Identifying Emotions from Walking Using Affective and Deep Features

Tanmay Randhavane, Uttaran Bhattacharya, Kyra Kapsaskis, Kurt Gray, Aniket Bera, and Dinesh Manocha

Abstract—We present a new data-driven model and algorithm to identify the perceived emotions of individuals based on their walking styles. Given an RGB video of an individual walking, we extract his/her walking gait in the form of a series of 3D poses. Our goal is to exploit the gait features to classify the emotional state of the human into one of four emotions: happy, sad, angry, or neutral. Our perceived emotion recognition approach uses deep features learned via LSTM on labeled emotion datasets. Furthermore, we combine these features with affective features computed from gaits using posture and movement cues. These features are classified using a Random Forest Classifier. We show that our mapping between the combined feature space and the perceived emotional state provides 80.07% accuracy in identifying the perceived emotions. In addition to classifying discrete categories of emotions, our algorithm also predicts the values of perceived valence and arousal from gaits. We also present an "*EWalk* (Emotion Walk)" dataset that consists of videos of walking individuals with gaits and labeled emotions. To the best of our knowledge, this is the first gait-based model to identify perceived emotions from videos of walking individuals.

1 Introduction

Emotions play a large role in our lives, defining our experiences and shaping how we view the world and interact with other humans. Perceiving the emotions of social partners helps us understand their behaviors and decide our actions towards them. For example, people communicate very differently with someone they perceive to be angry and hostile than they do with someone they perceive to be calm and content. Furthermore, the emotions of unknown individuals can also govern our behavior, (e.g., emotions of pedestrians at a road-crossing or emotions of passengers in a train station). Because of the importance of perceived emotion in everyday life, automatic emotion recognition is a critical problem in many fields such as games and entertainment, security and law enforcement, shopping, human-computer interaction, human-robot interaction, etc.

Humans perceive the emotions of other individuals using verbal and non-verbal cues. Robots and AI devices that possess speech understanding and natural language processing capabilities are better at interacting with humans. Deep learning techniques can be used for speech emotion recognition and can facilitate better interactions with humans [1].

Understanding the perceived emotions of individuals using non-verbal cues is a challenging problem. The non-verbal cues humans use to perceive emotions include both facial expressions and body movements. With a more extensive availability of data, considerable research has focused on using facial expressions to understand emotion [2]. However, recent studies in psychology question the communicative purpose of facial expressions and doubt the quick, automatic process of perceiving emotions from these expressions [3]. There are situations when facial expressions can be unreliable, such as with "mock" or "referential expressions" [4]. Facial expressions can also be unreliable depending on whether an audience is present [5].

Research has shown that body expressions are also crucial in emotion expression and perception [6]. For example,

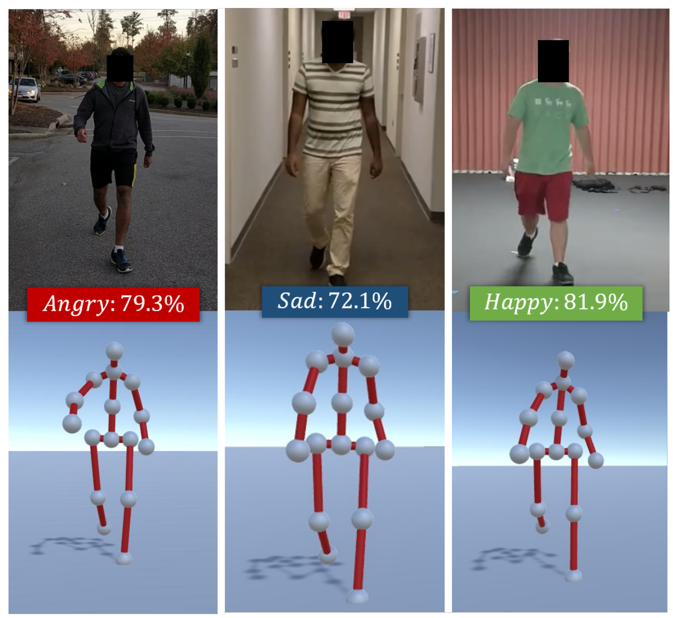


Fig. 1: **Identiying Perceived Emotions**: We present a novel algorithm to identify the perceived emotions of individuals based on their walking styles. Given an RGB video of an individual walking (top), we extract his/her walking gait as a series of 3D poses (bottom). We use a combination of deep features learned via an LSTM and affective features computed using posture and movement cues to then classify into basic emotions (e.g., happy, sad, etc.) using a Random Forest Classifier.

- T. Randhavane is with the Department of Computer Science, University of North Carolina, Chapel Hill, NC, 27514. E-mail: tanmay@cs.unc.edu
- A. Bera, K. Kapsaskis and K. Gray are with University of North Carolina at Chapel Hill.
- U. Bhattacharya and D. Manocha are with University of Maryland at College Park.

when presented with bodies and faces that expressed either anger or fear (matched correctly with each other or as mismatched compound images), observers are biased towards body expression [7]. Aviezer et al.'s study [8] on positive/negative valence in tennis players showed that faces alone were not a diagnostic predictor of valence, but the body alone or the face and body together can be predictive.

Specifically, body expression in walking, or an individual's gait, has been proven to aid in the perception of emotions. In an early study by Montepare et al. [9], participants were able to identify sadness, anger, happiness, and pride at a significant rate by observing affective features such as increased arm swinging, long strides, a greater foot landing force, and erect posture. Specific movements have also been correlated with specific emotions. For example, sad movements are characterized by a collapsed upper body and low movement activity [10]. Happy movements have a faster pace with more arm swaying [11].

Main Results: We present an automatic emotion identification approach for videos of walking individuals (Figure 1). We classify walking individuals from videos into happy, sad, angry, and neutral emotion categories. These emotions represent emotional states that last for an extended period and are more abundant during walking [12]. We extract gaits from walking videos as 3D poses. We use an LSTM-based approach to obtain deep features by modeling the long-term temporal dependencies in these sequential 3D human poses. We also present spatiotemporal affective body features representing the posture and movement of walking humans. We combine these affective features with LSTM-based deep features and use a Random Forest Classifier to classify them into four categories of emotion. We observe an improvement of 13.85% in the classification accuracy over other gait-based perceived emotion classification algorithms.

We also present a new dataset, *"Emotion Walk (EWalk),"* which contains videos of individuals walking in both indoor and outdoor locations. Our dataset consists of 1384 gaits and the perceived emotions labeled using Mechanical Turk.

Some of the novel components of our work include:

1. A novel data-driven mapping between the affective features extracted from a walking video and the perceived emotions.

2. A novel emotion identification algorithm that combines affective features and deep features, obtaining 80.07% accuracy.

3. A new public domain dataset, *EWalk*, with walking videos, gaits, and labeled emotions.

The rest of the paper is organized as follows. In Section 2, we review the related work in the fields of emotion modeling, bodily expression of emotion, and automatic recognition of emotion using body expressions. In Section 3, we give an overview of our approach and present the affective features. We provide the details of our LSTM-based approach to identifying perceived emotions from walking videos in Section 4. We compare the performance of our method with state-of-the-art methods in Section 5. We present the *EWalk* dataset in Section 6.

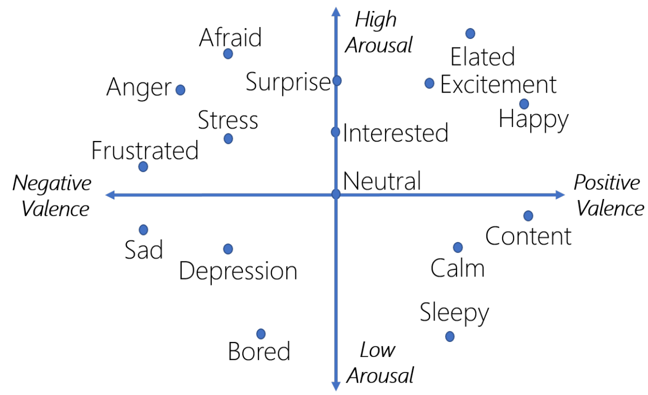


Fig. 2: All discrete emotions can be represented by points on a 2D affect space of Valence and Arousal [13], [14].

2 Related Work

In this section, we give a brief overview of previous works on emotion modeling, emotion expression using body posture and movement, and automatic emotion recognition.

2.1 Emotion Modeling

In previous literature, emotions were modeled as discrete categories or as points in a continuous space of affective dimensions. In the continuous space representation, emotions are treated as points on a 2D space of arousal and valence dimensions [14]. Sometimes an additional dimension of action tendency [6] or dominance [15] is also used to represent emotions in a 3D space. Mikels et al. [16] and Morris [17] investigated a mapping between the continuous model and discrete emotional models. For example, discrete emotions of anger, happiness, and pride are related to high arousal, whereas sadness, relief, and contentment are related to low arousal (Figure 2). Many affective computing approaches have used biometric signals for detecting affective dimensions of arousal and valence [18], [19]. In this paper, we identify four discrete emotions (happy, angry, sad, and neutral) from the walking motions and gaits while also identifying the values of valence and arousal. A combination of these four emotions can be used to represent other emotions [16].

