective reference of emotional states, *i.e.*, to assume that winning a match would lead to a more positive emotion status than losing one. Because this dataset is to analyze MGs and further recognize suppressed emotions without using sensitive biometric data, we suggest the researchers who work on estimating emotional states but concern the privacy issues could use this dataset as a benchmark.

4. Unsupervised Learning for MG Recognition

After the dataset building, a challenging issue is discovered and should be carefully discussed. Compared with the controlled recording circumstance with the fixed or planed

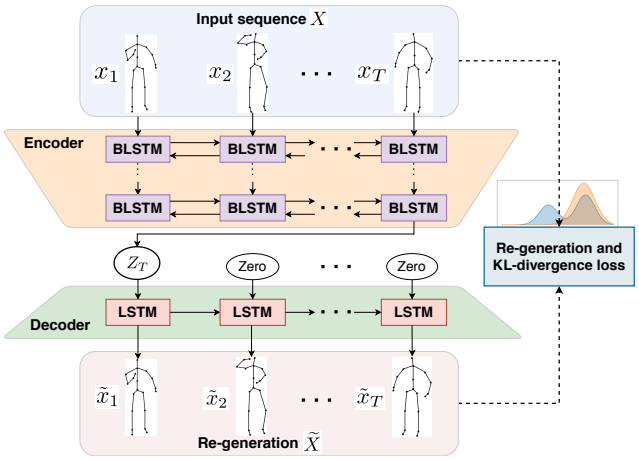

Figure 3. Framework of the proposed unsupervised encoder-decoder network.

number of samples, the imbalanced data issue is hard to avoid in the condition of an in-the-wild setting. In other words, iMiGUE dataset has a long-tailed distribution with class imbalance issue (see Fig. 2 (b)), which may raise a challenge for fully-supervised learning models and cause a significant performance drop under the extreme label bias.

As an intuitive substitution of a fully supervised method, the unsupervised model is advantageous since it does not require human-labeled data. In this paper, following the Seq2Seq unsupervised learning routine [76, 77], we introduce an encoder-decoder model to learn discriminative information of MG (pose or key-points-based) sequences without using labelled data. There are a few key differences between our method and previous unsupervised models: 1) We introduce the mutual information to control characteristics of the representation by matching to a prior distribution adversarially, namely, the Kullback-Leibler (KL) divergence is utilized to act as a measure of non-linear statistical dependence between input sequence and reconstructed one which facilitates the model to learn inherent action/gesture representations. While most of existing methods only rely on traditional element-wise loss, *e.g.*, mean square error (L_2 distance) [97] [77] and mean absolute error (L_1) [77]. 2) Unlike other Seq2Seq type encoder-decoder with a fixed-length scheme [76] [77] that only read in parts of the input sequence, we provide a flexible strategy which enables the encoder to read in the whole input sequence, aiming to utilize the complete context information and capture the long-term dynamics in sequences with arbitrary lengthes. To realize the above functions, we extend the preliminary model of sequential variational autoencoder (S-VAE) [68], which is a variant of VAE whose encoder-decoder are implemented by bidirectional LSTM (BLSTM). Different to the vanilla VAE, S-VAE can handle sequential data and capture latent patterns from the whole input sequence.

The framework of the proposed unsupervised S-VAE (U-S-VAE) is illustrated in Fig. 3. The encoder of U-S-VAE is a multi-layer BLSTM in which the input is a whole sequence of body key-points (pose) corresponding to an MG $X = (x_1, x_2, ..., x_T)$. After the last frame is read in, the hidden state Z_T is passed to the decoder which acts as the holistic summary of X. During the decoding phase, a simple LSTM decoder receives the Z_T at the first time-step and further re-generates the whole input sequence, denoted as $\widetilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_T)$. In particular, we train U-S-VAE with a joint loss function:

$$\mathcal{L}_{\text{joint}} = \mathcal{L}_{\text{reg}} + \lambda \mathcal{L}_{\text{KL}}, \tag{2}$$

