Take an Emotion Walk: Perceiving Emotions from Gaits Using Hierarchical Attention Pooling and Affective Mapping

Uttaran Bhattacharya[0000−0003−2141−9276],[1], Christian Roncal[0000−0003−2772−5071],[1], Trisha Mittal[0000−0003−3558−6518],[1], Rohan Chandra[0000−0003−4843−6375],[1], Kyra Kapsaskis[0000−0003−2063−6156],[2], Kurt Gray[0000−0001−5816−2676],[2], Aniket Bera[0000−0002−0182−6985],[1], and Dinesh Manocha[0000−0001−7047−9801],[1]

[1] University of Maryland, College Park, MD 20742, USA
[2] University of North Carolina, Chapel Hill, NC 27599, USA

Abstract. We present an autoencoder-based semi-supervised approach to classify perceived human emotions from walking styles obtained from videos or motion-captured data and represented as sequences of 3D poses. Given the motion on each joint in the pose at each time step extracted from 3D pose sequences, we hierarchically pool these joint motions in a bottom-up manner in the encoder, following the kinematic chains in the human body. We also constrain the latent embeddings of the encoder to contain the space of psychologically-motivated affective features underlying the gaits. We train the decoder to reconstruct the motions per joint per time step in a top-down manner from the latent embeddings. For the annotated data, we also train a classifier to map the latent embeddings to emotion labels. Our semi-supervised approach achieves a mean average precision of 0.84 on the Emotion-Gait benchmark dataset, which contains both labeled and unlabeled gaits collected from multiple sources. We outperform current state-of-art algorithms for both emotion recognition and action recognition from 3D gaits by 7%–23% on the absolute. More importantly, we improve the average precision by 10%–50% on the absolute on classes that each makes up less than 25% of the labeled part of the Emotion-Gait benchmark dataset.

1 Introduction

Humans perceive others' emotions through verbal cues such as speech [53,29], text [63,12], and non-verbal cues such as eye-movements [57], facial expressions [19], tone of voice, postures [4], walking styles [34], etc. Perceiving others' emotions shapes people's interactions and experiences when performing tasks in collaborative or competitive environments [6]. Given this importance of perceived emotions in everyday lives, there has been a steady interest in developing

This project has been supported by ARO grant W911NF-19-1-0069.
Code and additional materials in project webpage: https://gamma.umd.edu/taew

automated techniques for perceiving emotions from various cues, with applications in affective computing, therapy, and rehabilitation [55], robotics [7,46], surveillance [3,52], audience understanding [66], and character generation [62].

While there are multiple non-verbal modalities for perceiving emotions, in our work, we only observe people's styles of walking or their gaits, extracted from videos or motion-captured data. Perceived emotion recognition using any non-verbal cues is considered to be a challenging problem in both psychology and AI, primarily because of the unreliability in the cues, arising from sources such as "mock" expressions [17], expressions affected by the subject's knowledge of an observer [20], or even self-reported emotions in certain scenarios [47]. However, gaits generally require less conscious initiation from the subjects and therefore tend to be more reliable cues. Moreover, studies in psychology have shown that observers were able to perceive the emotions of walking subjects by observing features such as arm swinging, stride lengths, collapsed upper body, etc. [42,34].

Gaits have been widely used in computer vision for many applications, including action recognition [69,59,37] and perceiving emotions [50,51,9,43]. However, there are a few key challenges in terms of designing machine learning methods for emotion recognition using gaits:

- Methods based on hand-crafted biomechanical features extracted from human gaits often suffer from low prediction accuracy [15,64].
- Fully deep-learned methods [50,9] rely heavily on sufficiently large sets of annotated data. Annotations are expensive and tedious to collect due to the variations in scales and motion trajectories [2], as well as the inherent subjectivity in perceiving emotions [9]. The benchmark dataset for emotion recognition, Emotion-Gait [9], has around $4,000$ data points of which more than 53% are unlabeled.
- Conditional generative methods are useful for data augmentation, but current methods can only generate data for short time periods [26,32] or with relatively low diversity [49,68,70,9].

On the other hand, acquiring poses from videos and MoCap data is cheap and efficient, leading to the availability of large-scale pose-based datasets [1,28,11,58]. Given the availability of these unlabeled gait datasets and the sparsity of gaits labeled with perceived emotions, there is a need to develop automatic methods that can utilize these datasets for emotion recognition.

Main Contributions: We present a semi-supervised network that accepts 3D pose sequences of human gaits extracted from videos or motion-captured data and predicts discrete perceived emotions, such as happy, angry, sad, and neutral. Our network consists of an unsupervised autoencoder coupled with a supervised classifier. The encoder in the unsupervised autoencoder hierarchically pools attentions on parts of the body. It learns separate intermediate feature representations for the motions on each of the human body parts (arms, legs, and torso) and then pools these features in a bottom-up manner to map them to the latent embeddings of the autoencoder. The decoder takes in these embeddings and reconstructs the motion on each joint of the body in a top-down manner.

We also perform affective mapping: we constrain the space of network-learned features to subsume the space of biomechanical affective features [54] expressed from the input gaits. These affective features contain useful information for distinguishing between different perceived emotions. Lastly, for the labeled data, our supervised classifier learns to map the encoder embeddings to the discrete emotion labels to complete the training process. To summarize, we contribute:

- **A semi-supervised network**, consisting of an autoencoder and a classifier, that are trained together to predict discrete perceived emotions from 3D pose sequences of gaits of humans.
- **A hierarchical attention pooling module** on the autoencoder to learn useful embeddings for unlabeled gaits, which improves the mean average precision (mAP) in classification by 1–17% on the absolute compared to state-of-the-art methods in both emotion recognition and action recognition from 3D gaits on the Emotion-Gait benchmark dataset.
- **Subsuming the affective features** expressed from the input gaits in the space of learned embeddings. This improves the mAP in classification by 7–23% on the absolute compared to state-of-the-art methods.

We observe the performance of our network improves linearly as more unlabeled data is used for training. More importantly, we report a 10–50% improvement on average precision on the absolute for emotion classes that have fewer than 25% labeled samples in the Emotion-Gait dataset [9].

2 Related Work

We briefly review prior work in classifying perceived emotions from gaits, as well as the related task of action recognition and generation from gaits.