2.2 Body Expression of Emotion

Affect expression combines verbal and nonverbal communication styles, including eye gaze and body expressions in addition to facial expressions, intonation, and other cues [20]. Facial expressions–like any element of emotional communication–do not exist in isolation. There is no denying that in certain cases such as with actors and caricatures, it is appropriate to assume affect based on the visual cues from the face, however, in day-to-day life, this doesn't account for body expressions. More specifically, the way a person walks, or their gait, has been proven to aid in the perception of that persons emotion.

The ability for body joints to express emotions has been studied in two pathways: posture and movement. Studies involving signals from posture and movement determined that both are used in the perception of emotion [21]. Expression of emotion in various activities such as knocking [22], dancing [23], playing musical instruments, walking [24], etc. has also been studied [6]. Kleinsmith et al. [6] identified affective dimensions that human observers use when discriminating between postures. Roether et al. [21] used a systematic approach and Omlor and Giese [25] identified spatiotemporal features that are specific to different emotions in gaits. Our approach is inspired by these studies and uses a combination of posture and movement features (i.e. affective features) to identify the perceived emotions from walking gaits.

2.3 Emotion Perception

It is important to distinguish between perceived emotions and actual emotions as we discuss the perception of emotions. One of the most obvious cues to another person's emotional state is his or her self-report [26]. People typically have access to their internal feelings [27] and when a person says that they are sad or disgusted they are conveying important information [27]. However, self-reports are not always available; for example, when people observe others remotely (e.g., via cameras) they do not have the ability to ask about their emotional state. Self-reports can also often be misleading, as when people report feeling "fine" when they are clearly feeling depressed, or when people try to deceive law enforcement agencies by conveying a false sense of calm after wrongdoing [28], [29].

Additionally, in our daily lives, we do not have access to factual information regarding the emotional states of others. We only have the information that we gather from social perception cues and these help guide us through our social interactions. Therefore, in this paper, we focus on these cues of social perception of emotions instead of using self-reported measures.

2.4 Automatic Emotion Recognition

With the increasing availability of technologies that capture body expression, there is considerable work on the automatic recognition of emotions from body expressions. Most works use a feature-based approach to identify emotion from body expressions automatically. These features are either extracted using purely statistical techniques or using techniques that are inspired by psychological studies. Some approaches focused on specific activities such as dancing, knocking [22], walking [24], games [6], etc., whereas some other approaches used a more generalized approach [30], [31]. Some approaches have combined both facial and body expressions [7]. Some approaches found emotions in body expressions with the help of neutral expressions [21]. Crenn et al. [32] generated neutral movements from expressive movements and then identified the emotion in the expressive movement. Karg et al. [24] examined gait information for person-dependent affect recognition using motion capture data of a single walking stride. Wang et al. [33] used a Kinect to capture the gaits and identify whether an individual is angry, happy, or neutral using four walk cycles. As is the case for most of these techniques, our approach is also founded on using psychology-based features to identify emotion in walking movements without using a neutral movement in real-time.

3 Approach

In this section, we describe our algorithm (Figure 3) for identifying perceived emotions from RGB videos.

3.1 Notation

For our formulation, we represent a human with a set of 16 joints, as shown in [34] (Figure 4). A pose $P \in \mathbb{R}^{48}$ of a human is a set of 3D positions of each joint $j_i, i \in \{1, 2, ..., 16\}$. For any RGB video V, we represent the gait extracted using 3D pose estimation as G. The gait G is a set of 3D poses $P_1, P_2, ..., P_\tau$ where τ is the number of frames in the input video V. We represent the extracted affective features of a gait G as F. Given the gait features F, we represent the predicted emotion by $e \in \{happy, angry, sad, neutral\}$. These four basic emotions represent emotional states that last for an extended period and are more abundant during walking [12]. These four emotions capture the spectrum of the affective space and a combination of them can be used to represent other emotions [16]

3.2 Overview

Our real-time perceived emotion prediction algorithm is based on a data-driven approach. We present an overview of our approach in Figure 3. During the offline training phase, we use multiple gait datasets and extract affective features. These affective features are based on psychological characterization [24], [30] and consist of both posture and movement features. We also extract deep features by training an LSTM network. We combine these deep and affective features and train a Random Forest classifier. At runtime, given an RGB video of an individual walking, we extract his/her gait in the form of a set of 3D poses using a state-of-the-art 3D human pose estimation technique [34]. We extract affective and deep features from this gait and identify the perceived emotion using the trained Random Forest classifier. We now describe each component of our algorithm in detail.

3.3 Affective Feature Computation

For an accurate prediction of an individual's affective state, both posture and movement features are essential [6]. Features in the form of joint angles, distances, and velocities, and space occupied by the body have been used for recognition of emotions and affective states from gaits [6], [30]. Based on these psychological findings, we compute affective features that include both the posture and the movement features.

We represent the extracted affective features of a gait $\mathbf{G}$ as a vector $F \in \mathbb{R}^{29}$. For feature extraction, we use a single stride from each gait corresponding to consecutive foot strikes of the same foot. We used a single cycle in our experiments because in some of the datasets (CMU, ICT, EWalk) only a single walk cycle was available. When multiple walk cycles are available, they can be used to increase accuracy.

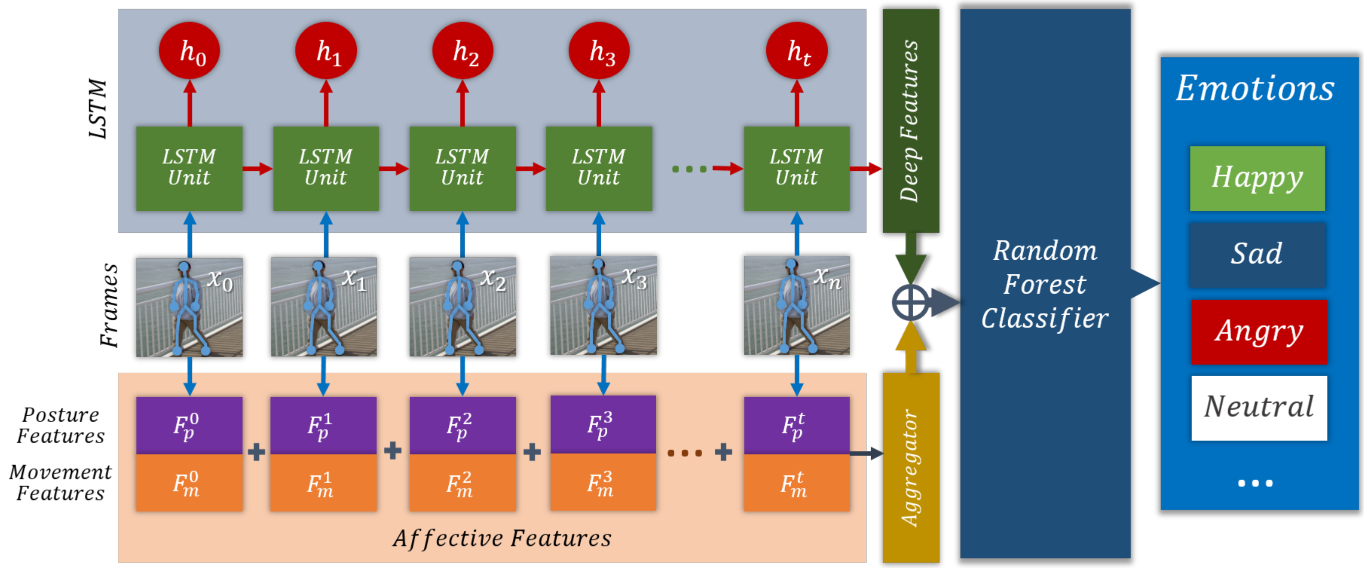


Fig. 3: **Overview:** Given an RGB video of an individual walking, we use a state-of-the-art 3D human pose estimation technique [34] to extract a set of 3D poses. These 3D poses are passed to an LSTM network to extract deep features. We train this LSTM network using multiple gait datasets. We also compute affective features consisting of both posture and movement features using psychological characterization. We concatenate these affective features with deep features and classify the combined features into 4 basic emotions using a Random Forest classifier.

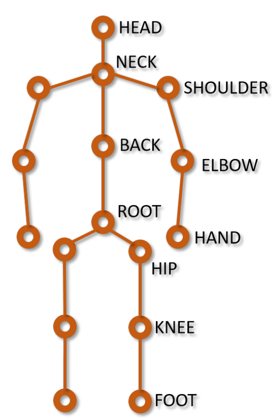


Fig. 4: **Human Representation**: We represent a human by a set of 16 joints. The overall configuration of the human is defined using these joint positions and is used to extract the features.

