Robotica (2023), **41**, pp. 1452–1465
doi:10.1017/S0263574722001813

RESEARCH ARTICLE

Data augmentation by separating identity and emotion representations for emotional gait recognition

Weijie Sheng[1,2], Xiaoyan Lu[3] and Xinde Li[2,3,*]

[1]Yangzhou Collaborative Innovation Research Institute Co., Ltd., Institute of Shenyang Aircraft Design and Research, Yangzhou, 225000, China, [2]Key Laboratory of Measurement and Control of CSE Ministry of Education, School of Automation, Southeast University, Nanjing, China, and [3]School of Cyber Science and Engineering, Southeast University, Nanjing, China
*Corresponding author. Email: xindeli@seu.edu.cn

Received: 27 June 2022; **Revised:** 26 October 2022; **Accepted:** 24 November 2022; **First published online:** 6 February 2023

Keywords: human–robot interaction, data augmentation, gait recognition, emotion recognition

Abstract

Human-centered intelligent human–robot interaction can transcend the traditional keyboard and mouse and have the capacity to understand human communicative intentions by actively mining implicit human clues (e.g., identity information and emotional information) to meet individuals' needs. Gait is a unique biometric feature that can provide reliable information to recognize emotions even when viewed from a distance. However, the insufficient amount and diversity of training data annotated with emotions severely hinder the application of gait emotion recognition. In this paper, we propose an adversarial learning framework for emotional gait dataset augmentation, with which a two-stage model can be trained to generate a number of synthetic emotional samples by separating identity and emotion representations from gait trajectories. To our knowledge, this is the first work to realize the mutual transformation between natural gait and emotional gait. Experimental results reveal that the synthetic gait samples generated by the proposed networks are rich in emotional information. As a result, the emotion classifier trained on the augmented dataset is competitive with state-of-the-art gait emotion recognition works.

1. Introduction

Human emotions can be perceived not only through explicit facial expressions [1], voice information [2], or text cues [3], but also through implicit body language, including eye movements [4], body postures [5], and gait traits [6]. Nonverbal communication plays a major role in recent human–robot interaction (HRI) [7]. Body language delivers nonverbal signals that can provide important cues for a person's mental and physiological state and intentions. Gait is a unique biometric trait that can be obtained from a distance without individuals' attention or cooperation [8]. Meanwhile, ref. [9] has reported that a human's walking pattern is difficult to imitate or intentionally deceive. Human gait conveys significant information that can be used to identify people and recognize emotions [10]. HRI can not only transfer mechanical power [11, 12] but also emotional signals [13] between the human and robotic machines. Emotion is a ubiquitous element of HRI. Compared to traditional emotion detection biometrics, such as facial expression, voice, and physiological signals, gait provides a new source and can be obtained from a long distance without the subject's cooperation. Gait fills the emotion recognition field gaps when other traits are infeasible in long-distance observation. Recent paper [14] presented a review of current gait emotion recognition research and possible future developments. There are many application scenarios based on gait-based emotion recognition such as psychology diagnosis, emotionally aware robot [13], customer services, interactive games, and virtual reality [15]. This field has great potential to be improved to a higher level to support a broader range of applications.

Understanding human emotion through facial expressions has been well studied [7]. However, the ability to rely on body language to perceive emotion becomes important when a person is not directly

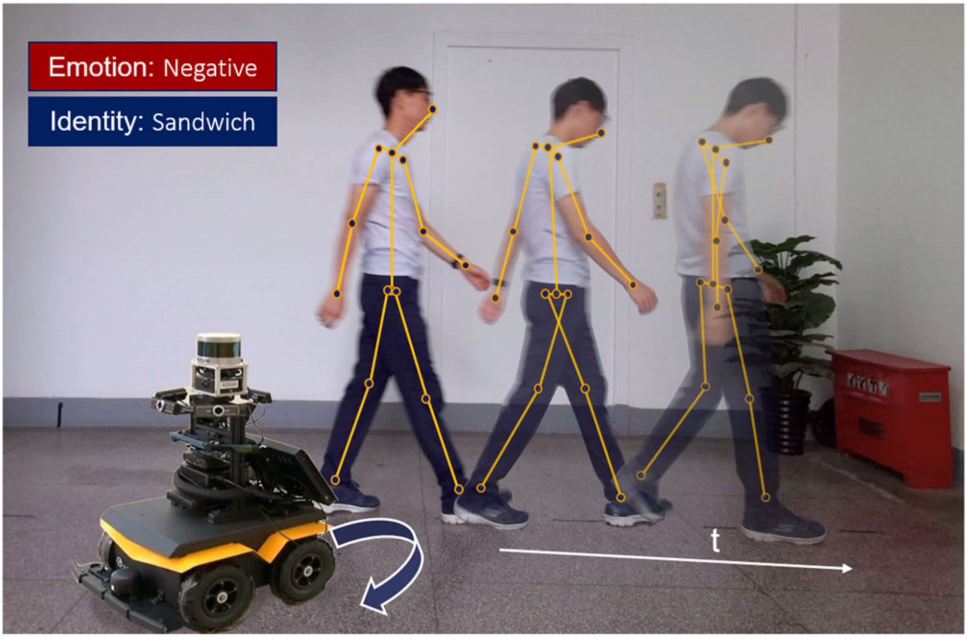


Figure 1. *We present a data augmentation method for gait emotion and identity recognition to perform emotionally aware robot navigation.*

facing the robot, or facial expressions are not visible from a distance. Recent work [16] observed that collaborative robots can improve interaction and performance by understanding the movement intentions of human operators. For example, in space-sharing application scenarios such as hospitals, airports, and shopping malls, robots can understand the intention of pedestrians through gait recognition of emotional states and determine whether to provide friendly navigation services or to wisely avoid causing untimely disturbances (as illustrated in Fig. 1). It is expected that the emotionally aware robot can navigate safely through crowds without causing discomfort to nearby pedestrians. Meanwhile, identity recognition is a prerequisite for robots to provide personalized services. Since each person's emotional expression will have individual differences, having personalized emotion understanding capability is the key to achieving intelligent HRI. Gait-based identity and emotion recognition as an aspect of nonverbal communication can help analyze and understand human intentions.

Previous work [17] discovered that variances in a person's emotional states during training and testing datasets can degrade the recognition performance in an identity verification task from gait. Moreover, several research [18, 19] have indicated that through multi-task learning (MTL), the emotion recognition task can benefit from training with secondary related tasks. However, most of the existing works learn identity representations and emotional feature separately and treat them independently to each other. In ref. [10], models trained with MTL for gait-based emotion and identity recognition have shown additional performance improvements. They believed gait-based identity and emotion recognition are interrelated tasks that are favorable for jointly learning. The MTL models entangle information between the tasks to capture the joint dependencies from the multi-labels of the training data [20]. However, there has also been a noticeable absence of studies on MTL for emotional gait, mainly due to the lack of gait datasets annotated with both emotion and identity labels.