Detecting Perceived Emotions from Gaits. Experiments in psychology have shown that observers were able to identify sadness, anger, happiness, and pride by observing gait features such as arm swinging, long strides, erect posture, collapsed upper body, etc. [44,39,42,34]. This, in turn, has led to considerable interest from both the computer vision and the affective computing communities in detecting perceived emotions from recorded gaits. Early works exploited different gait-based affective features to automatically detect perceived emotions [31,64,15,16]. More recent works combined these affective features with features learned from recurrent [50] or convolutional networks [9] to significantly improve classification accuracies.

Action Recognition and Generation. There are large bodies of recent work on both gait-based supervised action recognition [13,67,69,73,61,59,60,37], and gait-based unsupervised action generation [70,68,26,49]. These methods make use of RNNs or CNNs, including GCNs, or a combination of both, to achieve high classification accuracies on benchmark datasets such as Human3.6M [28], Kinetics [11], NTU RGB-D [58], and more. On top of the deep-learned networks, some methods have also leveraged the kinematic dependencies between joints and bones [59], dynamic movement-based features [60], and long-range temporal

dependencies [37], to further improve performance. A comprehensive review of recent methods in kinect-based action recognition is available in [65].

RNN and CNN-based approaches have been extended to semi-supervised classification as well [24,48,30,72]. These methods have also added constraints on limb proportions, movement constraints, and exploited the autoregressive nature of gait prediction to improve their generative and classification components.

Generative methods have also exploited full sequences of poses to directly generate full test sequences [71,10]. Other approaches have used constraints on limb movements [2], action-specific trajectories [26], and the structure and kinematics of body joints [49], to improve the naturalness of generated gaits.

In our work, we learn latent embeddings from gaits by exploiting the kinematic chains in the human body [5] in a hierarchical fashion. Inspired by prior works in emotion perception from gaits, we also constrain our embeddings to contain the space of affective features expressed from gaits, to improve our average precision, especially on the rarer classes.

3 Approach

Given both labeled and unlabeled 3D pose sequences for gaits, our goal is classify all the gaits into one or more discrete perceived emotion labels, such as happy, sad, angry, etc. We use a semi-supervised approach to achieve this, by combining an autoencoder with a classifier, as shown in Fig. 2. We denote the set of trainable parameters in the encoder, decoder, and classifier with θ, ψ, and ϕ respectively. We first extract the rotation per joint from the first time step to the current time step in the input sequences (details in Sec. 3.2). We then pass these rotations through the encoder, denoted with $f_\theta(\cdot)$, to transform the input rotations into features in the latent embedding space. We pass these latent features through the decoder, denoted with $f_\psi(\cdot)$, to generate reconstructions of the input rotations. If training labels are available, we also pass the encoded features through the fully-connected classifier network, denoted with $f_\phi(\cdot)$, to predict the probabilities of the labels. We define our overall loss function as

$$\mathcal{C}(\theta,\phi,\psi) = \sum_{i=1}^{M} I_y^{(i)} \mathcal{C}_{CL}\left(y^{(i)}, f_{\phi\circ\theta}\left(D^{(i)}\right)\right) + \mathcal{C}_{AE}\left(D^{(i)}, f_{\psi\circ\theta}\left(D^{(i)}\right)\right), \quad (1)$$

where $f_{b\circ a}(\cdot) := f_b(f_a(\cdot))$ denotes the composition of functions, $I_y^{(i)}$ is an indicator variable denoting whether the i^{th} data point has an associated label $y^{(i)}$, M is the number of gait samples, $\mathcal{C}_{CL}$ denotes the classifier loss detailed in Sec. 3.3, and $\mathcal{C}_{AE}$ denotes the autoencoder loss detailed in Sec. 3.4. For brevity of notation, we will henceforth use $\hat{y}^{(i)} := f_{\phi\circ\theta}\left(D^{(i)}\right)$ and $\hat{D}^{(i)} := f_{\psi\circ\theta}\left(D^{(i)}\right)$.

3.1 Representing Emotions

The Valence-Arousal-Dominance (VAD) model [41] is used for representing emotions in a continuous space. This model assumes three independent axes for valence, arousal, and dominance values, which collectively indicate an observed

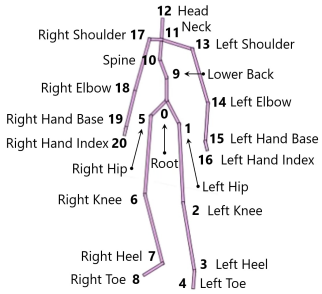

Fig. 1: **3D pose model.** The names and numbering of the 21 joints in the pose follow the nomenclature in the ELMD dataset [23].

Table 1: **Affective Features.** List of the 18 pose affective features that we use to describe the affective feature space for our network.

Angles between	shoulders at lower back
	hands at root
	left shoulder and hand at elbow
	right shoulder and hand at elbow
	head and left shoulder at neck
	head and right shoulder at neck
	head and left knee at root
	head and right knee at root
	left toe and right toe at root
	left hip and toe at knee
	right hip and toe at knee
Distance ratios between	left hand index (LHI) to neck and LHI to root
	right-hand index (RHI) to neck and RHI to root
	LHI to RHI and neck to root
	left toe to right toe and neck to root
Area(Δ) ratios between	Δ shoulders to lower back and Δ shoulders to root
	Δ hands to lower back and Δ hands to root
	Δ hand indices to neck and Δ toes to root

emotion. Valence indicates how pleasant (vs. unpleasant) the emotion is, arousal indicates how much the emotion is tied to high (vs. low) physiological intensity, and dominance indicates how much the emotion is tied to the assertion of high (vs. low) social status. For example, discrete emotion terms such as happy indicate high valence, medium arousal, and low dominance, angry indicate low valence, high arousal, and high dominance, and sad indicate low valence, low arousal, and low dominance.

On the other hand, these discrete emotion terms are easily understood by non-expert annotators and end-users. As a result, most existing datasets for supervised emotion classification consist of discrete emotion labels, and most supervised methods report performance on predicting these discrete emotions. In fact, discrete emotions can actually be mapped back to the VAD space through various known transformations [40,25]. Given these factors, we choose to use discrete emotion labels in our work as well. We also note that human observers have been reported to be most consistent in perceiving emotions varying primarily on the arousal axis, such as happy, sad, and angry [56,22]. Hence we work with the four emotions, happy, sad, angry, and neutral.

