Journal on Interactive Systems, 2026, 17:1
ISSN: 2763-7719 • doi: 10.5753/jis.2026.6799

Research Paper

What are Hard Samples for Gait Emotion Recognition?

Thifany Ketuli Silva de Souza [Federal University of Pernambuco | *tkss@cin.ufpe.br*]
Willams de Lima Costa [Federal University of Pernambuco | *wlc2@cin.ufpe.br*]
Veronica Teichrieb [Federal University of Pernambuco | *vt@cin.ufpe.br*]

Voxar Labs, Centro de Informática, Universidade Federal de Pernambuco, Av. Jorn. Aníbal Fernandes, Cidade Universitária, Recife, PE, 50670-901, Brazil.

Abstract. Emotion recognition is a relevant capability for understanding human behavior, enabling interactive systems to support natural interaction with their users. Among the diverse ways the literature tackles this task, gait emotion recognition (extracting the perceived emotion from how someone walks) remains challenging due to limited data availability and inconsistent annotations that often overlap, leading to ambiguous scenarios and harming training and validation processes. Given this challenge, we pose the following question: What are hard samples for gait emotion recognition? To answer this question, we investigate sample complexity in gait-based emotion recognition and its impact on learning dynamics, focusing on ambiguous and hard-to-learn samples using dataset cartography techniques that map samples based on model correctness and confidence. Our findings suggest that easier samples are those that highlight clear emotional cues, as expected, whereas ambiguous and difficult samples include restrained movements, static frames, and annotation inconsistencies. These insights can inform dataset curation and model training strategies for more reliable gait emotion recognition.

Keywords: Gait Analysis, Emotion Recognition, Dataset Cartography, Human Behavior Recognition, Biomechanics

Edited by: Rafael Rieder | **Received:** 03 November 2025 • **Accepted:** 15 April 2026 • **Published:** 29 April 2026

1 Introduction

Developing interfaces that enable people to communicate with technology in ways that feel more like everyday life is a concept of natural interaction between humans and machines. This approach aims to make these interactions more fluid so that users do not perceive them as dealing with a machine [Valli, 2007]. User-centered techniques to enable this type of interaction usually rely on human behavior understanding through gestures, expressions, posture, and other nonverbal communication cues. Complementing this perspective, research in behavioral psychology has investigated how gait-related parameters could be used to describe social aspects, including emotion, through features such as stride length and arm swing [Montepare *et al*., 1987]. In this context, these findings highlight that gait analysis can be a valuable source for assessing nonverbal body language and information about human behavior, while also providing insights into someone's emotional state.

Despite progress in recognizing human behavior patterns, modeling these processes for computational methods is still challenging due to their complexity and variability. The quality of data representation remains a challenge, notwithstanding improvements in deep learning techniques, such as spatio-temporal convolutional networks (e.g., STEP and ST-Gait++ [Bhattacharya *et al*., 2024]), which have improved emotion classification based on gait. Advancements in this area face critical limitations in data resources, including a scarcity of high-quality and diverse public databases, inaccurate annotations of emotional states, and biases derived from collection in artificial environments and uneven demographics. Consequently, current computational models often exhibit limited generalization capacity and difficulty addressing the natural variability of human behavior.

On the other hand, the challenge of differentiating emotional cues persists, as we cannot directly measure these states and must instead infer them from nonverbal indicators. The identification of these emotions is often hampered by ambiguities and overlaps between categories, leading to confusion between certain emotions. In the context of gait analysis, it is common for emotions of similar intensity, such as joy and anger, or sadness and fear, to share kinematic and intensity parameters, leading to classification ambiguities [Roether *et al*., 2009; Halovic and Kroos, 2018; Reynolds *et al*., 2019]. Furthermore, factors such as the level of physiological arousal displayed and individual variability in recognition increase the complexity of the process, limiting the accuracy of emotional identification [Reynolds *et al*., 2019; Lopez *et al*., 2017]. In this context, some strategies have been increasingly explored in affective computing, such as multimodal representation approaches that aim to integrate complementary cues to provide richer contextual information and reduce classification errors by combining gait with additional signals, thereby broadening the representation space and improving learning when movement patterns alone are ambiguous. Nevertheless, understanding the determinants of difficulty within gait data that hinder emotion differentiation remains essential to overcoming the limitations imposed on gait studies by affective computing, whether unimodal or multimodal.