3.3.1 Posture Features

We compute the features $F_{p,t} \in \mathbb{R}^{12}$ related to the posture P_t of the human at each frame t using the skeletal representation (computed using TimePoseNet Section 4.6). We list the posture features in Table 1. We define posture features of the following types:

- Volume: According to Crenn et al. [30], body expansion conveys positive emotions while a person has a more compact posture during negative expressions. We model this by the volume $F_{volume,t} \in \mathbb{R}$ occupied by the bounding box around the human.
- Area: We also model body expansion by areas of triangles between the hands and the neck and between the feet and the root joint $F_{area,t} \in \mathbb{R}^2$.
- Distance: Distances between the feet and the hands can also be used to model body expansion $F_{distance,t} \in \mathbb{R}^4$.
- Angle: Head tilt is used to distinguish between happy and sad emotions [24], [30]. We model this by the angles extended by different joints at the neck $F_{angle,t} \in \mathbb{R}^5$.

TABLE 1: **Posture Features**: We extract posture features from an input gait using emotion characterization in visual perception and psychology literature [24], [30].

Type	Description
Volume	Bounding box
Angle	At neck by shoulders
	At right shoulder by neck and left shoulder
	At left shoulder by neck and right shoulder
	At neck by vertical and back
	At neck by head and back
Distance	Between right hand and the root joint
	Between left hand and the root joint
	Between right foot and the root joint
	Between left foot and the root joint
	Between consecutive foot strikes (stride length)
Area	Triangle between hands and neck
	Triangle between feet and the root joint

We also include stride length as a posture feature. Longer stride lengths convey anger and happiness and shorter stride lengths convey sadness and neutrality [24]. Suppose

TABLE 2: **Movement Features**: We extract movement features from an input gait using emotion characterization in visual perception and psychology literature [24], [30].

Type	Description
Speed	Right hand
	Left hand
	Head
	Right foot
	Left foot
Acceleration Magnitude	Right hand
	Left hand
	Head
	Right foot
	Left foot
Movement Jerk	Right hand
	Left hand
	Head
	Right foot
	Left foot
Time	One gait cycle

we represent the positions of the left foot joint j_{lFoot} and the right foot joint j_{rFoot} in frame t as $\vec{p}(j_{lFoot}, t)$ and $\vec{p}(j_{rFoot}, t)$ respectively. Then the stride length $s \in \mathbb{R}$ is computed as:

$$s = \max_{t \in 1..\tau} ||\vec{p}(j_{lFoot}, t) - \vec{p}(j_{rFoot}, t)|| \quad (1)$$

We define the posture features $F_p \in \mathbb{R}^{13}$ as the average of $F_{p,t}, t = \{1, 2, .., \tau\}$ combined with the stride length:

$$F_p = \frac{\sum_t F_{p,t}}{\tau} \cup s, \quad (2)$$

3.3.2 Movement Features

Psychologists have shown that motion is an important characteristic for the perception of different emotions [6]. High arousal emotions are more associated with rapid and increased movements than low arousal emotions. We compute the movement features $F_{m,t} \in \mathbb{R}^{15}$ at frame t by considering the magnitude of the velocity, acceleration, and movement jerk of the hand, foot, and head joints using the skeletal representation. For each of these five joints $j_i, i = 1, ..., 5$, we compute the magnitude of the first, second, and third finite derivatives of the position vector $\vec{p}(j_i, t)$ at frame t. We list the movement features in Table 2.

Since faster gaits are perceived as happy or angry whereas slower gaits are considered sad [24], we also include the time taken for one walk cycle ($gt \in \mathbb{R}$) as a movement feature. We define the movement features $F_m \in \mathbb{R}^{16}$ as the average of $F_{m,t}, t = \{1, 2, .., \tau\}$:

$$F_m = \frac{\sum_t F_{m,t}}{\tau} \cup gt, \quad (3)$$

3.3.3 Affective Features

We combine posture and movement features and define affective features F as: $F = F_m \cup F_p$.

4 Perceived Emotion Identification

We use a *vanilla* LSTM network [35] with a cross-entropy loss that models the temporal dependencies in the gait data. We chose an LSTM network to model deep features of walking because it captures the geometric consistency and temporal dependency among video frames for gait modeling [36]. We describe the details of the training of the LSTM in this section.

4.1 Datasets

We used the following publicly available datasets for training our perceived emotion classifier:

- **Human3.6M** [37]: This dataset consists of 3.6 million 3D human images and corresponding poses. It also contains video recordings of 5 female and 6 male professional actors performing actions in 17 scenarios including taking photos, talking on the phone, participating in discussions, etc. The videos were captured at 50 Hz with four calibrated cameras working simultaneously. Of these, there are motion-captured gaits from 14 videos of the subjects walking.
- **CMU** [38]: The CMU Graphics Lab Motion Capture Database contains motion-captured videos of humans interacting among themselves (*e.g.*, talking, playing together), interacting with the environment (*e.g.*, playgrounds, uneven terrains), performing physical activities (*e.g.*, playing sports, dancing), enacting scenarios (*e.g.*, specific behaviors), and locomoting (*e.g.*, running, walking). In total, there are motion captured gaits from 49 videos of subjects walking with different styles.
- **ICT** [39]: This dataset contains motion-captured gaits from walking videos of 24 subjects. The videos were annotated by the subjects themselves, who were asked to label their own motions as well as motions of other subjects familiar to them.
- **BML** [12]: This dataset contains motion-captured gaits from 30 subjects (15 male and 15 female). The subjects were nonprofessional actors, ranging between 17 and 29 years of age with a mean age of 22 years. For the walking videos, the actors walked in a triangle for 30 sec, turning clockwise and then counterclockwise in two individual conditions. Each subject provided 4 different walking styles in two directions, resulting in 240 different gaits.
- **SIG** [40]: This is a dataset of 41 synthetic gaits generated using local mixtures of autoregressive (MAR) models to capture the complex relationships between the different styles of motion. The local MAR models were developed in real-time by obtaining the nearest examples of given pose inputs in the database. The trained model were able to adapt to the input poses with simple linear transformations. Moreover, the local MAR models were able to predict the timings of synthesized poses in the output style.
- **EWalk (Our novel dataset)**: We also collected videos and extracted 1136 gaits using 3D pose estimation. We present details about this dataset in Section 6.

The wide variety of these datasets includes acted as well as non-acting and natural-walking datasets (CMU, ICT) where the subjects were not told to assume an emotion. These datasets provide a good sample of real-world scenarios.

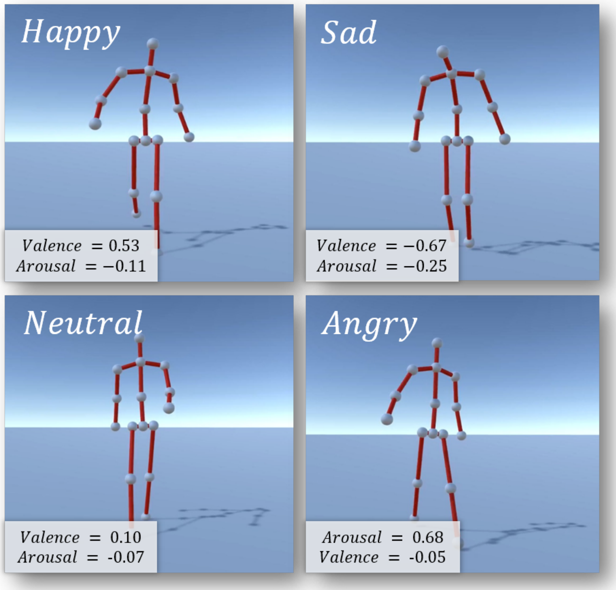


Fig. 5: **Gait Visualizations:** We show the visualization of the motion-captured gaits of four individuals with their classified emotion labels. Gait videos from 248 motion-captured gaits were displayed to the participants in a web based user study to generate labels. We use that data for training and validation.

4.2 Perceived Emotion Labeling

We obtained the perceived emotion labels for each gait using a web-based user study.

4.2.1 *Procedure*

We generated visualizations of each motion-captured gait using a skeleton mesh (Figure 5). For the *EWalk* dataset, we presented the original videos to the participants when they were available. We hid the faces of the actors in these videos to ensure that the emotions were perceived from the movements of the body and gaits, not from the facial expressions.

4.2.2 *Participants*

We recruited 688 participants (279 female, 406 male, $\overline{age} = 34.8$) from Amazon Mechanical Turk and the participant responses were used to generate perceived emotion labels. Each participant watched and rated 10 videos from one of the datasets. The videos were presented randomly and for each video we obtained a minimum of 10 participant responses.