Deep learning models often require a great quantity of data for training to obtain good predictions or classification performance. Nevertheless, the procedure of collecting gait samples is often costly and time-consuming, making it very difficult to obtain a well-annotated dataset with sufficient samples [21]. It is particularly prominent in gait emotion recognition tasks because the annotation of emotional categories is ambiguous and vulnerable to subjective factors [22]. To reduce the impact of personal

https://doi.org/10.1017/S0263574722001813 Published online by Cambridge University Press

subjectivity, it is often necessary to recruit multiple annotators to strengthen the annotation reliability. However, in some cases, it is impossible to ensure the accuracy of the results even for experienced annotators [23]. Therefore, the insufficient data problem severely hinders the application of gait emotion recognition in reality.

With the increasing applications of deep learning in emotion recognition tasks, data augmentation via generative adversarial networks (GANs) on the training set to augment the original to obtain improvement for recognition results may offer a solution for this challenge. Using a data augmentation strategy similar to ours, ref. [24] recorded hundreds of annotated gait videos and augmented them with synthetic gaits built on conditional variational autoencoder (CVAE) to increase the emotion classification accuracy.

Traditional methods for data augmentation are generally based on GANs or autoencoders, such as conditional GANs (cGANs) [25] or conditional VAE (CVAE) [26]. The decoder of CVAE produces random samples from a conditional distribution and generates synthetic data to learn different distributions for the specific categories [27]. Pix2pix [28] can generate high-quality image results in the case of paired training data using a cGAN to implement the mapping function. To train with unpaired data, CycleGAN [29], DiscoGAN [30], MUNIT [31], and StarGAN [32] exploit cycle consistency to constrain the training process. Applying data enhancement to gait emotion recognition, ref. [24] designed a gait generation network STEP, based on CVAE to generate thousands of synthetic samples.

Motivated by the achievements of emotional conversion in voice [33, 34] and face expression [35], we propose the emotional gait conversion approach to transform natural gaits into emotional gaits by separating identity and emotion representations for data augmentation. The contributions of this work can be summarized as follows:

- We introduce a MTL discriminator for gait identity and emotion joint learning, which takes into account nonverbal communication clues to enhance HRI.
- We propose a novel emotional gait conversion model with adversarial loss and cycle consistency loss to realize the mutual transformation between natural gait and emotional gait.
- We propose two kinds of data augmentation strategies by the emotional conversion model to increase the amount and diversity of the existing restricted dataset.
- We present an augmented synthetic dataset of human emotional gait, validated by a multitask classifier and achieved a corresponding 2.1% and 6.8% absolute increase in identity recognition and emotion recognition, respectively.

2. The proposed method

The main idea of this work is to increase the amount and diversity of the original limited dataset by transforming natural gaits into emotional gaits. We first extract gait trajectories from the original videos to represent the discriminative gait features. Then two autoencoders are trained to separate latent identity embedding and emotion-specific embedding using two auxiliary classifiers to guarantee the minimal mutual information related to each other. In the second stage, we propose a novel cycle consistency GAN to realize the synthesis of the separated identity and emotion features from different samples. After carrying out this data generation process, we can train an enhanced gait emotion classifier on the augmented dataset to obtain a significantly improved performance. Figure 2 illustrates how we incorporate our data augmentation method for gait emotion and identity recognition into an end-to-end emotionally guided navigation pipeline.

2.1. Gait trajectories generation

In this work, the gait data were recorded by two Microsoft Azure Kinect DK sensors placed in front and on the side of the subjects. Kinect DK is a convenient body tracking toolkit to capture RBG image,

https://doi.org/10.1017/S0263574722001813 Published online by Cambridge University Press

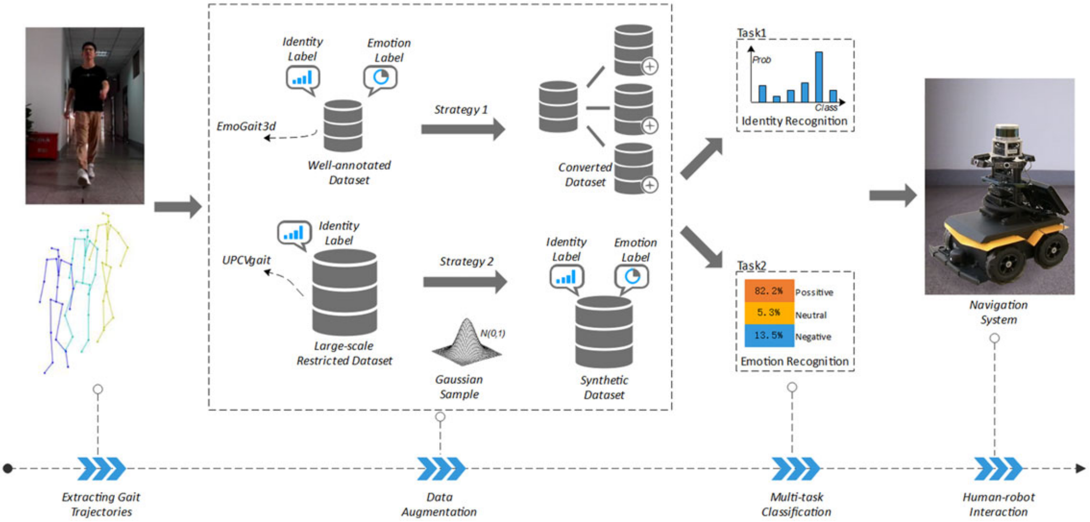


Figure 2. *An overview of the pipeline for emotionally aware robot navigation system using gait-based dataset augmentation method. The well-annotated dataset is augmented by the emotional conversion strategy. The large-scale restricted dataset is augmented by adapting the Gaussian sampling to generate different variants of emotion-labeled synthetic samples.*

depth information, and human skeleton coordinates all at once, reducing the need for sophisticated model extraction processes. By the body tracking function, we can extract a real-time data stream of the body joints, represented by 25 joint coordinates in a 3D space. We selected 20 joints with relatively large ranges of motion to represent the gait movement. Then, we concatenated the coordinates of each joint to form a continuous trajectory by the motion across time. Finally, to eliminate the impact of the distance variations between people and cameras, we normalized the coordinates using the distance between a subject’s hip and neck.