where $\mathcal{L}_{\text{reg}}$ is the element-wise-based re-generation loss ($||X - \widetilde{X}||_2^2$) which is responsible for ensuring overall structure similarity between the input and the reconstructed one. Importantly, the KL-divergence is introduced to ensure closer approximation to the joint distribution and the product of the marginals. The motivation is to train a representation-learning encoder-decoder to maximize the mutual information between the inputs and the re-generated. The λ controls the weight of KL-divergence loss. As shown in Fig. 3, the decoder LSTM reads in the Z_T as the first-frame data to initiate its states. In each of the next time step, instead of the masked ground truth [97] or any other meaningful information as an input, zeros are being fed into the decoder. This operation aims to weaken the decoder which cannot get any information for prediction and it exclusively relies on the state Z_T passed by the encoder. In other words, this strategy enforces the encoder to learn the latent features and represent them with the final state transferred to the decoder. After this unsupervised network is trained, the latent state Z_T of the encoder can be used for classifying MG. Similar to [77], for the feature vectors Z_T of all sequences in the training set, a K-nearest neighbors (KNN) classifier is used to assign classes.

5. Experiments

5.1. Benchmark Evaluations

To have standard evaluations for all the reported results on the iMiGUE dataset, a two-level criteria has been defined. More specifically, on the MG recognition level, we utilized the cross-subject evaluation protocol which divides the 72 subjects into a training group of 37 subjects and a testing group of 35 subjects. The training and testing sets have 13 936 and 4 563 MG samples, respectively. The IDs of training and testing subjects can be found in the supplementary materials. On the emotion classification level, we selected 102 videos (51 win and 51 lose matches) as the training set, and 100 videos (50 win and 50 loss matches) as the test set. The player's emotional states (positive/negative) with the result of win or loss, are classified via analysis of MGs. The details of training and testing pro-

Methods		Model+Modality	Accuracy	
			Top-1	Top-5
Super-vised	S-VAE [68]	RNN + Pose	27.38	60.44
	LSTM		32.36	72.93
	BLSTM		32.39	71.34
	ST-GCN [90]	GCN + Pose	46.97	84.09
	2S-GCN [69]		47.78	88.43
	Shift-GCN [10]		51.51	88.18
	GCN-NAS [58]		53.90	89.21
	MS-G3D [46]		54.91	<u>89.98</u>
	C3D [79]	3DCNN + RGB	20.32	55.31
	R3D-101 [27]		25.27	59.39
	I3D [8]		34.96	63.69
	TSN [86]	2DCNN + RGB	51.54	85.42
	TRN [98]		<u>55.24</u>	89.17
	TSM [43]		**61.10**	**91.24**
Unsup-ervised	P&C [77]	Encoder-Decoder + Pose	31.67	64.93
	U-S-VAE Z (Ours)		32.43	64.30

Table 2. Comparison of MG recognition accuracy (%) with state-of-the-art algorithms on the iMiGUE dataset (best: bold, second best: underlined).

tocols (video IDs) can be found in the supplementary material. Specially, to benefit the community of skeleton or pose-based gesture recognition, we provide the pose data of every frame, achieved by using the OpenPose toolbox [7].

5.2. Implementation details

In the proposed U-S-VAE, we set the following architecture: Encoder: 1-Layer BLSTM with N = 256 units for each direction. Decoder: 1-Layer LSTM with N = 256 units. The learning rate is 0.0002 with a decay factor of 0.999 for every five training epochs. The network is trained till the loss converges such that the training loss tends to be stable.

A series of experiments are conducted on the iMiGUE dataset on a PC with a Titan RTX GPU. All training configurations follow the original papers unless stated otherwise.

5.3. Clip-level Micro-gesture Recognition

In order to evaluate supervised learning-based methods' performance on iMiGUE, 14 state-of-the-art algorithms are selected which can be simply categorized into four groups, namely, body key-points-based RNN (*i.e.*, BLSTM, LSTM, and S-VAE [68]), and GCN (*i.e.*, ST-GCN [90], 2S-GCN [69], Shift-GCN [10], GCN-NAS [58], and MS-G3D [46]), RGB-based 3DCNN (*i.e.*, C3D [79], R3D-101 [27], and I3D [8]), and 2DCNN with temporal reasoning (*i.e.*, TSN [86], TRN [98], and TSM [43]). We further evaluate the effectiveness of the proposed U-S-VAE by comparing it with existing unsupervised methods. In fact, only a few pose (skeleton)-based unsupervised models were proposed, *e.g.*, LongT GAN [97] and P&C [77]. Here, we report the results of P&C since its implement code is publicly available. It is noted that all models follow the same evaluation protocol mentioned above for a fair comparison. In Table 2, we present the performances of these baseline networks.