Therefore, the objective of our work is to investigate potential determinants of sample difficulty by analyzing training dynamics (class confidence/variability) using a well-accepted baseline model [Yan *et al*., 2018], which also serves as a basis for other works in the literature [Bhattacharya *et al*., 2024; Zhai *et al*., 2024; Li *et al*., 2025; Lima *et al*., 2024; Zhou *et al*., 2025]. We seek to recognize patterns across different emotions that may reveal their underlying similarities and

to explore how these instances behave during deep learning training. Our work aims to analyse, using a dataset cartography approach to visually separate our data into regions of difficulty, if there are characteristics of gait sequences associated with individual sample hardness in our proposed training dynamic, and collect some examples to investigate if there is an association between biomechanical characteristics of the data with inter-class overlap, spatio-temporal cues, dataset composition changes, etc. The contributions of this work are:

- Assessing the interactions between intra-class and extra-class instances, what influences the training dynamics, and whether intrinsic characteristics account for them in this context.
- Examining the relationship between easy, hard, and ambiguous gait samples, perceiving the patterns observed, and explaining the similarities and differences among them.
- Discussing if there are specific characteristics that contribute to the stagnation of SOTA results, analyzing the inherent challenges involved in emotion recognition.

We organized this study as follows: we present a literature review of gait emotion recognition and datasets in the Section 2; Section 3 describes the methods adopted for exploring gait analysis; Section 4 describes our experimental setup; Section 5 presents the results obtained and discuss the main points raised, limitations and suggestions for future works; finally, Section 6 presents the conclusions of the study, summarizing the findings and pointing to possible future directions.

2 Literature Review

In this Section, we present the main concepts, foundations, and related works supporting this study to contextualize the analysis performed.

2.1 Datasets for gait emotion recognition

The CMU Motion Capture (MoCap)[1] repository is a large-scale collection of human motion recordings produced by the Carnegie Mellon University Graphics Lab as a dataset for computer graphics, animation, biomechanics, and affective computing research. Over 140 subjects performed 2,605 motion samples collected with approximately 41 reflective markers to capture 3D joint trajectories. The samples include walking sequences performed under different conditions, including neutral, happy, sad, and angry walking. Although widely used, limitations include the lack of naturalistic environments and occasional missing marker data.

Bhattacharya *et al*. [2024] introduced the Emotion-Gait dataset as a benchmark for human gait recognition using ST-GCN models. A total of 2,177 natural gait points were reunited, of which the authors collected 342 samples, while the remaining 1,835 came from the Edinburgh Locomotion MOCAP Database (ELMD) [Habibie *et al*., 2017]. Domain experts labeled both parts of the dataset. Table 1 shows the samples distribution according to the four emotion labels present on Emotion-gait Dataset: Angry, Neutral, Happy and Sad.

[1]Available on `https://mocap.cs.cmu.edu/info.php`. Accessed on 26 April 2026

Table 1. Emotion-gait class distribution according to its four emotion labels: Angry, Neutral, Happy and Sad.

Class				
Angry	**Neutral**	**Happy**	**Sad**	**Total**
1160	487	332	198	2177

The OU-ISIR Gait Database Comprising the Treadmill Dataset [Makihara *et al*., 2012] is one of the largest and most comprehensive repositories of human gait data, developed by the Institute of Scientific and Industrial Research (ISIR) at Osaka University. It contains 4,007 recorded subjects, making it the largest treadmill-based gait dataset available. The database provides gait sequences across multiple walking speeds, ranging from slow to fast, with each subject recorded under at least two speed conditions.

Despite their relevance, these datasets share some common limitations. They are often recorded in controlled or artificial environments (laboratories or treadmills), which reduces ecological validity. In addition, there can be missing or noisy data, subjective annotation processes, and a lack of representation of the full diversity of emotional and situational contexts encountered daily. These limitations may affect data quality and the replicability of results in real-world conditions. In this context, our study aims to investigate whether common patterns can compromise generalization and robustness in practical applications, examine whether intrinsic characteristics in gait data point to these limitations, and determine how they affect the recognition process.

2.2 Methods for gait emotion recognition

Earlier works on gait-based emotion recognition focused on two main approaches: appearance-based and biomechanical parameter-based. Appearance-based models focused on segmentation to align silhouettes and extract features for gait analysis. On the other hand, biomechanical models aimed to model the body skeleton based on proportions, joint trajectories, etc. [Iwashita *et al*., 2013; Liao *et al*., 2020]. Although these techniques were innovative in gait modeling, they showed severe limitations regarding their ability to correlate local and global features. Moving forward, the state-of-the-art began to show interest in graph-based approaches for their ease of mapping local and global features. These techniques often relied on Graph Convolutional Networks (GCNs) [Kipf and Welling, 2017]. Yan *et al*. [2018] introduced Spatial-Temporal Graph Convolutional Networks (ST-GCN) as a deep learning architecture for dynamic recognition of human body skeletons. This method incorporates graph convolution based on the spatial relations among body joints and their temporal evolution.

