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 is based on using deep features learned using LSTM on labeled emotion datasets. Furthermore, we combine these features with affective features computed from the 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 74.10% accuracy in identifying the perceived emotions. 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 behavior and decide our actions towards them. For example, people interact very differently with someone they perceive to be angry and hostile than they do with someone they perceive to be calm and contented. Furthermore, the emotions of unknown individuals can also govern our behavior, (e.g., emotions of pedestrians on a road-crossing or emotions of passengers in a
 Humans perceive the emotions of other individuals using verbal and non-verbal cues. Robots and intelligent 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 interaction with humans [49, 15].

Understanding the perceived emotions of individuals using non-verbal cues is a challenging problem. Non-verbal cues humans use to perceive emotions include both facial expressions and body movements. Due to a more extensive availability of data, considerable research has focused on using facial expressions to understand emotion [48, 18, 42]. However, recent studies in the psychology literature question the communicative purpose of facial expressions and doubt the quick, automatic process of perceiving emotions from these expressions [47]. There are instances when facial expressions can be unreliable, such as with “Mock” or “Referential Expressions” [16]. Whether or not an audience is present also impacts the association between positive emotional states and smiling [19, 30].

Research has shown that body expressions are also crucial in emotion expression and perception [13, 27, 35, 51, 43]. For example, 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 [35]. Aviezer et al.’s study [3] on intense 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 were.

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. [46], 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, greater foot landing force, and erect posture. Specific movements have also been categorized with specific emotions [22]. For example, sad movements are characterized by a collapsed upper body and low movement activity [53]. Happy movements have a faster pace with more arm swaying [37].

**Main Results:** We present an end-to-end automatic emotion identification approach for videos of walking individuals. 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 of time and are more abundant in the walking activity [34]. 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 humans in walking. We combine these affective features with LSTM-based deep features and use a Random Forest Classifier to classify them into 4 emotion categories. We obtain an accuracy of 74.10% with our combined affective and deep features. We observe an improvement of 7.88% over other gait-based perceived emotion classification algorithms.

We also present a new dataset, “Emotion Walk (EWalk)”, which contains more than 80 videos of individuals walking in both indoor and outdoor locations. Our dataset consists of extracted gaits and the perceived emotions labeled using Mechanical Turk. Some of the novel components 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 and obtains 74.10% 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 field of emotion modeling, the 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 present the EWalk dataset in Section 5. We compare the performance of our method with state-of-the-art methods in Section 5.

**2. Related Work**

In this section, we give a brief overview of previous work 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. Kleinsmith et al. [29] used a 3D space with arousal, valence, and action tendency as the three affective dimensions. In the PAD model of emotions, Mehrabian [36] used a 3D continuous space of Pleasure, Arousal, and Dominance, whereas Ekman and Wallace [17] treated emotion as a point on a 2D continuous space of arousal and valence. Mikels et al. [38] and Morris [39] investigated a mapping between the PAD model and discrete emotional models. For example, discrete emotions of anger, happiness, pride are
Figure 2. Overview: We present an overview of our emotion classification algorithm. Given an RGB video of an individual walking, we use a state-of-the-art 3D human pose estimation technique [10] 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 motion-captured gait datasets. We also compute affective features consisting of both posture and movement features using psychological characterization [8, 26]. We concatenate these affective features with deep features and classify the combined features into 4 basic emotions using a Random Forest classifier.

related to high arousal, whereas sadness, relief, and contentment are related to low arousal. In this paper, we identify four discrete emotions (happy, angry, sad, and neutral) from the walking motion and gaits while also identifying the values of the affective dimensions of valence and arousal.

2.2. Body Expression of 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 [2, 44]. Expression of emotion in various activities such as knocking [21], dancing [14], playing musical instruments [11, 6], walking [26], etc. has also been studied [27]. Kleinsmith et al. [29] identified affective dimensions of arousal, valence, and action tendency that human observers use when discriminating between postures. Roether et al. [44] used a systematic approach and Omlor and Giese [41] 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. 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 emotions from body expressions automatically. These features are either extracted using purely statistical techniques or using techniques that are inspired from psychological studies. Some approaches focused on specific activities such as dancing, knocking [21], walking [26], games [28], etc., whereas some other approaches used a more generalized approach [8, 54]. Some approaches have combined both facial and body expressions [43, 51, 35, 50, 55, 7, 23, 25]. Some approaches found emotions in body expressions with the help of neutral expressions [44]. Crenn et al. [9] generated neutral movements from expressive movements and then identified the emotion in the expressive movement. Karg et al. [26] examined gait information for person-dependent affect recognition using motion capture data of a single walking stride. As is the case for most of these techniques, our approach is based 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 for identifying perceived emotions from RGB videos.