From Table 2, we can summarize several observations:

Figure 4. Examples of challenging recognition of the micro-gestures.

1) Almost all of methods' accuracy (Top-1) stuck under 60 percentage, which could verify that recognizing such hardly noticeable MGs (*e.g.*, a short-timing "Shake shoulders" as shown in Fig. 4 (a)) is a very challenging task. Due to the subtle differences between MGs (*e.g.*, "Covering face" vs. "Touching forehead", "Touching neck" vs. "Covering suprasternal notch" as shown in Fig. 4 (b) and (c)), visual or structural appearances in the form of RGB or pose contribute significantly less than that in a regular gesture (action) recognition task. 2) 3DCNN and RNN-based models' Top-1 performance are lower than 35%, which is not surprising as fully-supervised learning models may have a significant performance drop with class imbalance issue. 3) Capturing temporal dynamics (temporal reasoning) is important as 2DCNN-based TSM and TRN outperform others by large margins. 4) Our method outperforms prior unsupervised learning model P&C. Although not using any labels, our performance is very competitive with the supervised 3DCNN and RNN-based methods.

5.4. Video-level Emotion Understanding

Evaluations also include quantitative analysis comparing performance of the-state-of-art methods for the task of video-level emotion understanding. Here, three RGB-based models with good performance on MG recognition, namely I3D [8], TRN [98], and TSM [43] are selected for comparison. Similarly, two pose-based methods ST-GCN [90] and MS-G3D [46] are chosen. Those models follow the same configurations with the clip-level recognition. For example, in the task of clip-level MG recognition, TSM [43] divides a clip (input) into 8 segments and samples one RGB frame from each segment to form the input. Now, in the video-level emotion understanding, only the input is changed to a video. We report the performances of these networks in Table 3, we can see the TSM [43] and TRN [98] only obtain an emotion classification accuracy of 53 percentage by using the RGB frames as input. The ST-GCN [90] and MS-G3D [46] yield the similar results with the pose data as input. It is

Methods	Model + Modality	Accuracy
TRN [98]	CNN + RGB	0.53
TSM [43]		0.53
I3D [8]		<u>0.57</u>
ST-GCN [90]	GCN + Pose	0.50
MS-G3D [46]		0.55
U-S-VAE + LSTM	RNN + Micro-gesture	0.55
TSM [43] + LSTM		**0.60**

Table 3. Comparison of emotion understanding accuracy (%) with methods on iMiGUE dataset (best: bold, second best: underlined).

noted that I3D [8] achieves the best score among those models. I3D concatenates outputs from multi-parallel branches at the same level but with different resolutions, which can provide richer representations in emotion classification.

In order to experimentally confirm how micro-gesture influences the classification accuracy of emotions, we feed the probability vectors yield by TSM [43] into a RNN network. This is aiming to train an emotion understanding model via the results of clip-level MG recognition. More specifically, a vector (*e.g.*, an output of the Softmax layer) that represents the probability distributions of a list of potential outcomes (possible MG classes), will be fed in a three-layered LSTM network (TSM [43] with LSTM). After all vectors (clips) of a video are fed in, the labels of positive or negative (winning or losing of the match) can be used to train the network to understand the emotional states behind a series of micro-gestures in a holistic way. The reason why we select the TSM is because it obtains the best accuracy score in clip-level MG recognition. For comparison, the output vectors of proposed U-S-VAE are also utilized to train a similar emotion understanding network. In Table 3, we report the emotion understanding results of these models, according to the protocol (video-level) described in Sec. 5.1. We can observe and conclude that micro-gesture is helpful for the emotion understanding. TSM with LSTM (TSM + LSTM) can achieve the best score, U-S-VAE with a 35% MG recognition accuracy (U-S-VAE + LSTM) can beat most of the compared methods which further verifies that the MG-based analysis can benefit the final emotion understanding.