3.2 Representing the Data

Given the 3D pose sequences for gaits, we first obtain the rotations per joint per time step. We denote a gait as $G = \left\{\left(x_j^t, y_j^t, z_j^t\right)\right\}_{j=1,t=1}^{J,T}$, consisting of the 3D positions of J joints across T time steps. We denote the rotation of joint j from the first time step to time step t as $R_j^t \in \mathbb{SO}(3)$. We represent these rotations as unit quaternions $q_j^t \in \mathbb{H} \subset \mathbb{R}^4$, where $\mathbb{H}$ denotes the space of unit 4D quaternions. As stated in [49], quaternions are free of the gimbal-lock problem, unlike other common representations such as Euler angles or exponential maps [21]. We enforce the additional unit norm constraints for these quaternions when training our autoencoder. We represent the overall input to our network as $D^{(i)} := \left\{q_j^t\right\}_{j=1,t=1}^{J,T} \in \mathbb{H}^{J \times T}$.

3.3 Using Perceived Emotions and Constructing Classifier Loss

Observers' perception of emotions in others depends heavily influenced by their own personal, social, and cultural experiences, making emotion perception an inherently subjective task [56,34]. Consequently, we need to keep track of the differences in the perceptions of different observers. We do this by assigning multi-hot emotion labels to each input gait.

We assume that the given labeled gait dataset consists of C discrete emotion classes. The raw label vector $L^{(i)}$ for the i^{th} gait is a probability vector where the l^{th} element denotes the probability that the corresponding gait is perceived to have the l^{th} emotion. Specifically, we assume $L^{(i)} \in [0,1]^C$ to be given as $L^{(i)} = \begin{bmatrix} p_1 \; p_2 \; \dots p_C \end{bmatrix}^\top$, where p_l denotes the probability of the l^{th} emotion and $l = 1, 2, \dots C$. In practice, we compute the probability of each emotion for each labeled gait in a dataset as the fraction of annotators who labeled the gait with the corresponding emotion. To perform classification, we need to convert each element in $L^{(i)}$ to an assignment in $\{0,1\}$, resulting in the multi-hot emotion label $y^{(i)} \in \{0,1\}^C$. Taking into account the subjectivity in perceiving emotions, we set an element l in $y^{(i)}$ to 1 if $p_l > \frac{1}{C}$, *i.e.*, the l^{th} perceived emotion has more than a random chance of being reported, and 0 otherwise. Since our classification problem is multi-class (typically, $C > 2$) as well as multi-label (as we use multi-hot labels), we use the weighted multi-class cross-entropy loss

$$\mathcal{C}_{CL}\left(y^{(i)}, \hat{y}^{(i)}\right) := -\sum_{l=1}^{C} w_l \left(y_l\right)^{(i)} \log \left(\hat{y}_l\right)^{(i)} \tag{2}$$

for our classifier loss, where $\left(y_l\right)^{(i)}$ and $\left(\hat{y}_l\right)^{(i)}$ denote the l^{th} components of $y^{(i)}$ and $\hat{y}^{(i)}$, respectively. We also add per-class weights $w_l = e^{-p_l}$ to make the training more sensitive to mistakes on the rarer samples in the labeled dataset.

3.4 Using Affective Features and Constructing Autoencoder Loss

Our autoencoder loss consists of three constraints: affective loss, quaternion loss, and angle loss.

Affective loss. Prior studies in psychology report that a person's perceived emotions can be represented by a set of scale-independent gait-based affective features [15]. We consider the poses underlying the gaits to be made up of $J = 21$ joints (Fig. 1). Inspired by [50], we categorize the affective features as follows:

- *Angles* subtended by two joints at a third joint. For example, between the head and the neck (used to compute head tilt), the neck, and the shoulders (to compute slouching), root and thighs (to compute stride lengths), etc.
- *Distance ratios* between two pairs of joints. For example, the ratio between the distance from the hand to the neck, and that from the hand to the root (to compute arm swings).
- *Area ratios* formed by two triplets of joints. For example, the ratio of the area formed between the elbows and the neck and the area formed between

the elbows and the root (to compute slouching and arm swings). Area ratios can be viewed as amalgamations of the angle- and the distance ratio-based features used to supplement observations from these features.

We present the full list of the $\mathcal{A} = 18$ affective features we use in Table 1. We denote the set of affective features across all time steps for the i^{th} gait with $a^{(i)} \in \mathbb{R}^{\mathcal{A} \times T}$. We then constrain a subset of the embeddings learned by our encoder to map to these affective features. Specifically, we construct our embedding space to be $\mathbb{R}^{\mathcal{E} \times T}$ such that $\mathcal{E} \geq \mathcal{A}$. We then constrain the first $\mathcal{A} \times T$ dimensions of the embedding, denoted with $\hat{a}^{(i)}$ for the i^{th} gait, to match the corresponding affective features $a^{(i)}$. This gives our affective loss constraint:

$$\mathcal{L}_{\text{aff}}\left(a^{(i)}, \hat{a}^{(i)}\right) := \left\| a^{(i)} - \hat{a}^{(i)} \right\|^2. \tag{3}$$

We use affective constraints rather than providing affective features as input because there is no consensus on the universal set of affective features, especially due to cross-cultural differences [18,56]. Thus, we allow the encoder of our autoencoder to learn an embedding space using both data-driven features and our affective features, to improve generalizability.

Quaternion loss. The decoder for our autoencoder returns rotations per joint per time step as quaternions $\left(\hat{q}_j^t\right)^{(i)}$. We then constrain these quaternions to have unit norm:

$$\mathcal{L}_{\text{quat}}\left(\left(\hat{q}_j^t\right)^{(i)}\right) := \left(\left\|\left(\hat{q}_j^t\right)^{(i)}\right\| - 1\right)^2. \tag{4}$$

We apply this constraint instead of normalizing the decoder output, since individual rotations tend to be small, which leads the network to converge all its estimates to the unit quaternion.

Angle loss. This is the reconstruction loss for the autoencoder. We obtain it by converting the input and the output quaternions to the corresponding Euler angles and computing the mean loss between them:

$$\mathcal{L}_{\text{ang}}\left(D^{(i)}, \hat{D}^{(i)}\right) := \left\| (D_X, D_Y, D_Z)^{(i)} - \left(\hat{D}_X, \hat{D}_Y, \hat{D}_Z\right)^{(i)} \right\|_F^2 \tag{5}$$

where $(D_X, D_Y, D_Z)^{(i)} \in [0, 2\pi]^{3J \times T}$ and $\left(\hat{D}_X, \hat{D}_Y, \hat{D}_Z\right)^{(i)} \in [0, 2\pi]^{3J \times T}$ denotes the set of Euler angles for all the joints across all the time steps for input $D^{(i)}$ and output $\hat{D}^{(i)}$, respectively, and $\|\cdot\|_F$ denotes the Frobenius norm.