4.2.3 *Analysis*

We asked each participant whether he/she perceived the gait video as happy, angry, sad, or neutral on 5-point Likert items ranging from Strongly Disagree to Strongly Agree. For each gait $\mathbf{G}_i$ in the datasets, we calculated the mean of all participant responses ($r^e_{i,j}$) to each emotion:

$$r^e_i = \frac{\sum_{j=1}^{n_p} r^e_{i,j}}{n_p}, \tag{4}$$

TABLE 3: **Correlation Between Emotion Responses**: We present the correlation between participants' responses to questions relating to the four emotions.

	Happy	Angry	Sad	Neutral
Happy	1.000	-0.268	-0.775	-0.175
Angry	-0.268	1.000	-0.086	-0.058
Sad	-0.775	-0.086	1.000	-0.036
Neutral	-0.175	-0.058	-0.036	1.000

where n_p is the number of participant responses collected and e is one of the four emotions: angry, sad, happy, neutral.

We analyzed the correlation between participants' responses to the questions relating to the four emotions (Table 3). A correlation value closer to 1 indicates that the two variables are positively correlated and a correlation value closer to -1 indicates that the two variables are negatively correlated. A correlation value closer to 0 indicates that two variables are uncorrelated. As expected, *happy* and *sad* are negatively correlated and *neutral* is uncorrelated with the other emotions.

Previous research in the psychology literature suggests that social perception is affected by the gender of the observer [41], [42], [43]. To verify that our results do not significantly depend on the gender of the participants, we performed a t-test for differences between the responses by male and female participants. We observed that the gender of the participant did not affect the responses significantly ($t = -0.952, p = 0.353$).

We obtained the emotion label e_i for $\mathbf{G}_i$ as follows:

$$e_i = e \mid r^e_i > \theta, \tag{5}$$

where $\theta = 3.5$ is an experimentally determined threshold for emotion perception.

If there are multiple emotions with average participant responses greater than $r^e_i > \theta$, the gait is not used for training.

4.3 Long Short-Term Memory (LSTM) Networks

LSTM networks [35] are neural networks with special units known as "memory cells" that can store data values from particular time steps in a data sequence for arbitrarily long time steps. Thus, LSTMs are useful for capturing temporal patterns in data sequences and subsequently using those patterns in prediction and classification tasks. To perform supervised classification, LSTMs, like other neural networks, are trained with a set of training data and corresponding class labels. However, unlike traditional feedforward neural networks that learn structural patterns in the training data, LSTMs learn feature vectors that encode temporal patterns in the training data.

LSTMs achieve this by training one or more "hidden" cells, where the output at every time step at every cell depends on the current input and the outputs at previous time steps. These inputs and outputs to the LSTM cells are controlled by a set of gates. LSTMs commonly have three kinds of gates: the input gate, the output gate, and the forget gate, represented by the following equations:

$$\textit{Input Gate } (i): \quad i_t = \sigma(W_i + U_i h_{t-1} + b_i) \tag{6}$$
$$\textit{Output Gate } (o): \quad o_t = \sigma(W_o + U_o h_{t-1} + b_o) \tag{7}$$
$$\textit{Forget Gate } (f): \quad f_t = \sigma(W_f + U_f h_{t-1} + b_f) \tag{8}$$

where $\sigma(\cdot)$ denotes the activation function and W_g, U_g and b_g denote the weight matrix for the input at the current time step, the weight matrix for the hidden cell at the previous time step, and the bias, on gate $g \in \{i, o, f\}$, respectively. Based on these gates, the hidden cells in the LSTMs are then updated using the following equations:

$$c_t = f_t \circ c_{t-1} + i_t \circ \sigma(W_c x_t + U_c h_t + b_c) \tag{9}$$
$$h_t = \sigma(o_t \circ c_t) \tag{10}$$

where $\circ$ denotes the Hadamard or elementwise product, c is referred to as the cell state, and W_c, U_c and b_c are the weight matrix for the input at the current time step, the weight matric for the hidden cell at the previous time step, and the bias, on c, respectively.

4.4 Deep Feature Computation

We used the LSTM network shown in Figure 3. We obtained deep features from the final layer of the trained LSTM network. We used the 1384 gaits from the various public datasets (Section 4.1). We also analyzed the extracted deep features using an LSTM encoder-decoder architecture with reconstruction loss. We generated synthetic gaits and observed that our LSTM-based deep features correctly model the 3D positions of joints relative to each other at each frame. The deep features also capture the periodic motion of the hands and legs.

4.4.1 Implementation Details

The training procedure of the LSTM network that we followed is laid out in Algorithm 1. For training, we used a mini-batch size of 8 (*i.e.*, $b = 8$ in Algorithm 1) and 500 training epochs. We used the Adam optimizer [44] with an initial learning rate of 0.1, decreasing it to $\frac{1}{10}$-th of its current value after 250, 375, and 438 epochs. We also used a momentum of 0.9 and a weight-decay of 5×10^{-4}. The training was carried out on an NVIDIA GeForce GTX 1080 Ti GPU.

Algorithm 1 LSTM Network for Emotion Perception

Input: N training gaits $\{\mathbf{G}_i\}_{i=1\dots N}$ and corresponding emotion labels $\{\mathrm{L}_i\}_{i=1\dots N}$.

Output: Network parameters θ such that the loss $\sum_{i=1}^{N} \|\mathrm{L}_i - f_\theta(\mathbf{G}_i)\|^2$ is minimized, where $f_\theta(\cdot)$ denotes the network.

```
procedure TRAIN
    for number of training epochs do
        for number of iterations per epoch do
            Sample mini-batch of b training gaits and corresponding labels
            Update the network parameters θ w.r.t. the b samples using backpropagation.
```

TABLE 4: **Performance of Different Classification Methods**: We analyze different classification algorithms to classify the concatenated deep and affective features. We observe an accuracy of 80.07% with the Random Forest classifier.

Algorithm (Deep + Affective Features)	Accuracy
LSTM + Support Vector Machines (SVM)	70.04%
LSTM + Stochastic Gradient Descent (SGD)	71.01%
LSTM + Random Forest	**80.07**%

4.5 Classification

We concatenate the deep features with affective features and use a Random Forest classifier to classify these concatenated features. Before combining the affective features with the deep features, we normalize them to a range of $[-1, 1]$. We use Random Forest Classifier with 10 estimators and a maximum depth of 5. We use this trained classifier to classify perceived emotions.

4.6 Realtime Perceived Emotion Recognition

At runtime, we take an RGB video as input and use the trained classifier to identify the perceived emotions. We exploit a real-time 3D human pose estimation algorithm, *TimePoseNet* [34], which uses a semi-supervised learning method that utilizes the more widely available 2D human pose data [45] to learn the 3D information.

TimePoseNet is a single person model and expects a sequence of images cropped closely around the person as input. Therefore, we first run a real-time person detector [46] on each frame of the RGB video and extract a sequence of images cropped closely around the person in the video V. The frames of the input video V are sequentially passed to *TimePoseNet*, which computes a 3D pose output for each input frame. The resultant poses $P_1, P_2, ..., P_\tau$ represent the extracted output gait G. We normalize the output poses so that the root position always coincides with the origin of the 3D space. We extract features of the gait G using the trained LSTM model. We also compute the affective features and classify the combined features using the trained Random Forest classifier.

5 Results

We provide the classification results of our algorithm in this section.

5.1 Analysis of Different Classification Methods

We analyze different classification techniques to classify the combined deep and affective features and compare the resulting accuracies in Table 4. We use the Random Forest classifier in the subsequent results because it provides the highest accuracy (80.07%) of all the classification methods. Additionally, our algorithm achieves 79.72% accuracy on the non-acted datasets (CMU and ICT), indicating that it performs equally well on acted and non-acted data.

5.2 Comparison with Other Methods

In this section, we present the results of our algorithm and compare it with other state-of-the-art methods. We compare the results with the following methods:

TABLE 5: **Accuracy**: Our method with combined deep and affective features classified with a Random Forest classifier achieves an accuracy of 80.07%. We observe an improvement of 13.85% over state-of-the-art emotion identification methods and an improvement of 24.60% over a baseline LSTM-based classifier.