5.5. Analysis and Discussion

To test the generalization capability of U-S-VAE, we provide its performance on different datasets, such as the NTU RGB+D 60 [67], which is a large scale dataset commonly used for testing action/gesture models. NTU RGB+D contains 60 action categories collected from 40 subjects. In video capturing, each action is recorded simultaneously by three cameras at different horizontal angles. As such, not merely provided the commonly cross-subject (C-Sub) protocol, the authors of NTU RGB+D also recommended the cross-view (C-View) evaluation. We follow this convention and report the recognition accuracy (Top-1) of the two protocols. Here, the results of RGB-D-based unsupervised methods are presented, including Shuffle & Learn [52], and models of Luo *et al.* [49] and Li *et al.* [39]. These models rely on the depth information which are not available in iMiGUE so that we cannot evaluate their performance on our dataset. Moreover, two pose-based models, the LongT GAN [97] and P&C [77] are compared for evaluation analysis. Because the code of LongT GAN is not released, we cannot report its result on iMiGUE. Finally, for ablation studies on encoder with different number of BLSTM layer, U-S-VAE with 2-layers BLSTM (U-S-VAE 2L) and 3-layers (U-S-VAE 3L) are chosen for comparison. Ablation studies on different features for classification are carried, and we select the cell states of BLSTM to serve as the feature vectors for testing (U-S-VAE 1L w C). In Table 4, we report these experimental results.

Unsupervised Methods	Modality	iMiGUE	NTU RGB+D	
		C-Sub	C-View	C-Sub
Shuffle & Learn [52]	RGB-D	-	40.90	46.20
Luo *et al.* [49]		-	<u>53.20</u>	<u>61.40</u>
Li *et al.* [39]		-	**63.90**	**68.10**
LongT GAN [97]	Pose	-	48.10	39.10
P&C [77]		31.67	**76.10**	<u>50.70</u>
U-S-VAE 3L		30.85	25.46	22.50
U-S-VAE 2L		30.30	55.13	37.03
U-S-VAE 1L w C		<u>32.04</u>	44.57	36.11
U-S-VAE 1L		**32.43**	<u>64.88</u>	**50.96**

Table 4. Ablation study of U-S-VAE with different datasets (best: bold, second best: underlined).

From the results of Table 4, we can summarize several observations: 1) The mutual information (KL-divergence) plays a strong role in learning latent representation since our model (U-S-VAE 1L) can achieve better performance than P&C [77] on two datasets with cross-subject protocol. 2) The score of P&C in cross-view evaluation is higher than ours, this is because P&C has a pre-processing step to implement a view-invariant transformation [77]. Besides, P&C also has an additional feature-level auto-encoder can benefit the classification. 3) In ablation studies, U-S-VAE 1L can achieve the best performance on both iMiGUE and NTU RGB+D datasets. U-S-VAE with hidden states Z_T can beat U-S-VAE with the cell states (U-S-VAE 1L w C).

6. Conclusions

In this paper, we propose iMiGUE, a new dataset focusing on micro-gestures study. This work not merely investigates representative methods at the MG recognition level, but also attempt to understand the emotional states by using those MGs. We hope these efforts could facilitate new advances in the emotion AI field. In the future, more efforts will be put on studying relationships between MG groups and emotional states. Also, sophisticated models will be explored to find the latent mapping among emotions and MGs in a more holistic way.

References

[1] Min SH Aung, Sebastian Kaltwang, Bernardino Romera-Paredes, Brais Martinez, Aneesha Singh, Matteo Cella, Michel Valstar, Hongying Meng, Andrew Kemp, Moshen Shafizadeh, et al. The automatic detection of chronic pain-related expression: requirements, challenges and the multimodal emopain dataset. *IEEE Trans. Affect. Comput.*, 7(4):435–451, 2015. 1

[2] Hillel Aviezer, Yaacov Trope, and Alexander Todorov. Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science*, 338(6111):1225–1229, 2012. 2

[3] Roger E Axtell. *Gestures: the do's and taboos of body language around the wor ld*. 1991. 2