Combining Eqs. 3, 4 and 5, we write the autoencoder loss $\mathcal{C}_{AE}(\cdot, \cdot)$ as

$$\mathcal{C}_{AE}\left(D^{(i)}, \hat{D}^{(i)}\right) := \mathcal{L}_{\text{ang}}\left(D^{(i)}, \hat{D}^{(i)}\right) + \lambda_{\text{quat}}\mathcal{L}_{\text{quat}} + \lambda_{\text{aff}}\mathcal{L}_{\text{aff}} \tag{6}$$

where λ_{quat} and λ_{aff} are the regularization weights for the quaternion loss constraint and the affective loss constraint, respectively. To keep the scales of $\mathcal{L}_{\text{quat}}$ and $\mathcal{L}_{\text{aff}}$ consistent, we also scale all the affective features to lie in $[0, 1]$.

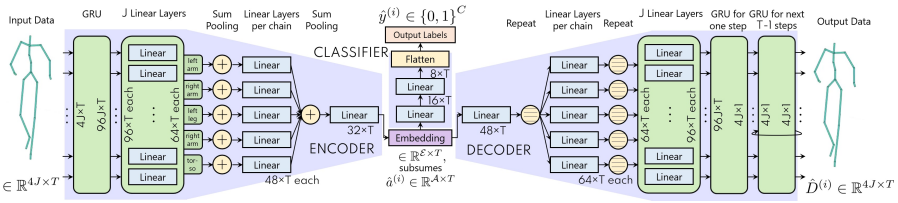

Fig. 2: **Our network for semi-supervised classification of discrete perceived emotions from gaits.** Inputs to the encoder are rotations on each joint at each time step, represented as 4D unit quaternions. The inputs are pooled bottom-up according to the kinematic chains of the human body. The embeddings at the end of the encoder are constrained to lie in the space of the mean affective features $\mathbb{R}^{\mathcal{A}}$. For labeled data, the embeddings are passed through the classifier to predict output labels. The linear layers in the decoder take in the embeddings and reconstruct the motion on each joint at a single time-step at the output of the first GRU. The second GRU in the decoder takes in the reconstructed joint motions at a single time step and predicts the joint motions for the next time step for $T-1$ steps.

4 Network Architecture and Implementation

Our network for semi-supervised classification of discrete perceived emotions from gaits, shown in Fig. 2, consists of three components, the encoder, the decoder, and the classifier. We describe each of these components and then summarize the training routine for our network.

4.1 Encoder with Hierarchical Attention Pooling

We first pass the sequences of joint rotations on all the joints through a two-layer Gated Recurrent Unit (GRU) to obtain feature representations for rotations at all joints at all time steps. We pass each of these representations through individual linear units. Following the kinematic chain of the human joints [5], we pool the linear unit outputs for the two arms, the two legs, and the torso in five separate linear layers. Thus, each of these five linear layers learns to focus attention on a different part of the human body. We then pool the outputs from these five linear layers into another linear layer, which, by construction, focuses attention on the motions of the entire body. For pooling, we perform vector addition as a way of composing the features at the different hierarchies.

Our encoder learns the hierarchy of the joint rotations in a bottom-up manner. We map the output of the last linear layer in the hierarchy to a feature representation in the embedding space of the encoder through another linear layer. In our case, the embedding space lies in $\mathbb{R}^{\mathcal{E}\times T}$ with $\mathcal{E}=32$, which subsumes the space of affective features $\mathbb{R}^{\mathcal{A}\times T}$ with $\mathcal{A}=18$, as discussed in Sec. 3.4.

4.2 Decoder with Hierarchical Attention Un-pooling

The decoder takes in the embedding from the encoder, repeats it five times for un-pooling, and passes the repeated features through five linear layers. The

outputs of these linear layers are features representing the reconstructions on the five parts, torso, two arms, and two legs. We repeat each of these features for un-pooling, and then collectively feed them into a GRU, which reconstructs the rotation on every joint at a single step. A subsequent GRU takes in the reconstructed joint rotations at a single time step and successively predicts the joint rotations for the next $T-1$ time steps.

4.3 Classifier for Labeled Data

Our classifier takes in the embeddings and passes it through a series of three linear layers, flattening the features between the second and the third linear layers. The output of the final linear layer, called "Output Labels" in Fig. 2, provides the label probabilities. To make predictions, we set the output for a class to be 1 if the label probability for that class was more than $\frac{1}{C}$, similar to the routine for constructing input labels discussed in Sec. 3.3.

4.4 Training Routine

We train using the Adam optimizer [33] with a learning rate of 0.001, which we decay by a factor of 0.999 per epoch. We apply the ELU activation [14] on all the linear layers except the output label layer, apply batch normalization [27] after every layer to reduce internal covariance-shift, and apply a dropout of 0.1 to prevent overfitting. On the second GRU in the decoder, which predicts joint rotations for T successive time steps, we use a curriculum schedule [8]. We start with a teacher forcing ratio of 1 on this GRU and at every epoch E, we decay the teacher forcing ratio by $\beta = 0.995$, *i.e.*, we either provide this GRU the input joint rotations with probability β^E, or the GRU's past predicted joint rotations with probability $1 - \beta^E$. Curriculum scheduling helps the GRU to gently transition from a teacher-guided prediction routine to a self-guided prediction routine, thereby expediting the training process.

We train our network for 500 epochs, which takes around 4 hours on an Nvidia GeForce GTX 1080Ti GPU with 12 GB memory. We use 80% of the available labeled data and all the unlabeled data for training our network, and validate its classification performance on a separate 10% of the labeled data. We keep the remaining 10% as the held-out test data. We also observed satisfactory performance when the weights λ_{quat} and λ_{aff} (in Eqn. 6) lie between 0.5 and 2.5. For our reported performances in Sec. 5.3, we used a value of 2 for both.

5 Results

We perform experiments with the Emotion-Gait benchmark dataset [9]. It consists of 3D pose sequences of gaits collected from a variety of sources and partially labeled with perceived emotions. We provide a brief description of the dataset in Sec. 5.1. We list the methods we compare with in Sec. 5.2. We then summarize the results of the experiments we performed with this dataset on all these methods in Sec. 5.3, and describe how to interpret the results in Sec. 5.4.

Table 2: **Average Precision scores.** Average precision (AP) per class and the mean average precision (mAP) over all the classes achieved by all the methods on the Emotion Gait dataset. Classes are Happy (H), Sad (S), Angry (A) and Neutral (N). Higher values are better. Bold indicates best, blue indicates second best.