Method	Accuracy
Baseline (Vanilla LSTM)	55.47%
Affective Features Only	68.11%
Karg et al. [24]	39.58%
Venture et al. [47]	30.83%
Crenn et al. [30]	66.22%
Crenn et al. [32]	40.63%
Daoudi et al. [48]	42.52%
Wang et al. [33]	53.73%
Our Method (Deep + Affective Features)	**80.07%**

- Karg et al. [24]: This method is based on using gait features related to shoulder, neck, and thorax angles, stride length, and velocity. These features are classified using PCA-based methods. This method only models the posture features for the joints and doesn't model the movement features.
- Venture et al. [47]: This method uses the auto-correlation matrix of the joint angles at each frame and uses similarity indices for classification. The method provides good intra-subject accuracy but performs poorly for the inter-subject databases.
- Crenn et al. [30]: This method uses affective features from both posture and movement and classifies these features using SVMs. This method is trained for more general activities like knocking and does not use information about feet joints.
- Daoudi et al. [48]: This method uses a manifold of symmetric positive definite matrices to represent body movement and classifies them using the Nearest Neighbors method.
- Crenn et al. [32]: This method synthesizes a neutral motion from an input motion and uses the difference between the input and the neutral emotion as the feature for classifying emotions. This method does not use the psychological features associated with walking styles.
- Wang et al. [33]: This method uses a Kinect to capture the gaits and identifies whether an individual is angry, happy, or neutral using four walk cycles using a feature-based approach.

We also compare our results to a baseline where we use the LSTM to classify the gait features into the four emotion classes. Table 5 provides the accuracy results of our algorithm and shows comparisons with other methods. These methods require input in the form of 3D human poses and then they identify the emotions perceived from those gaits. For this experiment, we extracted gaits from the RGB videos of the *EWalk* dataset and then provided them as input to the state-of-the-art methods along with the motion-captured gait datasets. Accuracy results are obtained using 10-fold cross-validation on various datasets (Section 4.1).

We also show the percentage of gaits that the LSTM+Random Forsest classifier correctly classified for each emotion class in Figure 6. As we can see, for every class,

True Class \ Predicted Class	Angry	Happy	Neutral	Sad
Angry	**80.47%**	7.81%	2.34%	9.38%
Happy	4.12%	**87.05%**	6.36%	2.47%
Neutral	1.29%	2.59%	**80.53%**	15.59%
Sad	6.61%	4.13%	10.74%	**78.52%**

Fig. 6: **Confusion Matrix**: For each emotion class, we show the percentage of gaits belonging to that class that were correctly classified by the LSTM+Random Forest classifier (green background) and the percentage of gaits that were misclassified into other classes (red background).

around 80% of the gaits are correctly classified, implying that the classifier learns to recognize each class equally well. Further, when the classifier does make mistakes, it tends to confuse neutral and sad gaits more than between any other class pairs.

5.3 Analysis of the Learned Deep Features

We visualize the scatter of the deep feature vectors learned by the LSTM network by projecting them in the top 3 principal component directions. This is shown in Figure 11. We observe that the data points are well-separated even in the projected dimension. By extension, this implies that the deep features are at least as well separated in their original dimension. Therefore, we can conclude that the LSTM network has learned meaningful representations of the input data that help it distinguish accurately between the different emotion classes.

Additionally, we show the saliency maps given by the network, for one correctly classified sample from each of the four emotion classes in Figure 14. The saliency maps show the activation on each of the joints during a single walk cycle. Red denotes high activation while black denotes no activation. Intuitively, the activated nodes at every frame are the nodes the network focuses on in that frame. Finally, based on the activation values of activated nodes in all the frames, the network determines the class label for the gait. We can observe from Figure 14 that the network focuses mostly on the hand joints (observing arm swinging), the feet joints (observing stride), and the head and neck joints (observing head jerk). Based on the speed and frequency of the movements of these joints, the network decides the class labels. For example, the activation values on the joints for anger (Figure 14a) are much higher than the ones for sadness (Figure 14c), which matches with the psychological studies of how angry and sad gaits typically look. This shows that the features learned by the network are representative of the psychological features humans tend to use when perceiving emotions from gaits.

6 Emotional Walk *(EWalk)* Dataset

In this section, we describe our new dataset of videos of individuals walking. We also provide details about the perceived emotion annotations of the gaits obtained from this dataset.

6.1 Data

The EWalk dataset contains 1384 gaits with emotion labels from four basic emotions: happy, angry, sad, and neutral (Figure 9). These gaits are either motion-captured or extracted from RGB videos. We also include synthetically generated gaits using state-of-the-art algorithms [40]. In addition to the emotion label for each gait, we also provide values of affective dimensions: valence and arousal.

6.2 Video Collection

We recruited 24 subjects from a university campus. The subjects were from a variety of ethnic backgrounds and included 16 male and 8 female subjects. We recorded the videos in both indoor and outdoor environments. We requested that they walk multiple times with different walking styles. Previous studies show that non-actors and actors are both equally good at walking with different emotions [21]. Therefore, to obtain different walking styles, we suggested that the subjects could assume that they are experiencing a certain emotion and walk accordingly. The subjects started 7m from a stationary camera and walked towards it. The videos were later cropped to include a single walk cycle.

6.3 Data Generation

Once we collect walking videos and annotate them with emotion labels, we can also use them to train generator networks to generate annotated synthetic videos. Generator networks have been applied for generating videos and joint-graph sequences of human actions such as walking, sitting, running, jumping, etc. Such networks are commonly based on either Generative Adversarial Networks (GANs) [49] or Variational Autoencoders (VAEs) [50].

GANs (Figure 7) are comprised of a generator that generates data from random noise samples and a discriminator that discriminates between real data and the data generated by the generator. The generator is considered to be trained when the discriminator fails to discriminate between the real and the generated data.

VAEs (Figure 8), on the other hand, are comprised of an encoder followed by a decoder. The encoder learns a latent embedding space that best represents the distribution of the real data. The decoder then draws random samples from the latent embedding space to generate synthetic data.

For temporal data such human action videos or joint-graph sequences, two different approaches are commonly taken. One approach is to individually generate each point in the temporal sequence (frames in a video or graphs in a graph sequence) respectively and then fuse them together in a separate network to generate the complete sequence. The methods in [51], [52], for example, use this approach. The network generating the individual points only considers the spatial constraints of the data, whereas the network fusing the points into the sequence only considers the temporal constraints of the data. The alternate approach is to train a single network by providing it both the spatial and temporal constraints of the data. For example, the approach used by Sijie et al. [53]. The first approach is relatively more lightweight, but it does not explicitly consider spatial temporal inter-dependencies in the data, such as the differences in the arm swinging speeds between angry and sad gaits. While the latter approach does take these inter-dependencies into account, it is also harder to train because of these additional constraints.

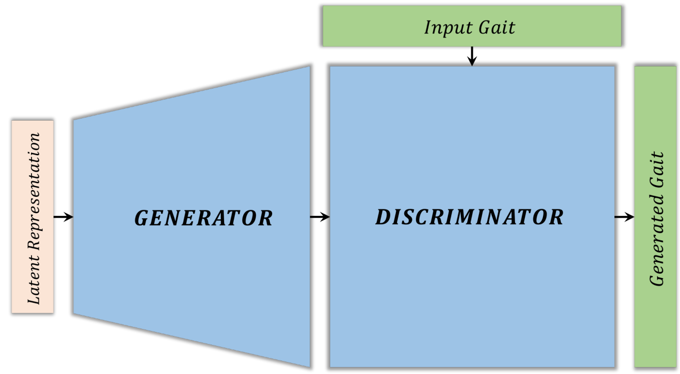


Fig. 7: **Generative Adversarial Networks (GANs)**: The network consists of a generator that generates synthetic data from random samples drawn from a latent distribution space. This is followed by a discriminator that attempts to discriminate between the generated data and the real input data. The objective of the generator is to learn the latent distribution space of the real data whereas the objective of the discriminator is to learn to discriminate between the real data and the synthetic data generated by the generator. The network is said to be learned when the discriminator fails to distinguish between the real and the synthetic data.

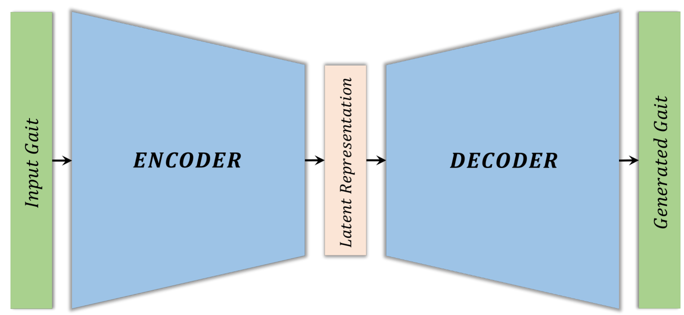


Fig. 8: **Variational Autoencoders (VAEs)**: The encoder consists of an encoder that transforms the input data to a latent distribution space. This is followed by a discriminator that draws random samples from the latent distribution space to generate synthetic data. The objective of the overall network is then to learn the latent distribution space of the real data, so that the synthetic data generated by the decoder belongs to the same distribution space as the real data.