[4] Marian Stewart Bartlett, Gwen Littlewort, Mark Frank, Claudia Lainscsek, Ian Fasel, and Javier Movellan. Fully automatic facial action recognition in spontaneous behavior. In *Proc. IEEE Int. Conf. Auto. Face Gesture Recognit.*, pages 223–230. IEEE, 2006. 1

[5] Yoann Baveye, Emmanuel Dellandrea, Christel Chamaret, and Liming Chen. LIRIS-ACCEDE: A video database for affective content analysis. *IEEE Trans. Affect. Comput.*, 6(1):43–55, 2015. 3

[6] Judee K Burgoon, David B Buller, and William Gill Woodall. *Nonverbal communication: The unspoken dialogue*. Harpercollins College Division, 1989. 2

[7] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 7291–7299, 2017. 7

[8] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 6299–6308, 2017. 3, 7, 8

[9] Ginevra Castellano, Santiago D Villalba, and Antonio Camurri. Recognising human emotions from body movement and gesture dynamics. In *Int. Conf. Affect. Comput. Intell. Interact.*, pages 71–82. Springer, 2007. 2

[10] Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with shift graph convolutional network. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 183–192, 2020. 3, 7

[11] Adrian K Davison, Cliff Lansley, Nicholas Costen, Kevin Tan, and Moi Hoon Yap. SAMM: A spontaneous micro-facial movement dataset. *IEEE Trans. Affect. Comput.*, 9(1):116–129, 2016. 1, 5

[12] Abhinav Dhall, Roland Goecke, Shreya Ghosh, Jyoti Joshi, Jesse Hoey, and Tom Gedeon. From individual to group-level emotion recognition: Emotiw 5.0. In *Proc. ACM Int. Conf. Multimodal Interaction*, pages 524–528, 2017. 1

[13] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 2625–2634, 2015. 3

[14] Ellen Douglas-Cowie, Roddy Cowie, Ian Sneddon, Cate Cox, Orla Lowry, Margaret Mcrorie, Jean-Claude Martin, Laurence Devillers, Sarkis Abrilian, Anton Batliner, et al. The humaine database: Addressing the collection and annotation of naturalistic and induced emotional data. In *Int. Conf. Affect. Comput. Intell. Interact.*, pages 488–500. Springer, 2007. 2, 3

[15] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1110–1118, 2015. 3

[16] Paul Ekman. *Telling lies: Clues to deceit in the marketplace, politics, and marriage (revised edition)*. WW Norton & Company, 2009. 2, 4, 5

[17] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 6202–6211, 2019. 3

[18] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1933–1941, 2016. 3

[19] Nesrine Fourati and Catherine Pelachaud. Emilya: Emotional body expression in daily actions database. In *LREC*, pages 3486–3493, 2014. 3

[20] Mihai Gavrilescu. Recognizing emotions from videos by studying facial expressions, body postures and hand gestures. In *23rd Telecommunications Forum Telfor*, pages 720–723. IEEE, 2015. 3

[21] Serge Ginger. *Gestalt therapy: the art of contact*. Karnac Books, 2007. 2

[22] Donald Glowinski, Antonio Camurri, Gualtiero Volpe, Nele Dael, and Klaus Scherer. Technique for automatic emotion recognition by body gesture analysis. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, pages 1–6. IEEE, 2008. 3

[23] Alex Gorban, Haroon Idrees, Yu-Gang Jiang, A Roshan Zamir, Ivan Laptev, Mubarak Shah, and Rahul Sukthankar. THUMOS challenge: Action recognition with a large number of classes, 2015. 2

[24] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The "Something Something" video database for learning and evaluating visual common sense. In *Proc. IEEE Int. Conf. Comput. Vis.*, volume 1, 2017. 2

[25] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image Vis. Comput.*, 28(5):807–813, 2010. 1

[26] Hatice Gunes and Massimo Piccardi. A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In *Proc. IAPR Int. Conf. Pattern Recognit.*, volume 1, pages 1148–1153. IEEE, 2006. 2, 3

[27] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and imagenet? In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 6546–6555, 2018. 3, 7

[28] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3334–3342, 2015. 1

[29] Hanbyul Joo, Tomas Simon, Mina Cikara, and Yaser Sheikh. Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 10873–10883, 2019. 1