2.2. Learning separated representations

Let $x \in \mathcal{X}$ be a gait trajectory sequence and $\mathcal{X}$ be the collection of all the trajectories in the training data. In stage 1, E_{id} denotes the identity encoder and E_{em} denotes the emotion encoder. To learn separated identity and emotion representations, we employ two classifiers C_{id} and C_{em} with adversarial learning constraints on the feature encoders. These constraints ensure that changes in one factor cannot be predicted from another factor to realize independence between them. Based on the adversarial training concept, E_{em} maximizes the retention of emotional information and discards identity information by minimizing the negative log probability to differentiate the identities. On the other hand, the classifier C_{em} is trained adversarially to induce the encoder E_{id} to extract only identity-related features. We thus apply the loss:

$$\begin{aligned}\mathcal{L}_{cls}^{em} = &\sum - \log P_{C_{em}}\left(c_{em}^{x} \mid E_{em}(x)\right)\\ &+ \sum \log P_{C_{em}}\left(c_{id}^{x} \mid E_{id}(x)\right)\end{aligned} \tag{1}$$

$$\begin{aligned}\mathcal{L}_{cls}^{id} = &\sum - \log P_{C_{id}}\left(c_{id}^{x} \mid E_{id}(x)\right)\\ &+ \sum \log P_{C_{id}}\left(c_{em}^{x} \mid E_{em}(x)\right)\end{aligned} \tag{2}$$

https://doi.org/10.1017/S0263574722001813 Published online by Cambridge University Press

To perform random sampling at test time, we restrict the emotion feature representation to a conditionally independent Gaussian distribution, by introducing KL divergence loss to match the posterior distribution $p(z_{em}|x)$ to the prior $N(0, I)$. We thus apply the loss:

$$\mathcal{L}_{KL} = E\left[KL\left(z_{em}|x\|N(0, 1)\right)\right] \tag{3}$$

where $KL(p\|q)$ represents the Kullback–Leibler Divergence score and quantifies the difference between two given probability distributions p and q.

The generator G is trained to generate x' which is a reconstruction of x from the concatenation of emotion representation z_{em} and identity representation z_{id}, given the original emotion label c^x and target emotion label $c^{x'}$:

$$x' = G(E_{id}(x), E_{em}(x)) \tag{4}$$

By using both original and target label as conditional information, this restriction encourages all the converted data to be close to real data. The mean absolute error is minimized in training the generator. So the reconstruction loss is given:

$$\mathcal{L}_{rec} = \sum \left\|x' - x\right\|_1 \tag{5}$$

The full objective in stage 1 is deployed by the following equation:

$$\mathcal{L}_1^{total} = \lambda_1^{rec}\mathcal{L}_{rec} + \lambda_1^{KL}\mathcal{L}_{KL} + \lambda_1^{em}\mathcal{L}_{cls}^{em} + \lambda_1^{id}\mathcal{L}_{cls}^{id} \tag{6}$$

which integrates the above losses and the hyperparameters λ_1s control the importance of each term. The encoders and the discriminators are trained alternatively.

2.3. Cycle-consistent GANs

Here, to learn an emotional gait conversion with paired emotional gait samples using the separated representation of identity in stage 1, we propose a cycle consistency technique to exploit the further features for cyclic reconstruction. Let $x, y \in \mathcal{X}$ be the two sampled gait trajectory sequences (as illustrated in Fig. 3). c_{em}^x and c_{id}^x denote the emotion label and identity label of sequence x, respectively, and c_{em}^y and c_{id}^y denote the labels of sequence y. We encode them into vector $\{v_{id}^x\}$ and $\{v_{id}^y, v_{em}^y\}$ by the pretrained encoders E_{em} and E_{id}. We then perform the generation process by reassembling the extracted identity vector v_{id}^x and the emotion vector v_{em}^y into a combined representation of a synthetic sample z:

$$z = G\left(v_{id}^x, v_{em}^y\right) \tag{7}$$

We further encode z into $\{v_{em}^z, v_{id}^z\}$. Then, a cycle consistency loss $\mathcal{L}_{cycl}^{id}$ for v_{id}^x, v_{id}^y, and v_{id}^z, the same structure as triplet loss [36], is designed to enforce identity preservation:

$$\mathcal{L}_{cycl}^{id} = \sum \left[\left\|v_{id}^z - v_{id}^x\right\|_2^2 - \left\|v_{id}^z - v_{id}^y\right\|_2^2 + \alpha\right]_+ \tag{8}$$

where α is the value of the margin in two terms. Another cycle consistency loss $\mathcal{L}_{cycl}^{em}$ between v_{em}^y and v_{em}^z is used to enforce emotion preservation:

$$L_{cycl}^{em} = \sum \left\|v_{em}^z - v_{em}^y\right\|_2^2 \tag{9}$$

We employ the reconstruction loss $\mathcal{L}_{rec}$ only when $c_{id}^x = c_{id}^y$:

$$\mathcal{L}_{rec} = \begin{cases} \sum \|z - x\|_1, & c_{id}^x = c_{id}^y \\ 0, & Otherwise \end{cases} \tag{10}$$

We also impose domain adversarial losses by a unified MTL discriminator D_{MTL} to discriminate between natural gaits and generated gaits in each conversion process and distinguish the generated data

https://doi.org/10.1017/S0263574722001813 Published online by Cambridge University Press

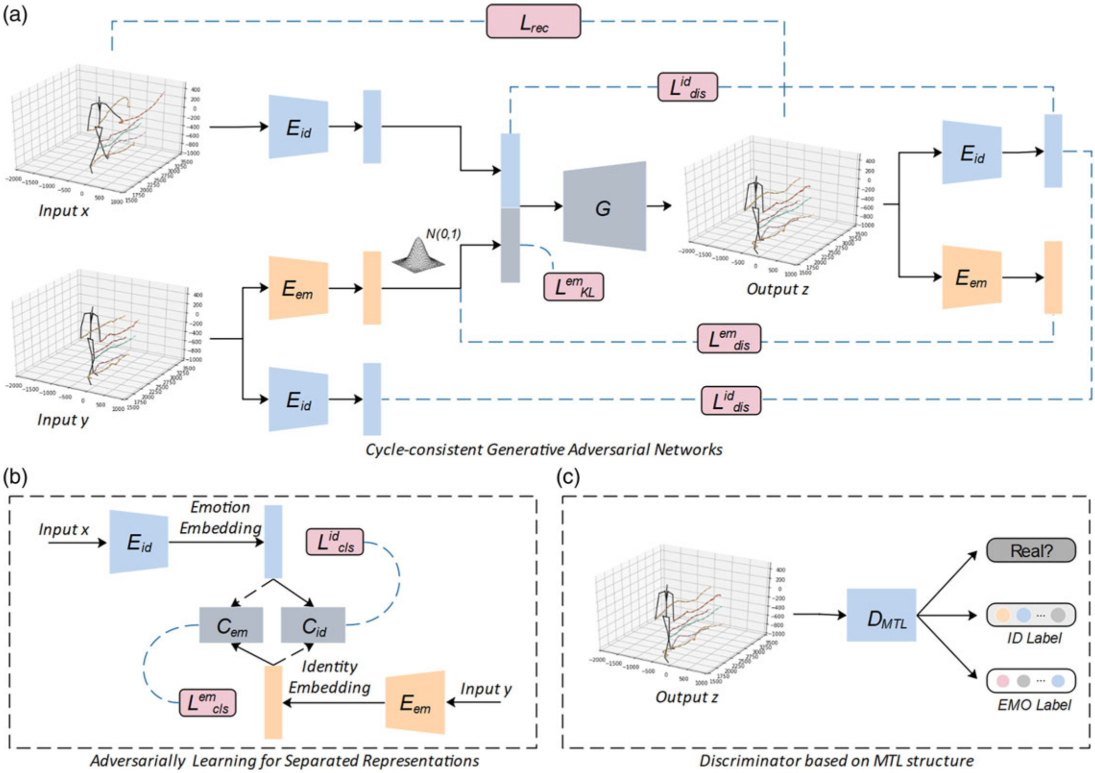


Figure 3. *The framework of the proposed adversarial learning network for emotional gait dataset augmentation.*

in both the emotion and identity domains. This adversarial MTL loss can be expressed as:

$$\begin{aligned}\mathcal{L}_{MTL} = & \sum (\log (D_{MTL}(x)) + \log (D_{MTL}(y))) \\ & + \sum \log (1 - D_{MTL}(z)) \\ & - \sum \log P_{D_{MTL}} \left(c^{y}_{em} \mid E_{em}(z)\right) \\ & - \sum \log P_{D_{MTL}} \left(c^{x}_{id} \mid E_{id}(z)\right)\end{aligned}$$