Method	AP				mAP
	H	S	A	N	
STGCN [69]	0.98	0.83	0.42	0.18	0.61
DGNN [59]	0.98	0.88	0.73	0.37	0.74
MS-G3D [59]	0.98	0.88	0.75	0.44	0.76
LSTM Network [50]	0.96	0.84	0.62	0.51	0.73
STEP [9]	0.97	0.88	0.72	0.52	0.77
Our Method	**0.98**	**0.89**	**0.81**	**0.71**	**0.84**

Table 3: **Ablation studies.** Comparing average precisions of ablated versions of our method. HP denotes Hierarchical Pooling, AL denotes the Affective Loss constraint. AP, mAP, H, S, A, N are reused from Table 2. Bold indicates best, blue indicates second best.

Method	AP				mAP
	H	S	A	N	
With only labeled data, no AL or HP	0.92	0.81	0.51	0.42	0.67
With only labeled data, HP and no AL	0.93	0.81	0.63	0.49	0.72
With only labeled data, AL and no HP	0.96	0.86	0.70	0.51	0.76
With only labeled data, AL and HP	0.97	0.86	0.72	0.55	0.78
With all data, no AL or HP	0.94	0.83	0.55	0.48	0.70
With all data, HP and no AL	0.96	0.85	0.70	0.60	0.78
With all data, AL and no HP	0.97	0.87	0.76	0.65	0.81
With all data, AL and HP	**0.98**	**0.89**	**0.81**	**0.71**	**0.84**

5.1 Dataset

The Emotion-Gait dataset [9] consists of gaits collected from various sources of 3D pose sequence datasets, including BML [38], Human3.6M [28], ICT [45], CMU-MoCap [1] and ELMD [23]. To maintain a uniform set of joints for the pose models collected from diverse sources, we converted all the models in Emotion-Gait to the 21 joint pose model used in ELMD [23]. We clipped or zero-padded all input gaits to have 240 time steps, and downsampled it to contain every 5^{th} frame. We passed the resultant 48 time steps to our network, we have *i.e.*, $T = 48$. In total, the dataset has $3,924$ gaits of which $1,835$ have emotion labels provided by 10 annotators, and the remaining $2,089$ are not annotated. Around 58% of the labeled data have happy labels, 32% have sad labels, 23% have angry labels, and only 14% have neutral labels (more details on the project webpage).

Histograms of Affective Features. We show histograms of the mean values of 6 of the 18 affective features we use in Fig. 3. The means are taken across the $T = 48$ time steps in the input gaits and differently colored for inputs belonging to the different emotion classes as per the annotations. We count the inputs belonging to multiple classes once for every class they belong to. For different affective features, different sets of classes have a high overlap of values while values of the other classes are well-separated. For example, there is a significant overlap in the values of the distance ratio between right-hand index to the neck and right-hand index to the root (Fig. 3, bottom left) for gaits belonging to sad and angry classes, while the values of happy and neutral are distinct from these. Again, for gaits in happy and angry classes, there is a high overlap in the ratio of the area between hands to lower back and hands to root (Fig. 3, bottom right), while the corresponding values for gaits in neutral and sad classes are distinct from these. The affective features also support observations in psychology corresponding to perceiving emotions from gaits. For example, slouching is generally considered to be an indicator of sadness [42]. Correspondingly, we can observe that the values

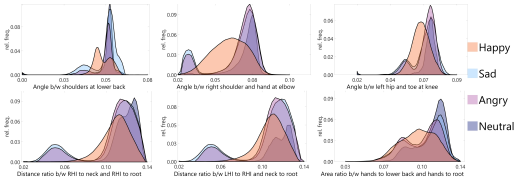

Fig. 3: **Conditional distribution of mean affective features.** Distributions of 6 of the 18 affective features, for the Emotion-Gait dataset, conditioned on the given classes Happy, Sad, Angry, and Neutral. Mean is taken across the number of time steps. We observe that the different classes have different distributions of peaks, indicating that these features are useful for distinguishing between perceived emotions.

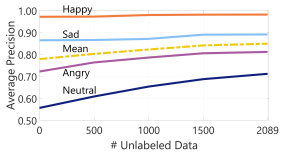

Fig. 4: **AP increases with adding unlabeled data.** AP achieved on each class, as well as the mean AP over the classes, increases linearly as we add more unlabeled data to train our network. The increment is most significant for the neutral class, which has the fewest labels in the dataset.

of the angle between the shoulders at the lower back (Fig. 3, top left) are lowest for sad gaits, indicating slouching.

5.2 Comparison Methods

We compare our method with the following state-of-the-art methods for both emotion recognition and action recognition from gaits. We choose to compare with action recognition methods because similar to these methods, we aim to learn a mapping from gaits to a set of labels (emotions instead of actions).

- **Emotion Recognition.** We compare with the network of [50], which combines affective features from gaits with features learned from an LSTM-based network taking pose sequences of gaits as input, to form hybrid feature vectors for classification. We also compare with STEP [9], which trains a spatial-temporal graph convolution-based network with gait inputs and affective features obtained from the gaits, and then fine-tunes the network with data generated from a graph convolution-based variational autoencoder.
- **Action Recognition.** We compare with recent state-of-the-art methods based on the spatial-temporal graph convolution network (STGCN) [69], the directed graph neural network (DGNN) [59], and the multi-scale graph convolutions with temporal skip connections (MS-G3D) [37]. STGCN computes spatial neighborhoods as per the bone structure of the 3D poses and temporal neighborhoods according to the instances of the same joints across time steps and performs convolutions based on these neighborhoods. DGNN computes directed acyclic graphs of the bone structure based on kinematic dependencies and trains a convolutional network with these graphs. MS-G3D performs multi-scale graph convolutions on the spatial dimensions and adds skip connections on the temporal dimension to model long-range dependencies for various actions.

For a fair comparison, we retrained all these networks from scratch with the labeled portion of the Emotion-Gait dataset, following their respective reported training parameters, and the same data split of $8:1:1$ as our network.

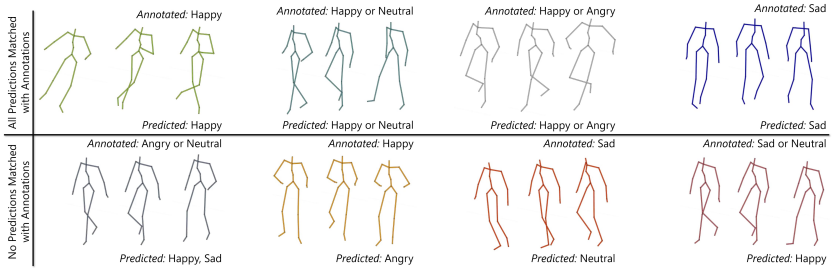

Fig. 5: **Comparing predictions with annotations.** The top row shows 4 gaits from the Emotion-Gait dataset where the predicted labels of our network exactly matched the annotated input labels. The bottom row shows 4 gaits where the predicted labels did not match any of the input labels. Each gait is represented by 3 poses in temporal sequence from left to right. We observe that most of the disagreements are between either happy and angry or between sad and neutral, which is consistent with general observations in psychology.