3.1. Emotion and Affective States

Most of the previous literature has modeled emotions as either discrete categories or as points in a continuous space of affective dimensions. Discrete categories include basic emotions such as anger, disgust, fear, joy, sadness,
and surprise as well as other emotions such as pride, depression, etc. On the other hand, Ekman and Wallace [17] used “affects” to represent emotions. *Affect* is a key characteristic of emotion and is defined as a 2-dimensional space of: (1) valence, the pleasure-displeasure dimension, and (2) arousal, the aroused/excited-sleep dimension [45]. All discrete emotions can be represented by points in this 2D affect space (Fig 3). In this paper, we use both discrete and continuous emotion representations. We use 4 basic emotions (happy, angry, sad, and neutral) representing emotional states that last for an extended period of time and are more abundant during walking [34]. These 4 emotions capture the spectrum of the affective space and a combination of them can be used to represent other emotions [38].

![Figure 3. Affect Space and Discrete Emotions: All discrete emotions can be represented by points on a 2D affect space of Valence and Arousal [32, 17].](image)

3.2. Notation

For our formulation, we represent a human with a set of 16 joints, as shown in 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 $\tau$ 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 $e$.

3.3. Overview

Our real-time perceived emotion prediction algorithm is based on a data-driven approach. We present an overview of our approach in Figure 2. During the offline training phase, we use multiple motion-captured gait datasets and extract affective features. These affective features are based on psychological characterization [8, 26] 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 [10]. 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.

3.4. Affective Feature Computation

For an accurate prediction of an individual’s affective state, both posture and movement features are essential [27]. Features in the form of joint angles, distances, and velocities of the joints, and space occupied by the body have also been used for recognition of emotions and affective states from gaits [8, 27]. Based on these psychological findings, we compute affective features that include both the posture and the movement features. We list these gait features in Table 1 and describe them in detail below.

We represent the extracted affective features of a gait $G$ 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.

3.4.1 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 4.4). We define posture features of the following type:

- **Volume**: According to Crenn et al. [8], body expansion conveys positive emotions whereas 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 $F_{area,t} \in \mathbb{R}^2$. 

![Figure 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.](image)
• 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 [8, 26]. We model this by the angles extended by different joints at the neck $F_{angle,t} \in \mathbb{R}^5$.

We also include stride length as a posture feature. Longer stride lengths convey anger and happiness and shorter stride lengths convey sadness and neutrality [26].

Suppose 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:

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

$$F_p = \frac{\sum_t F_{p,t}}{\tau} \cup s,$$  \hspace{1cm} (2)

### 3.4.2 Movement Features

Psychologists have shown that motion is an important characteristic for the perception of different emotions [4]. High arousal emotions have rapid and increased movements compared to the 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 the 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 $t$.

Since faster gaits are perceived as happy or angry whereas slower gaits are considered sad [26], 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 $F_{m,t}, t = \{1, 2, .., \tau\}$:

$$F_m = \frac{\sum_t F_{m,t}}{\tau} \cup gt,$$  \hspace{1cm} (3)

### 3.4.3 Affective Features

We combine posture and movement features and define affective features $F$ as:

$$F = F_m \cup F_p,$$  \hspace{1cm} (4)

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volume</td>
<td>Bounding box</td>
<td>Posture Features</td>
</tr>
<tr>
<td>Angle</td>
<td>At neck by shoulders</td>
<td></td>
</tr>
<tr>
<td></td>
<td>At right shoulder by neck and left shoulder</td>
<td></td>
</tr>
<tr>
<td></td>
<td>At left shoulder by neck and right shoulder</td>
<td></td>
</tr>
<tr>
<td></td>
<td>At neck by vertical and back</td>
<td></td>
</tr>
<tr>
<td></td>
<td>At neck by head and back</td>
<td></td>
</tr>
<tr>
<td>Distance</td>
<td>Between right hand and the root joint</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Between left hand and the root joint</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Between right foot and the root joint</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Between left foot and the root joint</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Between consecutive foot strikes (stride length)</td>
<td></td>
</tr>
<tr>
<td>Area</td>
<td>Triangle between hands and neck</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Triangle between feet and the root joint</td>
<td></td>
</tr>
<tr>
<td>Speed</td>
<td>Right hand</td>
<td>Movement Features</td>
</tr>
<tr>
<td></td>
<td>Left hand</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hand</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Right foot</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Left foot</td>
<td></td>
</tr>
<tr>
<td>Acceleration Magnitude</td>
<td>Right hand</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Left hand</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hand</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Right foot</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Left foot</td>
<td></td>
</tr>
<tr>
<td>Movement Jerk</td>
<td>Right hand</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Left hand</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hand</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Right foot</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Left foot</td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>One gait cycle</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Our Affective Feature Computation Algorithm: We extract affective features from an input gait using emotion characterization in visual perception and psychology literature [26, 8]. Since both posture and movement features are essential for an accurate prediction of an individual’s affective state [27], we compute affective features that include both the posture and the movement features.