Here, we also restrict the emotion attribute representation to a conditionally independent Gaussian distribution, by introducing KL divergence loss L_{KL}. The overall loss is a weighted sum of the above losses:

$$\begin{aligned}\mathcal{L}^{total}_{2} = & \lambda^{rec}_{2} \mathcal{L}_{rec} + \lambda^{MTL}_{2} \mathcal{L}_{MTL} + \lambda^{id}_{2} \mathcal{L}^{id}_{cycl} \\ & + \lambda^{emo}_{2} \mathcal{L}^{em}_{cycl} + \lambda^{KL}_{2} \mathcal{L}_{KL}\end{aligned} \tag{11}$$

where hyperparameters λ_2s are the regularization weights.

2.4. Gait-based recognition with data augmentation

According to its own specific defects of the training datasets, we design two strategies for data augmentation. For the small-scale dataset with complete labels, data augmentation is implemented by disentangling and composing the emotion and identity feature vector from different people, as illustrated

https://doi.org/10.1017/S0263574722001813 Published online by Cambridge University Press

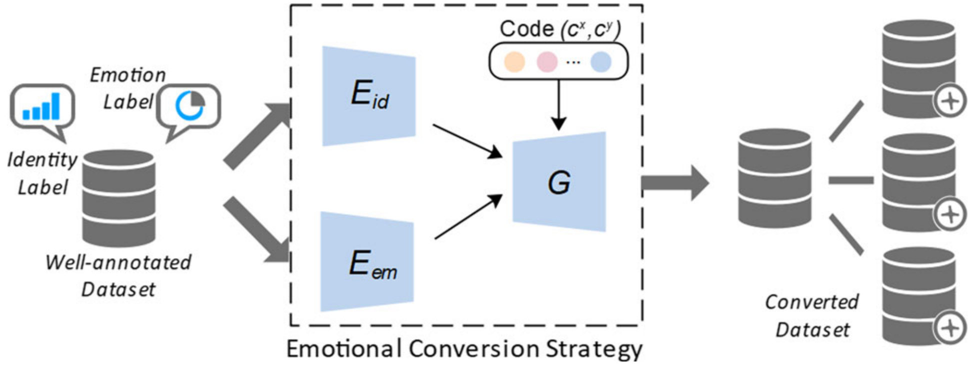


Figure 4. *Data augmentation by emotional conversion strategy. Data augmentation is implemented by disentangling and composing the emotion and identity feature vector from different people to improve the scale and variability of the original dataset.*

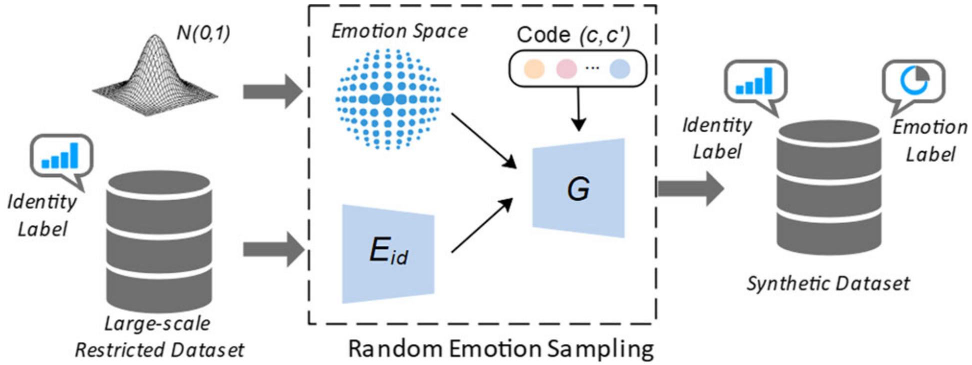


Figure 5. *Data augmentation by random emotion sampling. Our model could generate specific emotion vectors from the common emotion space by adapting the Gaussian stochastic sampling. With the random emotion vector, we can generate different variants of emotion-labeled synthetic samples to derive an augmentation for the target restricted dataset.*

in Fig. 4. In this strategy, we synthesize each target samples with three alternative emotion vectors and the specific identity vectors to generate the same amount of each emotional samples. For the large-scale dataset with restricted labels, data augmentation is implemented by random emotion sampling, which is shown in Fig. 5. With the random emotion vector, we can generate different variants of emotion-labeled samples to increase the amount and diversity of the original dataset.

After applying data augmentation strategies, we can easily train a multitask discriminator on the augmented and original dataset as our recognition model and then assess the quality of these synthetic samples through the discriminator. As illustrated in Fig. 3(c), the discriminator D_{MTL} attempts to discriminate between natural gaits and generated gaits in each conversion process and distinguish the generated data in both of the emotion and identity domains.

3. Experiment

3.1. Data preparation

To evaluate our approach and measure the quality of the synthetic dataset, we conducted several experiments for verification tasks on the public UPCV gait (K1&K2) dataset and multi-class labeled EmoGait3d dataset.

https://doi.org/10.1017/S0263574722001813 Published online by Cambridge University Press

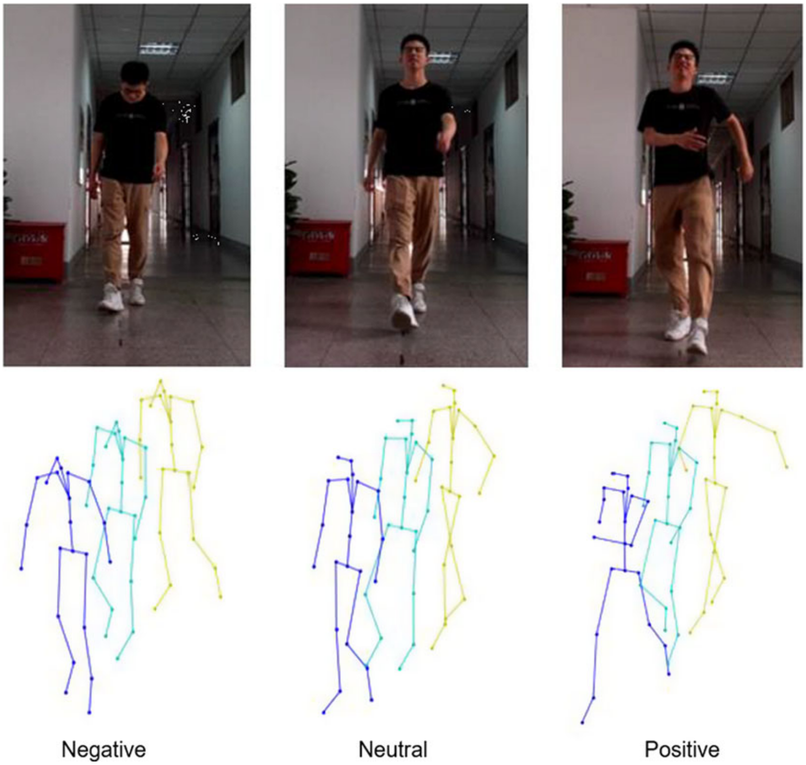


Figure 6. *Images and skeleton joints of three different emotion from the EmoGait3d dataset.*