[30] Takeo Kanade, Jeffrey F Cohn, and Yingli Tian. Comprehensive database for facial expression analysis. In *Proc. IEEE Int. Conf. Auto. Face Gesture Recognit.*, pages 46–53. IEEE, 2000. 1

[31] Dahun Kim, Donghyeon Cho, and In So Kweon. Self-supervised video representation learning with space-time cubic puzzles. In *AAAI Conf. Artif. Intell.*, volume 33, pages 8545–8552, 2019. 3

[32] Michael Kipp and Jean-Claude Martin. Gesture and emotion: Can basic gestural form features discriminate emotions? In *Int. Conf. Affect. Comput. Intell. Interact. Workshops*, pages 1–8. IEEE, 2009. 3

[33] Andrea Kleinsmith, Nadia Bianchi-Berthouze, and Anthony Steed. Automatic recognition of non-acted affective postures. *IEEE Trans. Syst. Man Cybern. B Cybern.*, 41(4):1027–1038, 2011. 3

[34] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE Trans. Affect. Comput.*, 3(1):18–31, 2011. 1

[35] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *Int. J. Comput. Vision*, 127(6-7):907–929, 2019. 1

[36] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. HMDB: a large video database for human motion recognition. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 2556–2563. IEEE, 2011. 2

[37] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 667–676, 2017. 3

[38] Chaolong Li, Zhen Cui, Wenming Zheng, Chunyan Xu, and Jian Yang. Spatio-temporal graph convolution for skeleton based action recognition. In *AAAI Conf. Artif. Intell.*, 2018. 3

[39] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Unsupervised learning of view-invariant action representations. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 1254–1264, 2018. 3, 8

[40] Shuai Li, Wanqing Li, Chris Cook, Ce Zhu, and Yanbo Gao. Independently recurrent neural network (IndRNN): Building a longer and deeper rnn. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 5457–5466, 2018. 3

[41] Xiaobai Li, Tomas Pfister, Xiaohua Huang, Guoying Zhao, and Matti Pietikäinen. A spontaneous micro-expression database: Inducement, collection and baseline. In *Proc. IEEE Int. Conf. Auto. Face Gesture Recognit.*, pages 1–6. IEEE, 2013. 1, 5

[42] Yanghao Li, Cuiling Lan, Junliang Xing, Wenjun Zeng, Chunfeng Yuan, and Jiaying Liu. Online human action detection using joint classification-regression recurrent neural networks. In *Proc. Eur. Conf. Comput. Vis.*, pages 203–220. Springer, 2016. 3

[43] Ji Lin, Chuang Gan, and Song Han. TSM: Temporal shift module for efficient video understanding. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 7083–7093, 2019. 3, 7, 8

[44] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal LSTM with trust gates for 3D human action recognition. In *Proc. Eur. Conf. Comput. Vis.*, pages 816–833. Springer, 2016. 3

[45] Jun Liu, Gang Wang, Ping Hu, Ling-Yu Duan, and Alex C Kot. Global context-aware attention LSTM networks for 3D action recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1647–1656, 2017. 3

[46] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 143–152, 2020. 3, 7, 8

[47] Patrick Lucey, Jeffrey F Cohn, Kenneth M Prkachin, Patricia E Solomon, and Iain Matthews. Painful data: The unbc-mcmaster shoulder pain expression archive database. In *Proc. IEEE Int. Conf. Auto. Face Gesture Recognit.*, pages 57–64. IEEE, 2011. 1

[48] Yu Luo, Jianbo Ye, Reginald B Adams, Jia Li, Michelle G Newman, and James Z Wang. ARBEE: Towards automated recognition of bodily expression of emotion in the wild. *Int. J. Comput. Vision*, 128(1):1–25, 2020. 3

[49] Zelun Luo, Boya Peng, De-An Huang, Alexandre Alahi, and Li Fei-Fei. Unsupervised learning of long-term motion dynamics for videos. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 2203–2212, 2017. 3, 8

[50] B. Mahasseni and S. Todorovic. Regularizing long short term memory with 3D human-skeleton sequences for action recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 3054–3062. IEEE, June 2016. 3

[51] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Trans. Affect. Comput.*, 3(1):5–17, 2011. 1