### 4. Perceived Emotion Identification

We use a vanilla LSTM network [20] that models the temporal dependencies in the gait data. We chose an LSTM network to model deep features of walking because it captures the temporal geometric consistency and dependency.
among video frames for gait modeling [33]. We describe the details of the training of the LSTM in this section.

4.1. Datasets

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

- Human3.6M [24]: This dataset is acquired by recording the performance of 5 female and 6 male subjects and contains motion-captured gaits from 14 videos of the subjects walking.
- CMU [1]: This dataset contains motion-captured gaits from 49 videos of subjects walking with different styles.
- ICT [40]: This dataset contains motion-captured gaits from walking videos of 24 subjects.
- BML [34]: This dataset contains motion-captured gaits from 30 subjects (15 male and 15 female). Each subject provided 4 different walking styles resulting in 120 different gaits.
- SIG [56]: This is a dataset of 41 synthetic gaits generated using a local mixtures of autoregressive models algorithm.
- EWalk (Our novel dataset): We also collected a dataset of 94 videos and extracted gaits using 3D pose estimation. We present details about this dataset in Section 5.

4.2. Perceived Emotion Labeling

We obtained the perceived emotion labels for each gait using a web-based user study. We generated visualizations of each motion-captured gait using a skeleton mesh 4. For the EWalk dataset where the original videos were available, we presented the original videos to the participants. 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, and not from the facial expressions. We recruited 688 participants (279 female, 406 male, 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.

We asked each participant whether he/she perceived the gait video as happy, angry, sad, or neutral on a 5-point Likert scale ranging from Strongly Disagree to Strongly Agree. For each gait $G_i$ in the datasets, we calculated the mean of all participant responses $r_{i,j}^e$ to each emotion:

$$r_{i}^{e} = \frac{\sum_{j} r_{i,j}^{e}}{n_p},$$

where $n_p$ is the number of participant responses collected and $e$ is one of the four emotions: Angry, Sad, Happy, Neutral. We obtained the emotion label $e_i$ for a $G_i$ as follows:

$$e_i = e \mid r_{i}^{e} > \theta,$$

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

If there are multiple emotions with average participant responses greater than $r_{i}^{e} > \theta$, the gait is not used for training.

4.3. Classification

We obtain deep features from the trained LSTM and concatenate them with affective features. We 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 classifier at realtime to classify perceived emotions.

4.4. Realtime Perceived Emotion Recognition

At runtime, we take an RGB video as an input and use the trained classifier to identify the perceived emotions. We exploit a 3D human pose estimation algorithm, Time-PoseNet [10], which takes the video $V$ as input and returns
a gait \( G \) which contains the poses \( P_1, P_2, \ldots, P_\tau \) where \( \tau \) is the number of frames in the input video \( V \). TimePoseNet uses a semi-supervised learning method that utilizes the more widely available 2D human pose data [31] 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 person detector [5] on 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 \( P_1, P_2, \ldots, P_\tau \) represent the extracted output gait \( G \) from the input video \( V \). 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. 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.

5.1. Video Collection

The data collection was approved by Institutional Review Boards and the Office of Human Research Ethics and the subjects were informed that as part of the study, they will be recorded walking in front of the camera and their faces will be hidden to protect their identity. After agreeing to participate, they were asked to walk four times with different walking styles. To help them walk with different walking styles, they were suggested that they could assume different walking styles. To help them walk with different walking styles, they were suggested that they could assume different walking styles. To help them walk with different walking styles, they were suggested that they could assume different walking styles. To help them walk with different walking styles, they were suggested that they could assume different walking styles. To help them walk with different walking styles, they were suggested that they could assume different walking styles. To help them walk with different walking styles, they were suggested that they could assume different walking styles. To help them walk with different walking styles, they were suggested that they could assume different walking styles. To help them walk with different walking styles, they were suggested that they could assume different walking styles.

The subjects started at a distance of 7m from a stationary camera and walked towards it. The videos were later cropped to include a single walk cycle.

We recruited 24 subjects from a university campus. The subjects were from a variety of ethnic backgrounds and included both male (16 participants) and female (8 participants) subjects. We also recorded the videos in both indoor and outdoor environments.