The UPCV gait dataset contains 60 subjects in total from two subsets: UPCV gait K1 [37] and UPCV gait K2 [38]. The former contains five gait sequences for 30 participants captured using the Microsoft Kinect V1 sensor, and the latter captured by the Kinect V2 sensor contains a total of 300 sequences from 30 walkers. Each person walks in a straight line at a normal speed. The sensor maintains a fixed viewpoint in the walking direction at a frame rate of 30 fps. While, samples in UPCV gait are only annotated with identity labels and hardly perceived for their emotion categories through walking characteristics. Here, we regard the dataset as a large-scale restricted dataset and annotate all the samples with the emotion label of a neutral state. Because each gait sequence has a varied temporal duration, we extract 32-frame subsequences with a three-frame interval from each original sequence. With the pose estimation algorithm, we estimate the joint coordinates from each continuous 32-frame image sequence to obtain a $32 \times 20 \times 3$ trajectories vector as a gait sample. In the UPCV gait dataset, we can get a set of 15,053 samples as the original dataset. By implementing the data augmentation of random emotion sampling, each neutral sample can be transferred into positive, neutral, and negative samples. We finally obtained a set of 15053×3 synthetic samples as the augmented dataset of UPCV gait.

The EmoGait3d dataset is built to validate the effectiveness of the MTL structure by jointly training on multiple gait-related tasks. It consists of 1484 real-world gait videos annotated with identity labels and emotion labels. We recruited 27 volunteers (10 female and 17 male, aged 18–35 years) from campuses and took RGB and depth videos with two Microsoft Azure Kinect DK sensors. Each participant was asked to walk multiple times under three emotions (shown in Fig. 6). Participants' emotions were elicited by watching emotional movie clips, which were selected prior to the experiments based on their questionnaires. After completing the data collection, subjects were required to rate their emotional state during walking with a value on a scale from 1 to 10. When the emotion evoked by the film was consistent with the subject's self-assessment emotion, and the rating score was higher than 8, the video could be labeled as the elicited emotion. Otherwise, it would be marked as an invalid video. With the proposed

https://doi.org/10.1017/S0263574722001813 Published online by Cambridge University Press

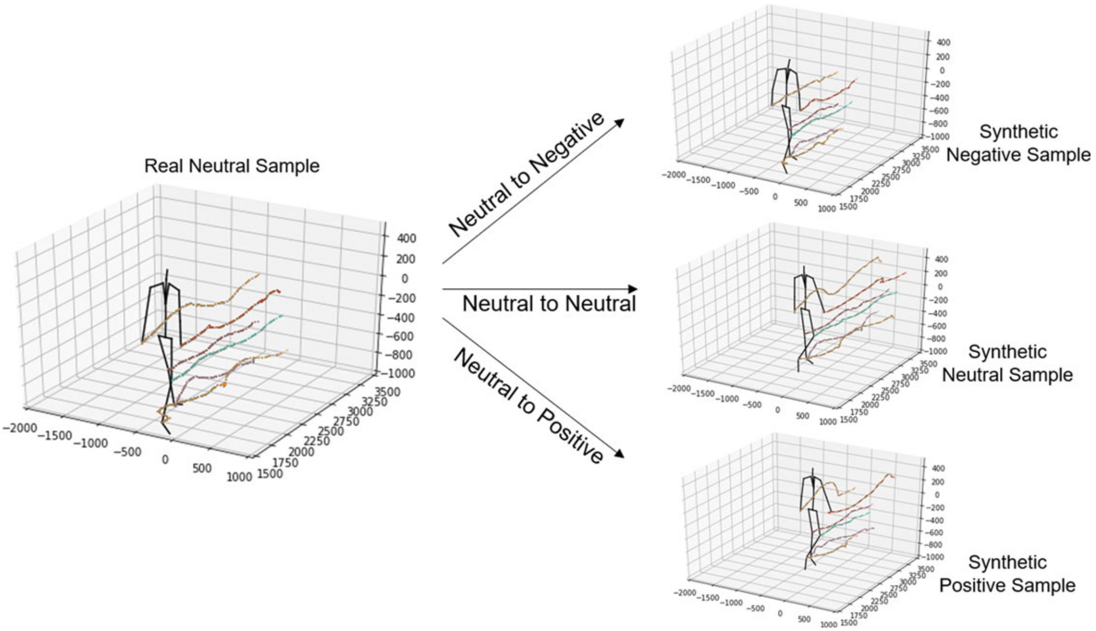


Figure 7. *Synthetic emotional gait trajectories. A real gait sample from the EmoGai3d database is represented on the left, and the synthetic target emotional gaits are shown on the right.*

data augmentation method, we generated 1484 × 3 synthetic emotional samples (shown in Fig. 7), by separating identity and emotion representations from the original EmoGait3d dataset for each of the three emotion categories.

3.2. Implementation details

The network architecture is illustrated in Fig. 3 with details listed in Table I. The encoders take 32-dimensional gait skeleton sequences as input and learn disentangled identity and emotion representations. In the emotion encoder, we apply instance normalization (IN) to removes the identity information while preserving the emotion information. The identity encoder provides the global identity information μ_i and σ_i to the generator by adaptive instance normalization (AdaIN) layer before activation. μ_e and σ_e denote the channel-wise mean and standard variation of the emotion feature vector e. The formula for a layer is given as follows:

$$AdaIN(e, i) = \sigma_i \left(\frac{e - \mu_e}{\sigma_e} \right) + \mu_i \tag{12}$$

The generator and encoders are implemented with recurrent layers and 1d convolutional layers to capture temporal dependencies and spatial patterns, respectively. Then, the temporal and spatial features are combined to represent a more discriminative embedding vector to feed the dense layers.