[52] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *Proc. Eur. Conf. Comput. Vis.*, pages 527–544. Springer, 2016. 3, 8

[53] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(2):502–508, 2019. 2

[54] Joe Navarro and Marvin Karlins. *What every body is saying*. HarperCollins, 2016. 4, 5

[55] Fatemeh Noroozi, Dorota Kaminska, Ciprian Corneanu, Tomasz Sapinski, Sergio Escalera, and Gholamreza Anbarjafari. Survey on emotional body gesture recognition. *IEEE Trans. Affect. Comput.*, 2018. 2

[56] Maja Pantic, Michel Valstar, Ron Rademaker, and Ludo Maat. Web-based database for facial expression analysis. In *Proc. IEEE Int. Conf. Multimed. Expo.*, pages 317–321. IEEE, 2005. 1

[57] Barbara Pease and Allan Pease. *The definitive book of body language: The hidden meaning behind people's gestures and expressions*. Bantam, 2008. 4, 5

[58] Wei Peng, Xiaopeng Hong, Haoyu Chen, and Guoying Zhao. Learning graph convolutional network for skeleton-based human action recognition by neural searching. In *AAAI Conf. Artif. Intell.*, pages 2669–2676, 2020. 3, 7

[59] Athanasios Psaltis, Kyriaki Kaza, Kiriakos Stefanidis, Spyridon Thermos, Konstantinos C Apostolakis, Kosmas Dimitropoulos, and Petros Daras. Multimodal affective state recognition in serious games applications. In *IEEE Int. Conf. Imaging Sys. Tech.*, pages 435–439. IEEE, 2016. 3

[60] Hiranmayi Ranganathan, Shayok Chakraborty, and Sethuraman Panchanathan. Multimodal emotion recognition using deep learning architectures. In *IEEE Winter Conf. Appl. Comput. Vis.*, pages 1–9. IEEE, 2016. 3

[61] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *Proc. IEEE Int. Conf. Auto. Face Gesture Recognit.*, pages 1–8. IEEE, 2013. 1

[62] Marcus Rohrbach, Anna Rohrbach, Michaela Regneri, Sikandar Amin, Mykhaylo Andriluka, Manfred Pinkal, and Bernt Schiele. Recognizing fine-grained and composite activities using hand-centric features and script data. *Int. J. Comput. Vision*, 119(3):346–373, 2016. 2

[63] Sriparna Saha, Shreyasi Datta, Amit Konar, and Ramadoss Janarthanan. A study on emotion recognition from body gestures using Kinect sensor. In *Int. Conf. Signal Process. Commun.*, pages 056–060. IEEE, 2014. 3

[64] Konrad Schindler, Luc Van Gool, and Beatrice De Gelder. Recognizing emotions expressed by body pose: A biologically inspired neural model. *Neural Networks*, 21(9):1238–1246, 2008. 2, 3

[65] Björn Schuller, Michel Valstar, Florian Eyben, Gary McKeown, Roddy Cowie, and Maja Pantic. Avec 2011–the first international audio/visual emotion challenge. In *Int. Conf. Affect. Comput. Intell. Interact.*, pages 415–424. Springer, 2011. 1

[66] Björn Schuller, Michel Valster, Florian Eyben, Roddy Cowie, and Maja Pantic. Avec 2012: the continuous audio/visual emotion challenge. In *Proc. ACM Int. Conf. Multimodal Interact.*, pages 449–456, 2012. 1

[67] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. NTU RGB+D: A large scale dataset for 3D human activity analysis. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1010–1019, 2016. 2, 8

[68] Henglin Shi, Xin Liu, Xiaopeng Hong, and Guoying Zhao. Bidirectional long short-term memory variational autoencoder. In *BMVC*, page 165, 2018. 6, 7

[69] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 12026–12035, 2019. 3, 7

[70] Chenyang Si, Wentao Chen, Wei Wang, Liang Wang, and Tieniu Tan. An attention enhanced graph convolutional LSTM network for skeleton-based action recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1227–1236, 2019. 3

[71] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Proc. Eur. Conf. Comput. Vis.*, pages 510–526. Springer, 2016. 2