6.4 Analysis

We presented the recorded videos to MTurk participants and obtained perceived emotion labels for each video using the method described in Section 4.2. Our data is widely

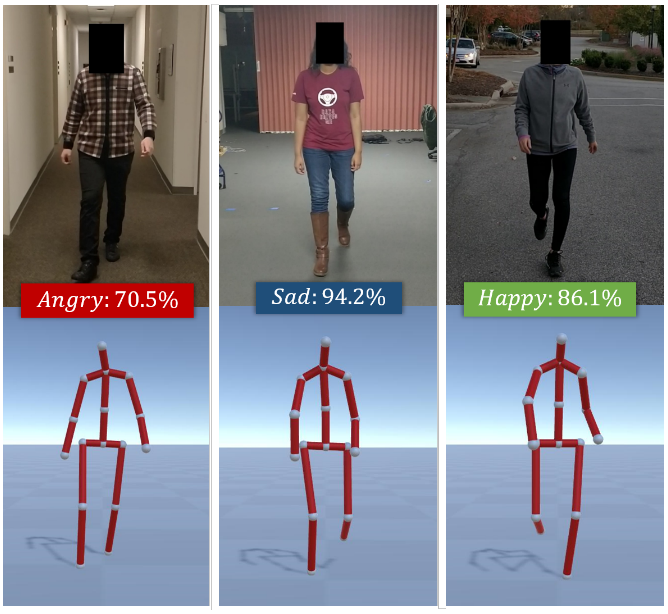


Fig. 9: **EWalk Dataset:** We present the EWalk dataset containing RGB videos of pedestrians walking and the perceived emotion label for each pedestrian.

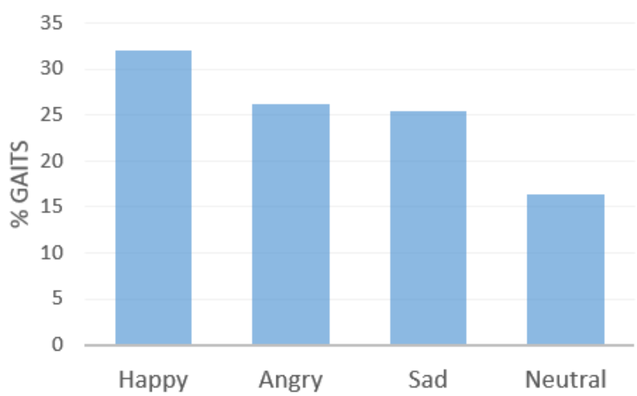


Fig. 10: **Distribution of Emotion in the Datasets:** We present the percentage of gaits that are perceived as belonging to each of the emotion categories (happy, angry, sad, or neutral). We observe that our data is widely distributed.

distributed across the four categories with the *Happy* category containing the most largest of gaits (32.07%) and the *Neutral* category containing the smallest number of gaits with 16.35% (Figure 10).

6.4.1 *Affective Dimensions*

We performed an analysis of the affective dimensions (i.e. valence and arousal). For this purpose, we used the participant responses to the questions about the happy, angry, and sad emotions. We did not use the responses to the question about the neutral emotion because it corresponds to the origin of the affective space and does not contribute to the valence and arousal dimensions. We performed a Principal Component Analysis (PCA) on the participant responses $[r_i^{happy}, r^{angry}, r^{sad}]$ and observed that the following two

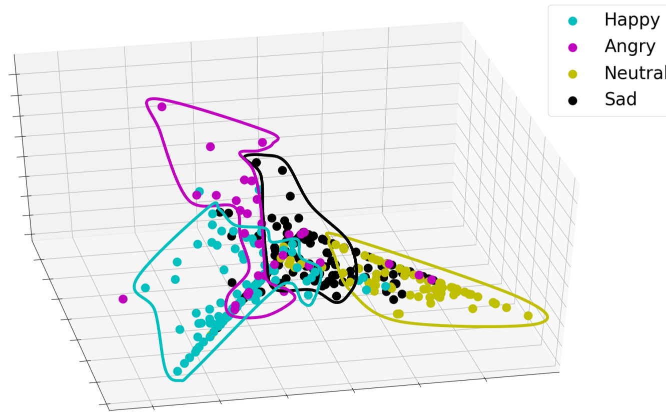


Fig. 11: **Scatter Plot of the Learned Deep Features**: These are the deep features learned by the LSTM network from the input data points, projected in the 3 principal component directions. The different colors correspond to the different input class labels. We can see that the features for the different classes are well-separated in the 3 dimensions. This implies that the LSTM network learns meaningful representations of the input data for accurate classification.

principal components describe 94.66% variance in the data:

$$\begin{bmatrix} PC1 \\ PC2 \end{bmatrix} = \begin{bmatrix} 0.67 & -0.04 & -0.74 \\ -0.35 & 0.86 & -0.37 \end{bmatrix} \tag{11}$$

We observe that the first component with high values of the *Happy* and *Sad* coefficients represents the *valence* dimension of the affective space. The second principal component with high values of the *Anger* coefficient represents the *arousal* dimension of the affective space. Surprisingly, this principal component also has a negative coefficient for the *Happy* emotion. This is because a calm walk was often rated as happy by the participants, resulting in low arousal.

6.4.2 *Prediction of Affect*

We use the principal components from Equation 11 to predict the values of the *arousal* and *valence* dimensions. Suppose, the probabilities predicted by the Random Forest classifier are $p(h), p(a)$, and $p(s)$ corresponding to the emotion classes $happy$, $angry$, and sad, respectively. Then we can obtain the values of $valence$ and $arousal$ as:

$$valence = \begin{bmatrix} 0.67 & -0.04 & -0.74 \end{bmatrix} \begin{bmatrix} p(h) & p(a) & p(s) \end{bmatrix}^T \tag{12}$$

$$arousal = \begin{bmatrix} -0.35 & 0.86 & -0.37 \end{bmatrix} \begin{bmatrix} p(h) & p(a) & p(s) \end{bmatrix}^T \tag{13}$$

7 Application: Virtual Character Generation

In this section, we present an application of our method that generates virtual characters with given desired emotions (Figure 12).

7.1 Overview

We provide an overview of our end-to-end approach to simulating virtual characters in Figure 13. We assume that the environment consists of static and dynamic obstacles. At the start of the simulation, we initialize the environment

Fig. 12: **Application**: Our gaits and their perceived emotion labels can be used to generate virtual characters with different emotions. We show a character that is generated using our approach to convey basic emotions: angry, happy, sad, and neutral.

state with positions and dimensions of the static obstacles and the current positions and velocities of the dynamic obstacles. We also initialize a Behavioral Finite State Machine (BFSM) based on the user input and the intended tasks. We set up a 3D model for each virtual character that is rigged using automatic rigging software and associate a hierarchical skeleton with appropriate joint values.

7.2 Behavioral Finite State Machine

We represent the behavioral state of the virtual characters in a BFSM and use it to control their behaviors. At runtime, we consider the environment state and the context of the current task and update the state of the BFSM that determines the virtual characters' behavior. This state also computes a goal position for each virtual character.

7.3 Global and Local Navigation

If the goal positions of virtual characters are different from their current positions, then a navigation algorithm is used to compute the trajectories to the new positions. To provide collision-free navigation in the presence of obstacles or other virtual characters, we utilize the multi-agent simulation framework, *Menge* [54]. In this framework, a global navigation step first breaks down the goal positions into intermediate goals that avoid collisions with the static obstacles in the environment. Next, a local navigation step uses a reciprocal collision avoidance (RVO) approach to avoid collisions with dynamic obstacles and provide navigation to the intermediate goals [55].

In this approach, we represent each agent on the 2D ground plane and generate smooth, stable, collision-free velocities. RVO is an agent-based approach that computes a collision-free 2D velocity for an agent given its preferred velocity, time horizon (t_{max}), and current positions and velocities of the all virtual agents in the environment. In other words, it computes a velocity that can generate a collision-free trajectory at time t_{max}. We update the character's location in the virtual world according to this collision-free trajectory at each frame.

7.4 Gait Generation

In addition to the goal position for each virtual character, the BFSM state also determines the desired emotion that each virtual character must convey. To achieve this, we use our gait-based approach to identify the perceived emotion. For each virtual character, we obtain a set of gaits that correspond to the desired emotion using our gait dataset and associated labels. We choose one of the gaits from this set and use it to update the joint positions of the agent in the virtual world. The selection of a gait can be made according to many criteria (such as personality or preferred walking speed).