The experiments are conducted on a system with two GTX TITAN XP GPUs. We first train the encoders to learn separated identification and emotion representations from 32-dimensional gait skeletal sequences. Then, the separated features are then combined to generate the synthetic emotional sample by dense layers. We use the Adam optimizer with a learning rate of 0.001. The batch size is set at 128. To reduce overfitting, we use the dropout approach with a dropout rate of 0.5. The discriminator and generator are updated with a 1:5 iteration frequency. We selected the parameters by using the early stopping criterion. If the validation error does not improve before the training epoch reaches the set value, the training procedure will be terminated earlier. We first pretrained the identity and emotion

Table I. *Network architecture. C-K indicates convolution layer with kernel size K. IN is instance normalization. ReLU indicates ReLU activation and FC indicates fully connected layer.*

Emotion encoder	
Conv1d, IN, ReLU	C-128-5
Downsample1d × 2, IN	C-256-5, C-512-5
Dense block1d × 4, IN, Res	FC-512
Recurrent layer	Bidirectional GRU-512
Combine layer	Dense output + Recurrent output
Dense layer × 2	FC-512, FC-256
Identity Encoder	
Conv1d, ReLU	C-128-5
Downsample1d × 2	C-256-5, C-512-5
Dense block1d × 4, Res	FC-512
Recurrent layer	Bidirectional GRU-512
Combine layer	Dense output + Recurrent output
Dense layer × 2	FC-512, FC-256
Decoder/Generator	
Conv1d, AdaIN, ReLU	C-512-3
Upsample1d × 2, AdaIN	C-512-5, C-256-5
Dense block1d × 4, AdaIN, Res	FC-256
Recurrent layer	Bidirectional GRU-256
Combine layer	Dense output + Recurrent output
Dense layer × 2	FC-256, FC-128
Classifier	
Conv1d, IN, ReLU	C-128-5
Downsample1d × 3, IN, ReLU	C-256-5, C-512-5, C-1024-5
Dense layer × 2, Softmax	FC-512, FC-N/FC-3
MTL Discriminator	
AT-GCN × 3, BN, ReLU	C-128, C-256, C-512
Dense block1d × 3, BN, Res	FC-512, FC-256, FC-128
Dense layer, Softmax	FC-N, FC-3, FC-1 (real/fake)

classifiers with $\mathcal{L}_{cls}^{emo}$ and $\mathcal{L}_{cls}^{id}$ in Eq. (1) and (2) for 10,000 mini-batches. Then we train the models in stage 1 and stage 2 successively for 30,000 mini-batches and 20,000 mini-batches. Also inference speed is an important aspect to evaluate the model. The preprocessing for pose estimation takes most of the time. The network inference procedure is relatively faster, which takes about 0.17 ms for each frame. Our model has low complexity and need to be optimized for real-world applications.

3.3. Objective evaluation

We evaluate the quality of the synthetic samples by comparing the recognition performance of the original and augmented EmoGait3d using the same setting of MTL classifiers. As shown in Table II, noticeable performance improvements of 2.1% and 6.8% can be observed by augmenting the original dataset. The experimental results show that samples generated by our model carry discriminative information that contributes to consistently higher performance for gait-based identity and emotion

https://doi.org/10.1017/S0263574722001813 Published online by Cambridge University Press

Table II. *Results of the identity and emotion classification on the original and augmented dataset.*

		Emo Acc (%)			
	Id Acc (%)	**Neg**	**Neu**	**Pos**	**Avg**
EmoGait3d	91.4	87.9	81.5	87.7	85.7
EmoGait3d augmented	93.5	93.7	90.2	93.9	92.5
UPCV gait	98.2	-	-	-	-
UPCV gait augmented	96.3	90.0	88.7	91.1	89.9
EmoGait3d+UPCV gait augmented	94.9	91.6	89.8	93.3	91.6

Table III. *Comparison of different generative models. Accuracies are computed using the same MTL classifier. The best results are marked in bold.*

		Emo Acc (%)			
	Id Acc (%)	**Neg**	**Neu**	**Pos**	**Avg**
CVAE [26]	92.3	88.3	84.6	88.6	87.2
CGAN [25]	90.8	87.9	79.7	89.4	85.7
CVAE-GAN [39]	93.2	91.0	86.5	89.8	89.1
Cycle-GAN [29]	93.7	92.1	88.2	91.1	90.4
MUNIT [31]	93.1	91.9	89.0	92.6	91.2
StarGAN v2 [32]	**93.9**	91.7	88.4	92.9	91.0
Ours(Stage 1)	92.2	86.9	85.8	89.1	87.3
Ours(Stage 2)	93.1	91.3	88.0	92.3	90.5
Ours(Stage 1 + 2)	93.5	**93.7**	**90.2**	**93.9**	**92.5**

recognition. There is no emotion annotation in the original UPCV gait dataset, so we cannot get the emotion recognition results. While after data augmentation, the UPCV gait dataset is transferred to an emotional gait dataset with no significant reduction in the discriminative identity features.

To highlight the effectiveness of our model, we also trained respective MTL classifiers for identity and emotion recognition using augmented data from CVAE, CGAN, CVAE-GAN, CycleGAN, StarGAN, and MUNIT and compared their performance, as shown in Table III. All the settings of baseline generative data augmentation approaches and classifiers are the same as ours for a fair comparison. The performance of our model obtains the best results of them. In contrast to these generative models, our model employs the separated features, and cycle consistency loss clearly outperforms all the others, especially for the gait emotion recognition task, which is 1.3% better than the baseline model MUNIT in average recognition accuracy. We can also observe that the model's performance without stage 1 or disentangle learning process significantly declines, which shows the prominent effect of the two-stage emotional gait conversion model intuitively.

Both CVAE and CGAN can generate synthetic data similar to the training data. For CVAE, the generated gait sample is relatively stable, but the curves tend to be straight lines to cheat the discriminator. For CGAN, the diversity of the generated sample is better, but the naturalness of the generated sample is poor. Since CVAE-GAN combines a variational autoencoder with GAN, the quality of the generated data is better than CVAE and CGAN. Without the cycle loss as Cycle-GAN, the CVAE-GAN model fails to capture the temporal details of gait trajectories. Due to the absence of a feature separating process, the performance of the synthetic sample generated by CycleGAN or StarGAN is also not ideal. MUNIT adopts a weaker form of cycle consistency constraint between the content and style spaces, the generated sample of which is deficient in temporal details.

https://doi.org/10.1017/S0263574722001813 Published online by Cambridge University Press