5.2.1 Evaluation Metric Since we deal with a multi-class, multi-label classification, we report the average precision (AP) achieved per class, which is the mean of the precision values across all values of recall between 0 and 1. We also report the mean AP, which is the mean of the APs achieved in all the classes.

5.3 Experiments

In our experiments, we ensured that the held-out test data were from sources different from the train and validation data in the Emotion-Gait dataset. We summarize the AP and the mean AP scores of all the methods in Table 2. Our method outperforms the next best method, STEP [9], by around 7% and outperforms the lowest-performing method, STGCN [69], by 23%, both on the absolute. We summarize additional results, including the interpretation of the data labels and our results in the VAD dimensions [41], on our project webpage.

Both the LSTM-based network and STEP consider per-frame affective features and inter-frame features such as velocities and rotations as inputs but do not explicitly model the dependencies between these two kinds of features. Our network, on the other hand, learns to embed a part of the features learned from joint rotations in the space of affective features. These embedded features, in turn, help our network predict the output emotion labels with more precision.

The action recognition methods STGCN, DGNN, and MS-G3D focus more on the movements of the leaf nodes, *i.e.*, hand indices, toes, and head. These nodes are useful for distinguishing between actions such as running and jumping but do not contain sufficient information to distinguish between perceived emotions.

Moreover, given the long-tail nature of the distribution of labels in the Emotion-Gait dataset (Sec. 5.1), all the methods we compare with have more than 0.95 AP in the happy and more than 0.80 AP in the sad classes, but perform much poorer on the angry and the neutral classes. Our method, by contrast, learns to map the joint motions to the affective features, which helps it achieve

around 10–50% better AP on the absolute on the angry and the neutral class while maintaining similarly high AP in the happy and the sad classes.

5.3.1 Ablation Studies We also perform ablation studies on our method to highlight the benefit of each of our three key components: using hierarchical pooling (HP) (Sec. 4.1), using the affective loss constraint (AL) (Eqn. 3), and using both labeled and unlabeled data in a semi-supervised manner (Eqn. 1). We summarize the observations of our ablation studies in Table 3.

First, we train our network only on the labeled dataset by removing the decoder part of our network and dropping the autoencoder loss from Eqn. 1. Without using either AL or HP, the network achieves an AP of 0.51 on angry and 0.42 on neutral, the two least populous classes. We call this our baseline network. Adding only the AL increases these two APs more from the baseline than adding only the HP. This is reasonable since hierarchical pooling helps the network learn generic differences in the pose sequences of different data, while the affective loss constraint helps the network to distinguish between pose structures specific to different perceived emotions. Adding both HP and AL increases the AP from the baseline even further. From these experiments, we can confirm that using either AL or HP improves the performance from the baseline, and their collective performance is better than their individual performances.

Next, we add in the decoder and use both labeled and unlabeled data for training our network, using the loss in Eqn. 1. Without both AL and HP, the network now achieves an AP of 0.55 on angry and 0.48 on neutral, showing appreciable improvements from the baseline. Also, as earlier, adding in only the AL shows more benefit on the network's performance than adding in only the HP. Specifically, adding in only the HP produces 1% absolute improvement in mean AP over STEP [9] (row 4 in Table 2) and 17% absolute improvement in mean AP over STGCN [69] (row 1 in Table 2). Adding in only the AL produces 4% absolute improvement in mean AP over STEP [9] (row 4 in Table 2) and 20% absolute improvement in mean AP over STGCN [69] (row 1 in Table 2). Adding in both, we get the final version of our network, which improves on the mean AP of STEP [9] by 7%, and the mean AP of STGCN [69] by 23%.

5.3.2 Performance Trend with Increasing Unlabeled Data In practice, it is relatively easy to collect unlabeled gaits from videos or using motion capture. We track the performance improvement of our network as we keep adding unlabeled data to our network, and summarize the results in Fig. 4. We observe that the mean AP improves linearly as we add more data. The trend does not indicate a saturation in AP for the angry and the neutral classes even after adding all the 2,089 unlabeled data. This suggests that the performance of our approach can increase further with more unlabeled data.

5.4 Interpretation of the Network Predictions

We show the qualitative results of our network in Fig. 5. The top row shows cases where the predicted labels for a gait exactly matched all the corresponding

annotated labels. We observe that the gaits with happy and angry labels in the annotation have more animated joint movements compared to the gaits with sad and neutral labels, which our network was able to successfully learn from the affective features. This is in line with established studies in psychology [41], which show that both happy and angry emotions lie high on the arousal scale, whereas neutral and sad are lower on the arousal scale. The bottom row shows cases where the predicted labels for a gait did not match any of the annotated labels. We notice that most disagreements arise either between sad and neutral labels or between happy and angry labels. This again follows the observation that both happy and angry gaits, higher on the arousal scale, often have more exaggerated joint movements, while both sad and neutral gaits, lower on the arousal scale, often have more reserved joint movements. There are also disagreements between happy and neutral labels for some gaits, where the joint movements in the happy gaits are not as exaggerated.

We also make an important distinction between the multi-hot input labels provided by human annotators and the multi-hot predictions of our network. The input labels capture the subjectivity in human perception, where different human observers perceive different emotions from the same gait based on their own biases and prior experiences [56]. The network, on the other hand, indicates that the emotion perceived from a particular gait data best fits one of the labels it predicts for that data. For example, in the third result from left on the top row in Fig. 5, five of the ten annotators perceived the gait to be happy, three perceived it to be angry, and the remaining two perceived it to be neutral. Following our annotations procedure in Sec. 4.3, we annotated this gait as an instance of both happy and angry. Given this gait, our network predicts a multi-hot label with 1's for happy and angry and 0's for neutral and sad. This indicates that the network successfully focused on the arousal in this gait, and found the emotion perceived from it to best match either happy or angry, and not match neutral and sad. We present more such results on our project webpage.