5.2. 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. In Figure 6, we present the percentage of gaits that are perceived as belonging to each of the emotion categories (Happy, Angry, Sad, or Neutral). Our data is widely distributed across the 4 categories with the Happy category containing the most number of gaits (32.07%) whereas 16.35% gaits were perceived as Neutral which is the smallest category.

![Figure 6. 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.](image)

5.2.1 Affective Dimensions

We also 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 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 \([r_{happy}, r_{angry}, r_{sad}]\) and observed that the following two principal components describe 94.66% variance in the data:

\[
\begin{bmatrix}
PC_1 \\
PC_2
\end{bmatrix} = \begin{bmatrix}
0.67 \\
-0.35
\end{bmatrix} \begin{bmatrix}
-0.04 \\
0.86
\end{bmatrix}
\begin{bmatrix}
-0.74 \\
-0.37
\end{bmatrix}
\]

(7)

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.

5.2.2 Prediction of Affect

We use the principal components from Equation 7 to predict the values of arousal and valence dimensions. Suppose, the predicted probabilities by the Random Forest classifier are \( p(h), p(a), \text{ and } p(s) \) corresponding to the emotion classes happy, angry, and sad respectively. Then, we can obtain...
Table 2. Performance of Different Classification Methods: We analyze different classification algorithms to classify the concatenated deep and affective features. We observe an accuracy of 74.10% with the Random Forest classifier.

<table>
<thead>
<tr>
<th>Algorithm (Deep + Affective Features)</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM + Support Vector Machines (SVM)</td>
<td>70.04%</td>
</tr>
<tr>
<td>LSTM + Stochastic Gradient Descent (SGD)</td>
<td>66.84%</td>
</tr>
<tr>
<td>LSTM + Random Forest</td>
<td>74.10%</td>
</tr>
</tbody>
</table>

6. Results

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

6.1. Analysis of Different Classification Methods

We analyze different classification techniques to classify the combined deep and affective features and compare the resultant accuracy in Table 2. We use Random Forest classifier in the subsequent results because it provides the highest accuracy of 74.10% among all the classification methods.

6.2. Comparison with Other Methods

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

- Karg et al. [26]: 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. [52]: This method uses auto-correlation matrix of the joint angles at each frame and uses similarity indices for classification. The obtained good intra-subject accuracy but performs poorly for inter-subject database.

- Crenn et al. [8]: 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. [12]: This method uses manifold of symmetric positive definite matrices to represent body movement and classify them using Nearest Neighbors method.

Table 3. Accuracy: Our method with combined deep and affective features classified with a Random Forest classifier achieves an accuracy of 74.10%. We observe an improvement of 7.88% over state-of-the-art emotion identification methods and an improvement of 18.63% over a baseline LSTM-based classifier.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (Vanilla LSTM)</td>
<td>55.47%</td>
</tr>
<tr>
<td>Karg et al. [26]</td>
<td>39.58%</td>
</tr>
<tr>
<td>Venture et al. [52]</td>
<td>30.83%</td>
</tr>
<tr>
<td>Crenn et al. [8]</td>
<td>66.22%</td>
</tr>
<tr>
<td>Crenn et al. [9]</td>
<td>40.63%</td>
</tr>
<tr>
<td>Daoudi et al. [12]</td>
<td>42.52%</td>
</tr>
<tr>
<td><strong>Our Method (Deep + Affective Features)</strong></td>
<td><strong>74.10%</strong></td>
</tr>
</tbody>
</table>

- Crenn et al. [9]: 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.

We also compare our results to a baseline where we use the LSTM to classify the gait features into the 4 emotion classes. Table 3 provides the accuracy results of our algorithm and comparisons with other methods. These methods require input in the form of 3D human poses and then 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 an 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).

7. Conclusion, Limitations, and Future Work

We presented an end-to-end method to classify perceived emotions of individuals based on their walking videos. Our method is based on learning deep features computed using LSTM and also exploits psychological characterization to compute affective features. 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 74.10%, which is an improvement of 18.63% over vanilla LSTM (i.e. only use of deep features) and offers an improvement of 7.88% over state-of-the-art emotion identification algorithm. Our approach is also the first approach to provide an realtime end-to-end 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 ethnic backgrounds in both indoor and outdoor environments.

There are some limitations our approach. The accuracy of our algorithm depends on the accuracy of the 3D human
pose estimation and gait extraction algorithms. Therefore, the 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 in case of occlusions in the video. We assume that the walking motion is natural and does not involve any accessories (e.g., suitcase, mobile phones, 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 face expressions.

References