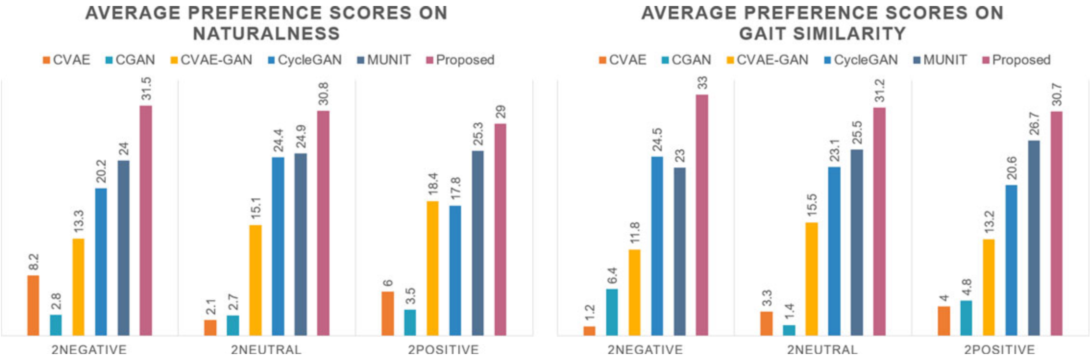
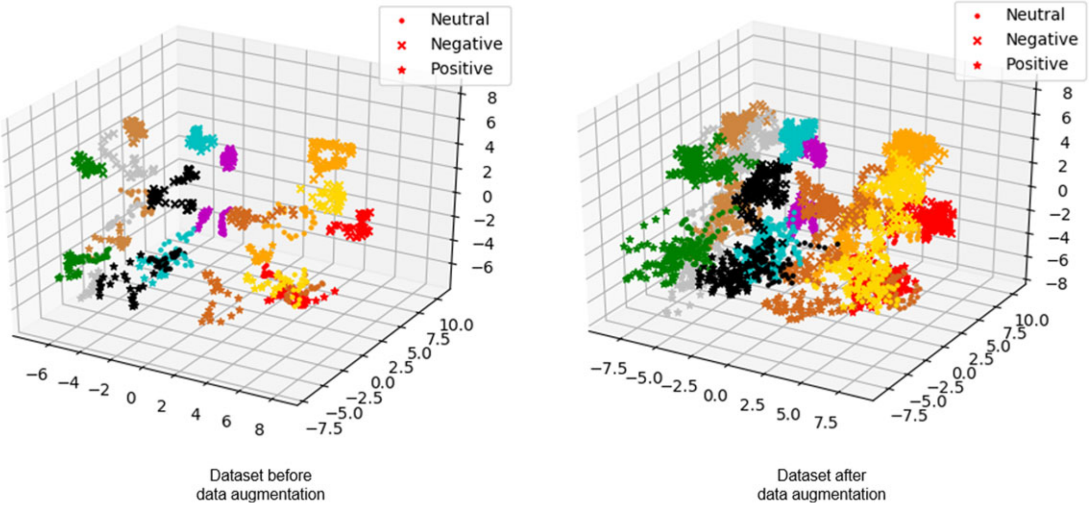


Figure 8. *Average preference scores on naturalness and similarity of synthetic samples of different generative models.*



Dataset before
data augmentation

Dataset after
data augmentation

Figure 9. *Visualization of the feature space after Principal Component Analysis (PCA) for the original and augmented EmoGait3d dataset. Three shapes of dots represent three kinds of emotional feature vectors, and the different colors correspond to different identities.*

3.4. Subjective evaluation and Discussion

We also performed subjective human evaluations for the synthetic gait. Twenty subjects were given pairs of converted samples in random order and asked which one they preferred in terms of two measures: the naturalness and the similarity in emotional characteristics of the converted gait trajectories. We computed the distance between 600 pairs of synthetic gait trajectories converted from 200 real samples. As shown in Fig. 8, we calculated average preference scores on these synthetic samples from source to target emotion. Higher values indicate higher quality of the synthetic sample after emotional conversion. The proposed model achieves the highest scores in terms of the naturalness and the similarity in emotional characteristics of the converted gait samples.

To evaluate the effect of our model, we further visualize the feature distribution of each emotion class from the original and enhanced EmoGait3d datasets. As shown in Fig. 9, we observe that almost all of the identity and emotion features for each type of synthetic sample are well generated, and the synthetic samples are well aligned with the authentic samples. It shows the effectiveness of learned features intuitively. The well-aligned data distributions are key in increasing the amount and diversity of the original EmoGait3d dataset to achieve improved accuracy for gait emotion recognition.

https://doi.org/10.1017/S0263574722001813 Published online by Cambridge University Press

4. Conclusion

This paper proposes a novel emotional gait conversion model with adversarial loss and cycle consistency loss as a data augmentation method to overcome the insufficient data problem for gait emotion recognition. Meanwhile, this is the first work to realize the mutual transformation between natural gait and emotional gait. By the emotional gait conversion model, we generated numerous synthetic gait samples that enhance the diversity of the original datasets. Experimental results show that our emotion classifiers are competitive with state-of-the-art gait emotion recognition systems by the augmented dataset. It is expected that the integration of emotion recognition as an aspect of nonverbal communication enhances HRI. We only identify three emotional states through gait information, while human emotions are extremely diverse. We will gather gait data from more emotions in the future to investigate the fine-grained space of gait-based emotions. Moreover, different modalities can complement each other to represent more discriminative features. We will try to incorporate appearance information to promote the performance of gait-based recognition.

Author contributions. Weijie Sheng and Xinde Li conceived and designed the study. Weijie Sheng and Xiaoyan Lu conducted data gathering. Weijie Sheng performed statistical analyses. Weijie Sheng and Xiaoyan Lu wrote the article.

Financial support. This work was supported in part by the National Natural Science Foundation of China under Grant 62233003 and 62073072, and in part by the Key Projects of Key R&D Program of Jiangsu Province under Grant BE2020006 and Grant BE2020006-1 and in part by Shenzhen Natural Science Foundation under Grant JCYJ20210324132202005 and JCYJ20220818101206014.

Conflicts of interest. The authors declare no conflicts of interest exist.

Ethical approval. Not applicable.

References

[1] L. Teijeiro-Mosquera, J.-I. Biel, J. L. Alba-Castro and D. Gatica-Perez, "What your face vlogs about: expressions of emotion and big-five traits impressions in youtube," *IEEE Trans. Affect. Comput.* **6**(2), 193–205 (2015).

[2] M. Korayem, S. Azargoshasb, A. Korayem and S. Tabibian, "Design and implementation of the voice command recognition and the sound source localization system for human–robot interaction," *Robotica* **39**(10), 1779–1790 (2021).

[3] N. Liu, T. Zhou, Y. Ji, Z. Zhao and L. Wan, "Synthesizing talking faces from text and audio: an autoencoder and sequence-to-sequence convolutional neural network," *Pattern Recognit.* **102**, 107231 (2020).

[4] S.-S. Yun, "A gaze control of socially interactive robots in multiple-person interaction," *Robotica* **35**(11), 2122–2138 (2017).

[5] X. Liu, K. N. Khan, Q. Farooq, Y. Hao and M. S. Arshad, "Obstacle avoidance through gesture recognition: Business advancement potential in robot navigation socio-technology," *Robotica* **37**(10), 1663–1676 (2019).

[6] P. Xue, B. Li, N. Wang and T. Zhu, "Emotion Recognition From Human Gait Features Based on DCT Transform," **In:** 5th International Conference on Human Centered Computing (HCC), vol. 11956 (2019) pp. 511–517.

[7] F. Göngör and Ö. Tutsoy, "Design and implementation of a facial character analysis algorithm for humanoid robots," *Robotica* **37**(11), 1850–1866 (2019).

[8] R. Jain, V. B. Semwal and P. Kaushik, "Stride segmentation of inertial sensor data using statistical methods for different walking activities," *Robotica*, 1–14 (2021).

[9] J. E. Cutting and L. T. Kozlowski, "Recognizing friends by their walk: Gait perception without familiarity cues," *Bull. Psychon. Soc.* **9**(5), 353–356 (1977).