6 Limitations and Future Work

Our work has some limitations. First, we consider only discrete emotions of people and do not explicitly map these to the underlying continuous emotion space given by the VAD model [41]. Even though discrete emotions are presumably easier to work with for non-expert end-users, we plan to extend our method to work in the continuous space of emotions, *i.e.*, given a gait, our network regresses it to a point in the VAD space that indicates the perceived emotions.

Second, our network only looks at gait-based features to predict perceived emotions. In the future, we plan to combine these features with cues from other modalities such as facial expressions and body gestures, that are often expressed in tandem with gaits, to develop more robust emotion perception methods. We also plan to look at higher-level information, such as the presence of other people in the vicinity, background context, etc. that are known to influence a person's emotions [35,36] to further sophisticate the performance of our network.

References

1. Cmu graphics lab motion capture database. http://mocap.cs.cmu.edu/ (2018)
2. Ahsan, U., Sun, C., Essa, I.: Discrimnet: Semi-supervised action recognition from videos using generative adversarial networks. arXiv preprint arXiv:1801.07230 (2018)
3. Arunnehru, J., Geetha, M.K.: Automatic human emotion recognition in surveillance video. In: ITSPMS, pp. 321–342. Springer (2017)
4. Babu, A.R., Rajavenkatanarayanan, A., Brady, J.R., Makedon, F.: Multimodal approach for cognitive task performance prediction from body postures, facial expressions and eeg signal. In: Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data. p. 2. ACM (2018)
5. Badler, N.I., Phillips, C.B., Webber, B.L.: Simulating humans: computer graphics animation and control. Oxford University Press (1993)
6. Barrett, L.F.: How emotions are made: The secret life of the brain. Houghton Mifflin Harcourt (2017)
7. Bauer, A., et al.: The autonomous city explorer: Towards natural human-robot interaction in urban environments. IJSR **1**(2), 127–140 (2009)
8. Bengio, S., Vinyals, O., Jaitly, N., Shazeer, N.: Scheduled sampling for sequence prediction with recurrent neural networks. In: Advances in Neural Information Processing Systems. pp. 1171–1179 (2015)
9. Bhattacharya, U., Mittal, T., Chandra, R., Randhavane, T., Bera, A., Manocha, D.: Step: Spatial temporal graph convolutional networks for emotion perception from gaits. In: AAAI. pp. 1342–1350 (2020)
10. Cai, H., Bai, C., Tai, Y.W., Tang, C.K.: Deep video generation, prediction and completion of human action sequences. In: ECCV. pp. 366–382 (2018)
11. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
12. Chen, Y., Hou, W., Cheng, X., Li, S.: Joint learning for emotion classification and emotion cause detection. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 646–651 (2018)
13. Choutas, V., Weinzaepfel, P., Revaud, J., Schmid, C.: Potion: Pose motion representation for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7024–7033 (2018)
14. Clevert, D.A., Unterthiner, T., Hochreiter, S.: Fast and accurate deep network learning by exponential linear units (elus). arXiv preprint arXiv:1511.07289 (2015)
15. Crenn, A., Khan, R.A., Meyer, A., Bouakaz, S.: Body expression recognition from animated 3d skeleton. In: IC3D. pp. 1–7. IEEE (2016)
16. Daoudi, M., Berretti, S., Pala, P., Delevoye, Y., Del Bimbo, A.: Emotion recognition by body movement representation on the manifold of symmetric positive definite matrices. In: ICIAP. pp. 550–560. Springer (2017)
17. Ekman, P., Friesen, W.V.: Head and body cues in the judgment of emotion: A reformulation. Perceptual and motor skills (1967)
18. Ekman, P., Friesen, W.V.: The repertoire of nonverbal behavior: Categories, origins, usage, and coding. semiotica **1**(1), 49–98 (1969)
19. Fabian Benitez-Quiroz, C., Srinivasan, R., Martinez, A.M.: Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In: CVPR (June 2016)

20. Fernández-Dols, J.M., Ruiz-Belda, M.A.: Expression of emotion versus expressions of emotions. In: Everyday conceptions of emotion, pp. 505–522. Springer (1995)
21. Grassia, F.S.: Practical parameterization of rotations using the exponential map. Journal of graphics tools **3**(3), 29–48 (1998)
22. Gross, M.M., Crane, E.A., Fredrickson, B.L.: Effort-shape and kinematic assessment of bodily expression of emotion during gait. Human movement science **31**(1), 202–221 (2012)
23. Habibie, I., Holden, D., Schwarz, J., Yearsley, J., Komura, T.: A recurrent variational autoencoder for human motion synthesis. In: Proceedings of the British Machine Vision Conference (BMVC) (2017)
24. Harvey, F.G., Roy, J., Kanaa, D., Pal, C.: Recurrent semi-supervised classification and constrained adversarial generation with motion capture data. Image and Vision Computing **78**, 42–52 (2018)
25. Hoffmann, H., Scheck, A., Schuster, T., Walter, S., Limbrecht, K., Traue, H.C., Kessler, H.: Mapping discrete emotions into the dimensional space: An empirical approach. In: 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC). pp. 3316–3320. IEEE (2012)
26. Holden, D., Saito, J., Komura, T.: A deep learning framework for character motion synthesis and editing. ACM Transactions on Graphics (TOG) **35**(4), 138 (2016)
27. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
28. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE transactions on pattern analysis and machine intelligence **36**(7), 1325–1339 (2013)
29. Jacob, A., Mythili, P.: Prosodic feature based speech emotion recognition at segmental and supra segmental levels. In: SPICES. pp. 1–5. IEEE (2015)
30. Kanazawa, A., Zhang, J.Y., Felsen, P., Malik, J.: Learning 3d human dynamics from video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5614–5623 (2019)
31. Karg, M., Kuhnlenz, K., Buss, M.: Recognition of affect based on gait patterns. Cybernetics **40**(4), 1050–1061 (2010)
32. Khodabandeh, M., Reza Vaezi Joze, H., Zharkov, I., Pradeep, V.: Diy human action dataset generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 1448–1458 (2018)
33. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
34. Kleinsmith, A., Bianchi-Berthouze, N.: Affective body expression perception and recognition: A survey. IEEE Transactions on Affective Computing **4**(1), 15–33 (2013)
35. Kosti, R., Alvarez, J., Recasens, A., Lapedriza, A.: Context based emotion recognition using emotic dataset. IEEE transactions on pattern analysis and machine intelligence (2019)
36. Lee, J., Kim, S., Kim, S., Park, J., Sohn, K.: Context-aware emotion recognition networks. arXiv preprint arXiv:1908.05913 (2019)
37. Liu, Z., Zhang, H., Chen, Z., Wang, Z., Ouyang, W.: Disentangling and unifying graph convolutions for skeleton-based action recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
38. Ma, Y., Paterson, H.M., Pollick, F.E.: A motion capture library for the study of identity, gender, and emotion perception from biological motion. Behavior research methods **38**(1), 134–141 (2006)