[72] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 568–576, 2014. 3

[73] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Trans. Affect. Comput.*, 3(1):42–55, 2011. 1

[74] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *Proc. AAAI Conf. Artif. Intell.*, pages 4263–4270, 2017. 3

[75] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *CRCV-TR-12-01*, 2012. 2

[76] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using LSTMs. In *Int. Conf. Mach. Learn.*, pages 843–852, 2015. 3, 6

[77] Kun Su, Xiulong Liu, and Eli Shlizerman. Predict & Cluster: Unsupervised skeleton based action recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 9631–9640, 2020. 3, 6, 7, 8

[78] Li Tao, Xueting Wang, and Toshihiko Yamasaki. Self-supervised video representation learning using inter-intra contrastive framework. In *Proc. ACM Int. Conf. Multimed.*, pages 2193–2201, 2020. 3

[79] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3D convolutional networks. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 4489–4497, 2015. 3, 7

[80] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 5552–5561, 2019. 3

[81] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 6450–6459, 2018. 3

[82] Michel Valstar and Maja Pantic. Induced disgust, happiness and surprise: an addition to the mmi facial expression

database. In *Proc. Int. Conf. Language Resources and Evaluation, Workshop EMOTION*, pages 65–70. Paris, France., 2010. 1

[83] Vivek Veeriah, Naifan Zhuang, and Guo-Jun Qi. Differential recurrent neural networks for action recognition. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 4041–4049, 2015. 3

[84] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. Social signal processing: Survey of an emerging domain. *Image Vis. Comput.*, 27(12):1743–1759, 2009. 1

[85] Jiangliu Wang, Jianbo Jiao, Linchao Bao, Shengfeng He, Yunhui Liu, and Wei Liu. Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 4006–4015, 2019. 3

[86] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(11):2740–2755, 2018. 3, 7

[87] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 7794–7803, 2018. 3

[88] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proc. Eur. Conf. Comput. Vis.*, pages 305–321, 2018. 3

[89] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 10334–10343, 2019. 3

[90] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI Conf. Artif. Intell.*, 2018. 3, 7, 8

[91] Wen-Jing Yan, Xiaobai Li, Su-Jing Wang, Guoying Zhao, Yong-Jin Liu, Yu-Hsin Chen, and Xiaolan Fu. Casme i-i: An improved spontaneous micro-expression database and the baseline evaluation. *PloS one*, 9(1):e86041, 2014. 1, 5

[92] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J Rosato. A 3d facial expression database for facial behavior research. In *Proc. IEEE Int. Conf. Auto. Face Gesture Recognit.*, pages 211–216. IEEE, 2006. 1

[93] Zhixuan Yu, Jae Shin Yoon, In Kyu Lee, Prashanth Venkatesh, Jaesik Park, Jihun Yu, and Hyun Soo Park. HUMBI: A Large Multiview Dataset of Human Body Expressions. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 2990–3000, 2020. 2

[94] Zitong Yu, Benjia Zhou, Jun Wan, Pichao Wang, Haoyu Chen, Xin Liu, Stan Z Li, and Guoying Zhao. Searching multi-rate and multi-modal temporal enhanced networks for gesture recognition. *arXiv preprint arXiv:2008.09412*, 2020. 3

[95] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 2117–2126, 2017. 3

[96] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, and Peng Liu. A high-resolution spontaneous 3d dynamic facial expression database. In *Proc. IEEE Int. Conf. Auto. Face Gesture Recognit.*, pages 1–6. IEEE, 2013. 1

[97] Nenggan Zheng, Jun Wen, Risheng Liu, Liangqu Long, Jianhua Dai, and Zhefeng Gong. Unsupervised representation learning with long-term dynamics for skeleton based action recognition. In *AAAI Conf. Artif. Intell.*, 2018. 3, 6, 7, 8

[98] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proc. Eur. Conf. Comput. Vis.*, pages 803–818, 2018. 3, 7, 8

[99] Wentao Zhu, Cuiling Lan, Junliang Xing, Wenjun Zeng, Yanghao Li, Li Shen, Xiaohui Xie, et al. Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks. In *Proc. AAAI Conf. Artif. Intell.*, volume 2, page 8, 2016. 3