[10] W. Sheng and X. Li, "Multi-task learning for gait-based identity recognition and emotion recognition using attention enhanced temporal graph convolutional network," *Pattern Recognit.* **114**(1), 107868 (2021).

[11] Z. Li, Z. Ren, K. Zhao, C. Deng and Y. Feng, "Human-cooperative control design of a walking exoskeleton for body weight support," *IEEE Trans. Ind. Inform.* **16**(5), 2985–2996 (2019).

[12] Z. Li, C. Xu, Q. Wei, C. Shi and C.-Y. Su, "Human-inspired control of dual-arm exoskeleton robots with force and impedance adaptation," *IEEE Trans. Syst. Man Cybernet. Syst*. **50**(12), 5296–5305 (2018).

[13] V. Narayanan, B. M. Manoghar, V. S. Dorbala, D. Manocha and A. Bera, "Proxemo: Gait-Based Emotion Learning and Multi-View Proxemic Fusion for Socially-Aware Robot Navigation," **In:** *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (IEEE, 2020) pp. 8200–8207.

[14] S. Xu, J. Fang, X. Hu, E. Ngai, Y. Guo, V. Leung, J. Cheng and B. Hu, "Emotion recognition from gait analyses: Current research and future directions, arXiv preprint arXiv:2003.11461 (2020).

https://doi.org/10.1017/S0263574722001813 Published online by Cambridge University Press

[15] U. Bhattacharya, N. Rewkowski, P. Guhan, N. L. Williams, T. Mittal, A. Bera and D. Manocha, "Generating Emotive Gaits for Virtual Agents Using Affect-Based Autoregression," **In:** *IEEE International Symposium on Mixed and Augmented Reality (ISMAR),* (IEEE, 2020b) pp. 24–35.

[16] G. Li, Z. Li and Z. Kan, "Assimilation control of a robotic exoskeleton for physical human-robot interaction," *IEEE Robot. Automat. Lett.* **7**(2), 2977–2984 (2022).

[17] R. Peri, S. Parthasarathy, C. Bradshaw and S. Sundaram, "Disentanglement for Audio-Visual Emotion Recognition Using Multitask Setup," **In:** *ICASSP, 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2021) pp. 6344–6348.

[18] J. Liang, Z. Liu, J. Zhou, X. Jiang, C. Zhang and F. Wang, "Model-protected multi-task learning," **In:** *IEEE Trans. Pattern Anal. Mach. Intell.* **44**(2), 1002–1019 (2020).

[19] B. Zhang, E. M. Provost and G. Essl, "Cross-corpus acoustic emotion recognition with multi-task learning: seeking common ground while preserving differences," *IEEE Trans. Affect. Comput.* **10**(1), 85–99 (2019).

[20] X. Yu, C. Xu, X. Zhang and L. Ou, "Real-time multitask multihuman–robot interaction based on context awareness," *Robotica* **40**(9), 1–27 (2022).

[21] W. Sheng and X. Li, "Siamese denoising autoencoders for joints trajectories reconstruction and robust gait recognition," *Neurocomputing* **395**, 86–94 (2020).

[22] L. Yi and M.-W. Mak, "Improving speech emotion recognition with adversarial data augmentation network," *IEEE Trans. Neur. Netw. Learn.* **33**(1), 172–184 (2020).

[23] C.-L. Huang, "Exploring Effective Data Augmentation with Tdnn-Lstm Neural Network Embedding for Speaker Recognition," **In:** *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (2019) pp. 291–295.

[24] U. Bhattacharya, T. Mittal, R. Chandra, T. Randhavane, A. Bera and D. Manocha, "Step: Spatial Temporal Graph Convolutional Networks for Emotion Perception From Gaits," **In:** *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34a (2020a) pp. 1342–1350.

[25] M. Mirza and S. Osindero. *Conditional Generative Adversarial Nets,* arXiv: Learning (2014).

[26] K. Sohn, H. Lee and X. Yan, "Learning Structured Output Representation Using Deep Conditional Generative Models," **In:** *NIPS 2015* (2015) pp. 3483–3491.

[27] J. Gao, D. Chakraborty, H. Tembine and O. Olaleye, "Nonparallel Emotional Speech Conversion," **In:** *Interspeech* (2019).

[28] P. Isola, J.-Y. Zhu, T. Zhou and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," **In:** *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017) pp. 1125–1134.

[29] J.-Y. Zhu, T. Park, P. Isola and A. A. Efros, "Unpaired Image-to-image Translation Using Cycle-Consistent Adversarial Networks," **In:** *IEEE International Conference on Computer Vision (ICCV)* (2017) pp. 2242–2251.

[30] T. Kim, M. Cha, H. Kim, J. K. Lee and J. Kim, "Learning to Discover Cross-Domain Relations with Generative Adversarial Networks," **In:** *International Conference on Machine Learning* (PMLR, 2017), pp. 1857–1865.

[31] X. Huang, M.-Y. Liu, S. Belongie and J. Kautz, "Multimodal Unsupervised Image-to-image Translation," **In:** *Proceedings of the European Conference on Computer Vision (ECCV)* (2018) pp. 172–189.

[32] Y. Choi, Y. Uh, J. Yoo and J.-W. Ha, "Stargan v2: Diverse Image Synthesis for Multiple Domains," **In:** *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020) pp. 8188–8197.

[33] G. Rizos, A. Baird, M. Elliott and B. Schuller, "Stargan for Emotional Speech Conversion: Validated by Data Augmentation of End-to-end Emotion Recognition," **In:** *ICASSP, 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2020) pp. 3502–3506.

[34] B.-H. Su and C.-C. Lee, "A Conditional Cycle Emotion Gan for Cross Corpus Speech Emotion Recognition," **In:** *IEEE Spoken Language Technology Workshop (SLT)* (2021) pp. 351–357.

[35] Q. Zhu, L. Gao, H. Song and Q. Mao, "Learning to disentangle emotion factors for facial expression recognition in the wild," *Int. J. Intell. Syst.* **36**(6), 2511–2527 (2021).

[36] F. Schroff, D. Kalenichenko and J. Philbin, "Facenet: A Unified Embedding for Face Recognition and Clustering," **In:** *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015) pp. 815–823.

[37] D. Kastaniotis, I. Theodorakopoulos, C. Theoharatos, G. Economou and S. Fotopoulos, "A framework for gait-based recognition using kinect," *Pattern Recogn. Lett.* **68**, 327–335 (2015).

[38] D. Kastaniotis, I. Theodorakopoulos, G. Economou and S. Fotopoulos, "Gait based recognition via fusing information from euclidean and riemannian manifolds," *Pattern Recogn. Lett.* **84**, 245–251 (2016).

[39] J. Bao, D. Chen, F. Wen, H. Li and G. Hua, "CVAE-GAN: Fine-grained Image Generation Through Asymmetric Training," **In:** *IEEE International Conference on Computer Vision (ICCV)* (2017) pp. 2764–2773.

Cite this article: W. Sheng, X. Lu and X. Li (2023). "Data augmentation by separating identity and emotion representations for emotional gait recognition", Robotica **41**, 1452–1465. https://doi.org/10.1017/S0263574722001813