39. Meeren, H.K., van Heijnsbergen, C.C., de Gelder, B.: Rapid perceptual integration of facial expression and emotional body language. Proceedings of NAS **102**(45), 16518–16523 (2005)
40. Mehrabian, A.: Analysis of the big-five personality factors in terms of the pad temperament model. Australian journal of Psychology **48**(2), 86–92 (1996)
41. Mehrabian, A., Russell, J.A.: An approach to environmental psychology. the MIT Press (1974)
42. Michalak, J., Troje, N.F., Fischer, J., Vollmar, P., Heidenreich, T., Schulte, D.: Embodiment of sadness and depressiongait patterns associated with dysphoric mood. Psychosomatic Medicine **71**(5), 580–587 (2009)
43. Mittal, T., Guhan, P., Bhattacharya, U., Chandra, R., Bera, A., Manocha, D.: Emoticon: Context-aware multimodal emotion recognition using frege's principle. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14234–14243 (2020)
44. Montepare, J.M., Goldstein, S.B., Clausen, A.: The identification of emotions from gait information. Journal of Nonverbal Behavior **11**(1), 33–42 (1987)
45. Narang, S., Best, A., Feng, A., Kang, S.h., Manocha, D., Shapiro, A.: Motion recognition of self and others on realistic 3d avatars. Computer Animation and Virtual Worlds **28**(3-4), e1762 (2017)
46. Narayanan, V., Manoghar, B.M., Dorbala, V.S., Manocha, D., Bera, A.: Proxemo: Gait-based emotion learning and multi-view proxemic fusion for socially-aware robot navigation. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2020. IEEE (2020)
47. Nisbett, R.E., Wilson, T.D.: Telling more than we can know: Verbal reports on mental processes. Psychological review **84**(3), 231 (1977)
48. Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M.: 3d human pose estimation in video with temporal convolutions and semi-supervised training. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7753–7762 (2019)
49. Pavllo, D., Grangier, D., Auli, M.: Quaternet: A quaternion-based recurrent model for human motion. arXiv preprint arXiv:1805.06485 (2018)
50. Randhavane, T., Bera, A., Kapsaskis, K., Bhattacharya, U., Gray, K., Manocha, D.: Identifying emotions from walking using affective and deep features. arXiv preprint arXiv:1906.11884 (2019)
51. Randhavane, T., Bera, A., Kapsaskis, K., Sheth, R., Gray, K., Manocha, D.: Eva: Generating emotional behavior of virtual agents using expressive features of gait and gaze. In: ACM Symposium on Applied Perception 2019. pp. 1–10 (2019)
52. Randhavane, T., Bhattacharya, U., Kapsaskis, K., Gray, K., Bera, A., Manocha, D.: The liar's walk: Detecting deception with gait and gesture. arXiv preprint arXiv:1912.06874 (2019)
53. Rao, K.S., Koolagudi, S.G., Vempada, R.R.: Emotion recognition from speech using global and local prosodic features. International Journal of Speech Technology (2013)
54. Riggio, H.R.: Emotional expressiveness. Encyclopedia of Personality and Individual Differences (2017)
55. Rivas, J.J., Orihuela-Espina, F., Sucar, L.E., Palafox, L., Hernández-Franco, J., Bianchi-Berthouze, N.: Detecting affective states in virtual rehabilitation. In: Proceedings of the 9th International Conference on Pervasive Computing Technologies for Healthcare. pp. 287–292. ICST (Institute for Computer Sciences, Social-Informatics and (2015)

56. Roether, C.L., Omlor, L., Christensen, A., Giese, M.A.: Critical features for the perception of emotion from gait. Journal of vision **9**(6), 15–15 (2009)
57. Schurgin, M., Nelson, J., Iida, S., Ohira, H., Chiao, J., Franconeri, S.: Eye movements during emotion recognition in faces. Journal of vision **14**(13), 14–14 (2014)
58. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1010–1019 (2016)
59. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Skeleton-based action recognition with directed graph neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7912–7921 (2019)
60. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 12026–12035 (2019)
61. Si, C., Chen, W., Wang, W., Wang, L., Tan, T.: An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1227–1236 (2019)
62. Starke, S., Zhang, H., Komura, T., Saito, J.: Neural state machine for character-scene interactions. ACM Transactions on Graphics **38**(6) (7 2019)
63. Strapparava, C., Mihalcea, R.: Learning to identify emotions in text. In: Proceedings of the 2008 ACM symposium on Applied computing. pp. 1556–1560. ACM (2008)
64. Venture, G., Kadone, H., Zhang, T., Grèzes, J., Berthoz, A., Hicheur, H.: Recognizing emotions conveyed by human gait. IJSR **6**(4), 621–632 (2014)
65. Wang, L., Huynh, D.Q., Koniusz, P.: A comparative review of recent kinect-based action recognition algorithms. arXiv preprint arXiv:1906.09955 (2019)
66. Wu, Z., Fu, Y., Jiang, Y.G., Sigal, L.: Harnessing object and scene semantics for large-scale video understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3112–3121 (2016)
67. Yan, A., Wang, Y., Li, Z., Qiao, Y.: Pa3d: Pose-action 3d machine for video recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7922–7931 (2019)
68. Yan, S., Li, Z., Xiong, Y., Yan, H., Lin, D.: Convolutional sequence generation for skeleton-based action synthesis. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4394–4402 (2019)
69. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: AAAI (2018)
70. Yang, C., Wang, Z., Zhu, X., Huang, C., Shi, J., Lin, D.: Pose guided human video generation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 201–216 (2018)
71. Yang, C., Wang, Z., Zhu, X., Huang, C., Shi, J., Lin, D.: Pose guided human video generation. In: ECCV. pp. 201–216 (2018)
72. Zhang, J.Y., Felsen, P., Kanazawa, A., Malik, J.: Predicting 3d human dynamics from video. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 7114–7123 (2019)
73. Zhang, S., Yang, Y., Xiao, J., Liu, X., Yang, Y., Xie, D., Zhuang, Y.: Fusing geometric features for skeleton-based action recognition using multilayer lstm networks. IEEE Transactions on Multimedia **20**(9), 2330–2343 (2018)