8 Conclusion, Limitations, and Future Work

We presented a novel method for classifying perceived emotions of individuals based on their walking videos. Our method is based on learning deep features computed using LSTM and exploits psychological characterization to compute affective features. The mathematical characterization

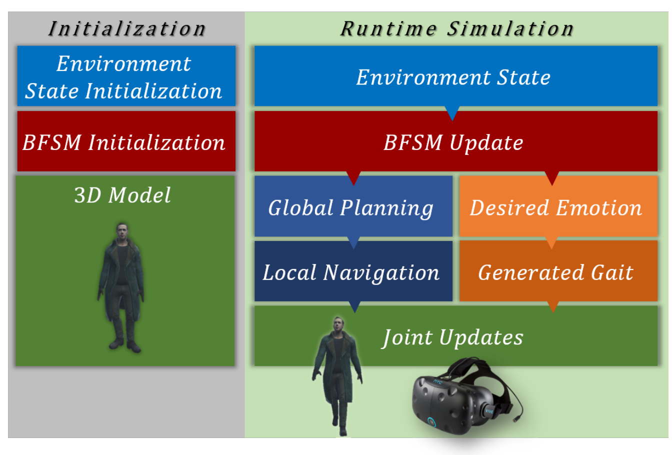
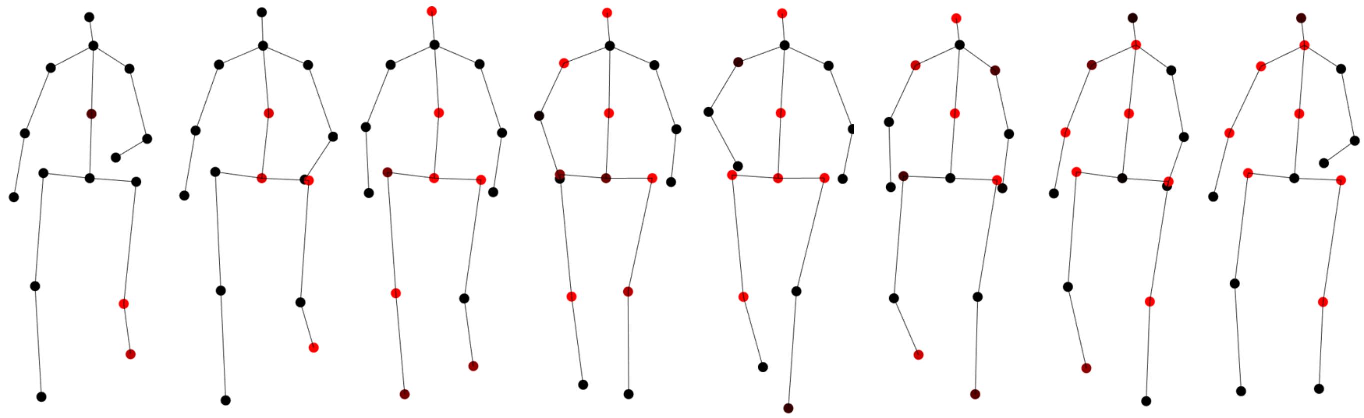
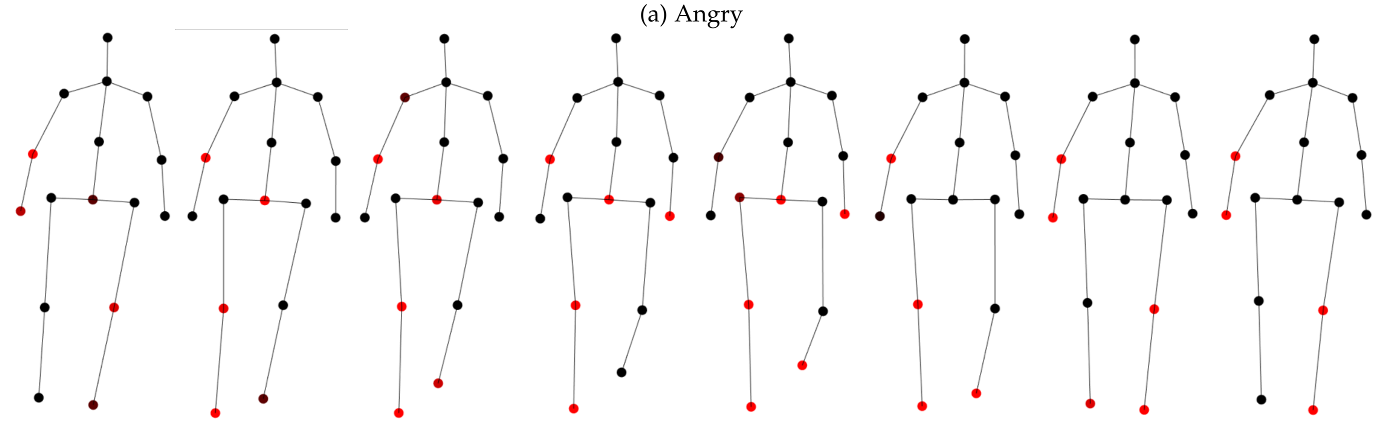


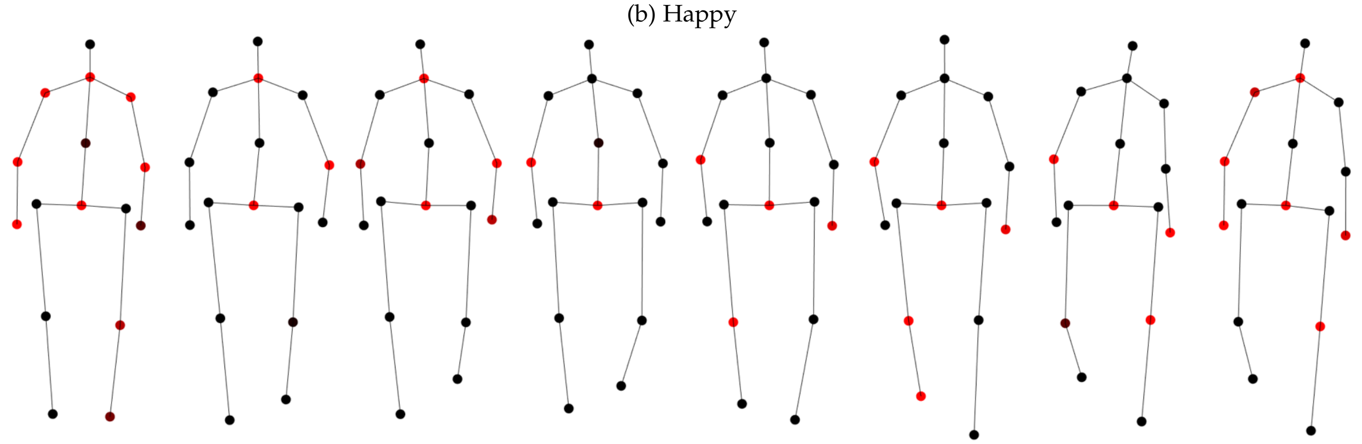
Fig. 13: **Virtual Character Generation**: We provide an overview of our end-to-end approach for simulating virtual characters. We represent the behavioral state of the virtual characters in a Behavioral Finite State Machine (BFSM) and use it to control their behavior based on the state of the environment, which consists of static and dynamic obstacles. We use our perceived emotion prediction to generate gaits for the virtual characters based on their desired emotions.



(a) Angry



(b) Happy



(c) Sad

Fig. 14: **Saliency Maps**: We present the saliency maps for examples from each of the four emotion classes, as learned by the network for a single walk cycle. The maps show activations on the joints during the walk cycle. Black represents no activation and red represents high activation. For all the emotion classes, the hand, feet and head joints have high activations, implying that the network deems these joints to be more important for determining the class. Moreover, the activation values on these joints for a high arousal emotion (*e.g.*, angry) are higher than those for a low arousal emotion (*e.g.*, sad), implying the network learns that higher arousal emotions lead to more vigorous joint movements.

of computing gait features also has methodological implications for psychology research. This approach explores the basic psychological processes used by humans to perceive emotions of other individuals using multiple dynamic and naturalistic channels of stimuli. We concatenate the deep and affective features and classify the combined features using a Random Forest Classification algorithm. Our algorithm achieves an absolute accuracy of 80.07%, which is an improvement of 24.60% over vanilla LSTM (i.e., using only deep features) and offers an improvement of 13.85% over state-of-the-art emotion identification algorithms. Our approach is also the first approach to provide a real-time pipeline for emotion identification from walking videos by leveraging state-of-the-art 3D human pose estimation. We also present a dataset of videos (EWalk) of individuals walking with their perceived emotion labels. The dataset is collected with subjects from a variety of ethnic backgrounds in both indoor and outdoor environments.

There are some limitations to our approach. The accuracy of our algorithm depends on the accuracy of the 3D human pose estimation and gait extraction algorithms. Therefore, emotion prediction may not be accurate if the estimated 3D human poses or gaits are noisy. Our affective computation requires joint positions from the whole body, but the whole body pose data may not be available if there are occlusions in the video. We assume that the walking motion is natural and does not involve any accessories (e.g., suitcase, mobile phone, etc.). As part of future work, we would like to collect more datasets and address these issues. We will also attempt to extend our methodology to consider more activities such as running, gesturing, etc. Finally, we would like to combine our method with other emotion identification algorithms that use human speech and facial expressions.

REFERENCES

[1] L. Devillers, M. Tahon *et al.*, "Inference of human beings emotional states from speech in human–robot interactions," *IJSR*, 2015.

[2] F. C. Benitez-Quiroz, R. Srinivasan *et al.*, "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *CVPR*, 2016.

[3] J. Russell, J. Bachorowski *et al.*, "Facial & vocal expressions of emotion," *Rev. of Psychology*, 2003.

[4] P. Ekman, "Facial expression and emotion," *American Psychologist*, vol. 48, no. 4, pp. 384–392, 1993.

[5] J. M. Fernandez-Dols, M. A. Ruiz-Belda *et al.*, "Expression of emotion versus expressions of emotions: Everyday conceptions of spontaneous facial behavior." *Everyday Conceptions of Emotion*, 1995.

[6] A. Kleinsmith, N. Bianchi-Berthouze *et al.*, "Affective body expression perception and recognition: A survey," *IEEE TAC*, 2013.

[7] H. Meeren, C. van Heijnsbergen *et al.*, "Rapid perceptual integration of facial expression and emotional body language," *PNAS*, 2005.

[8] H. Aviezer, Y. Trope *et al.*, "Body cues, not facial expressions, discriminate between intense positive and negative emotions," 2012.

[9] J. Montepare, S. Goldstein *et al.*, "The identification of emotions from gait information," *Journal of Nonverbal Behavior*, 1987.

[10] H. Wallbott, "Bodily expression of emotion," *European Journal of Social Psychology*, 1998.

[11] J. Michalak, N. Troje *et al.*, "Embodiment of sadness and depression-gait patterns associated with dysphoric mood," *Psychosomatic Medicine*, 2009.

[12] Y. Ma, H. M. Paterson *et al.*, "A motion capture library for the study of identity, gender, and emotion perception from biological motion," *Behavior research methods*, 2006.

[13] A. Loutfi, J. Widmark *et al.*, "Social agent: Expressions driven by an electronic nose," in *VECIMS*. IEEE, 2003.

[14] P. Ekman and W. V. Friesen, "Head and body cues in the judgment of emotion: A reformulation," *Perceptual and motor skills*, 1967.

[15] A. Mehrabian, "Basic dimensions for a general psychological theory implications for personality, social, environmental, and developmental studies," 1980.

[16] J. Mikels, B. Fredrickson *et al.*, "Emotional category data on images from the international affective picture system," *Behavior research methods*, 2005.

[17] J. Morris, "Observations: Sam: the self-assessment manikin; an efficient cross-cultural measurement of emotional response," *JAR*, 1995.

[18] H. Yates, B. Chamberlain *et al.*, "Arousal detection for biometric data in built environments using machine learning," in *IJCAI AIAC*, 2017.

[19] M. Atcheson, V. Sethu *et al.*, "Gaussian process regression for continuous emotion recognition with global temporal invariance," in *IJCAI AIAC*, 2017.

[20] R. Picard, "Toward agents that recognize emotion," *Proc. of IMAGINA*, pp. 153–165, 1998.

[21] C. Roether, L. Omlor *et al.*, "Critical features for the perception of emotion from gait," *Vision*, 2009.

[22] M. Gross, E. Crane *et al.*, "Methodology for assessing bodily expression of emotion," *Journal of Nonverbal Behavior*, 2010.

[23] M. De Meijer, "The contribution of general features of body movement to the attribution of emotions," *Journal of Nonverbal behavior*, 1989.

[24] M. Karg, K. Kuhnlenz *et al.*, "Recognition of affect based on gait patterns," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 2010.

[25] L. Omlor, M. Giese *et al.*, "Extraction of spatio-temporal primitives of emotional body expressions," *Neurocomputing*, 2007.

[26] M. D. Robinson and G. L. Clore, "Belief and feeling: Evidence for an accessibility model of emotional self-report," *Psychological Bulletin*, vol. 128, no. 6, pp. 934–960, 2002.

[27] L. F. Barrett, "Feelings or words? understanding the content in self-report ratings of experienced emotion," *Journal of Personality and Social Psychology*, vol. 87, no. 2, p. 266, 2004.

[28] R. E. Nesbett and T. D. Wilson, "Telling more than we can know: Verbal reports on mental processes," *Psychological Review*, vol. 84, no. 3, p. 231, 1977.

[29] K. S. Quigley, K. A. Lindquist, and L. F. Barrett, "Inducing and measuring emotion and affect: Tips, tricks, and secrets," *Handbook of research methods in social and personality psychology*, vol. ed Judd CJ, p. 220252, 2014.

[30] A. Crenn, A. Khan *et al.*, "Body expression recognition from anim. 3d skeleton," in *IC3D*, 2016.

[31] Wang, Enescu *et al.*, "Adaptive realtime emotion recognition from body movements," *TiiS*, 2016.

[32] A. Crenn, A. Meyer *et al.*, "Toward an efficient body expression recognition based on the synthesis of a neutral movement," in *ICMI*, 2017.

[33] J. Wang, B. Li *et al.*, "Automatic emotion recognition based on non-contact gaits information," in *Adv Methods and Techs in AI, Simulation, and HCI*, 2019.

[34] R. Dabral, A. Mundhada *et al.*, "Learning 3d human pose from structure and motion," in *ECCV*, 2018.

[35] K. Greff, R. K. Srivastava *et al.*, "Lstm: A search space odyssey," *IEEE transactions on neural networks and learning systems*, 2017.

[36] Y. Luo, J. Ren *et al.*, "Lstm pose machines," in *CVPR*, 2018.

[37] C. Ionescu, D. Papava *et al.*, "Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *TPAMI*, 2014.

[38] "Cmu graphics lab motion capture database," http://mocap.cs.cmu.edu/, 2018.

[39] S. Narang, A. Best *et al.*, "Motion recognition of self and others on realistic 3d avatars," *Computer Animation & Virtual Worlds*, 2017.

[40] S. Xia, C. Wang *et al.*, "Realtime style transfer unlabeled heterogeneous human motion," *TOG*, 2015.

[41] L. L. Carli, S. J. LaFleur, and C. C. Loeber, "Nonverbal behavior, gender, and influence." *Journal of personality and social psychology*, vol. 68, no. 6, p. 1030, 1995.

[42] J. Forlizzi, J. Zimmerman, V. Mancuso, and S. Kwak, "How interface agents affect interaction between humans and computers," in *Proceedings of the 2007 conference on Designing pleasurable products and interfaces*. ACM, 2007, pp. 209–221.

[43] N. C. Krämer, B. Karacora, G. Lucas, M. Dehghani, G. Rüther, and J. Gratch, "Closing the gender gap in stem with friendly male instructors? on the effects of rapport behavior and gender of a virtual agent in an instructional interaction," *Computers & Education*, vol. 99, pp. 1–13, 2016.

[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[45] T. Lin, M. Maire *et al.*, "Microsoft coco: Common objects in context," in *ECCV*, 2014.

[46] Z. Cao, T. Simon *et al.*, "Realtime 2d pose estimation using part affinity fields," in *CVPR*, 2017.

[47] G. Venture, H. Kadone *et al.*, "Recognizing emotions conveyed by human gait," *International Journal of Social Robotics*, 2014.

[48] M. Daoudi, "Emotion recognition by body movement representation on the manifold of symmetric positive definite matrices," in *ICIAP*, 2017.

[49] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[50] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.
[51] C. Yang, Z. Wang, X. Zhu, C. Huang, J. Shi, and D. Lin, "Pose guided human video generation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 201–216.
[52] H. Cai, C. Bai, Y.-W. Tai, and C.-K. Tang, "Deep video generation, prediction and completion of human action sequences," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 366–382.
[53] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
[54] S. Curtis, A. Best, and D. Manocha, "Menge: A modular framework for simulating crowd movement," *Collective Dynamics*, vol. 1, pp. 1–40, 2016.
[55] J. Van Den Berg, S. J. Guy, M. Lin, and D. Manocha, "Reciprocal n-body collision avoidance," in *Robotics Research*, 2011.

Tanmay Randhavane Tanmay Randhavane is a graduate student at the Department of Computer Science, University of North Carolina, Chapel Hill, NC, 27514.

Uttaran Bhattacharya Kyra Kapsaskis is a full time research assistant in the Department of Psychology and Neuroscience, University of North Carolina, Chapel Hill, NC, 27514.

Kyra Kapsaskis Kyra Kapsaskis is a full time research assistant in the Department of Psychology and Neuroscience, University of North Carolina, Chapel Hill, NC, 27514.

Kurt Gray Kurt Gray is a Associate Professor with the Department of Psychology and Neuroscience, University of North Carolina, Chapel Hill, NC, 27514.

Aniket Bera Aniket Bera is a Research Assistant Professor at the Department of Computer Science, University of North Carolina, Chapel Hill, NC, 27514.

Dinesh Manocha Dinesh Manocha is a Paul Chrisman Iribe Chair of Computer Science and Electrical & Computer Engineering at the University of Maryland at College Park, MD, 20740.