Among these works, Yin *et al*. [2022] introduced the MSA-GCN network, formed by selective adaptive spatiotemporal convolution and a cross-scale mapping fusion mechanism. Even though it improved the baseline, it still showed significant ambiguity between the Happy/Sad classes and the Neutral/Sad classes.

Given these limitations, Bhattacharya *et al*. [2024] proposed STEP, a model based on ST-GCNs, along with a new benchmark, Emotion-Gait, as we previously discussed. This method surpassed previous works and established a new base-

line; however, some misclassification issues remain between emotions with similar gait patterns, such as Sad/Neutral and Angry/Neutral. These confusions may stem from slow movements, low arm swings, tense body postures, or fast-paced steps. Other contributing factors include the subjectivity of labeling, individual differences in emotional expression, and natural similarities between emotions.

Lima *et al*. [2024] extends this approach by proposing an improved version of the ST-GCN model, which adds adaptive learning mechanisms. This approach improved accuracy while reducing convergence time, yielding performance gains across all classes; however, it underperformed on the Sad class.

Recent works have moved away from the ST-GCN baseline, focusing on new approaches such as BPM-GCN Zhai *et al*. [2024], which introduces a two-stream approach for posture and movement with adaptive fusion; MDT-GCN Li *et al*. [2025], which introduces multi-adaptive fusion and bifocal attention for improved spatiotemporal representations; and Zhou *et al*. [2025], combining GCNs and Transformers in parallel flows. Although there is an apparent increase in accuracy, these latter results still show confusion between classes.

The state-of-the-art shows consistent high overall accuracy, but notable performance disparities across certain classes. These results highlight the need to explore the boundaries between these emotions, seeking to understand the reasons behind the difficulty in separating them and whether there are features that discretize them into rigid categories.

3 Methodology

To explore why current state-of-the-art models exhibit these performance disparities, we propose the hypothesis that differences in how the model learns each sample may affect how individual instances influence the training dynamics and, consequently, the difficulty the model encounters when handling each sample. For this, we propose an experiment to extract learning representations during the training procedure.

For this experiment, we use an ST-GCN-based architecture [Yan *et al*., 2018] as baseline due to its structural alignment with skeleton-based modeling that offers good interpretability at the joint-temporal level, and its widespread adoption in literature for gait emotion recognition [Bhattacharya *et al*., 2024; Lima *et al*., 2024].

Consider an undirected spatial-temporal graph $G = (V, E)$ on a skeleton sequence with N joints and T frames. Each node $v_{ti} \in V$ represents the $i-th$ joint at time step t. The set of nodes $V = \{v_{ti} | t = 1, \ldots, T, i = 1, \ldots, N\}$ represents all joints covered in the skeleton sequence. The edge E is composed of two subsets: intra-skeleton connection at each frame, denoted as $E_s = \{v_{ti}v_{tj} | (i, j) \in H\}$, where H is the set of natural human body connections (e.g., elbow-wrist, knee-hip); inter-frame edges, denoted as $E_F = \{v_{ti}v_{(t+1)i}\}$, where all edges connect the same joint across consecutive frames and describe its trajectory over time. Following their implementation, we apply an architecture with six blocks combining spatial and temporal convolution, with channel dimensions increasing (64, 128, 256), followed by adaptive pooling and a fully connected layer.

3.1 Dataset Cartography

Understanding data behavior throughout the training process helps us understand how a machine learning model learns and the challenges it faces in addressing a problem. In this context, Swayamdipta *et al*. [2020] proposed a data-mapping technique that emphasizes each sample's behavior in the training dynamics. This method focuses on the evolution of the model predictions over the learning process.

Given a train set $D = \{(x_i, y_i^*)\}_{i=1}^N$, where x_i corresponds to the input and y_i^* the actual label, we consider a parameterized model θ, which is iterated through E epochs. When combined in a bidirectional plan, the model's performance on individual instances during training can provide information about each difficulty in the learning process. The vertical axis represents confidence levels, based on the probability distributions, while the horizontal axis indicates variability. Additionally, the sample distribution illustrates visual areas that can categorize samples as easy, complex, or ambiguous to learn.

Confidence is the mean probability the model assigns to the correct label. In other words, examples always classified as the actual label have a high correctness probability and, consequently, high confidence. On the other hand, low confidence means wrong classification over time. The confidence is formulated as follows:

$$\hat{\mu_i} = \frac{1}{E} \sum_{e=1}^{E} p_{\theta^{(e)}}(y_i^* | x_i) \tag{1}$$

where $p_{\theta^{(e)}}$ corresponds to the probability distribution at the end of the epoch e.

The standard deviation of confidence across epochs corresponds to the variability. This metric is essential for evaluating sample stability over time. Low variability means the model is stable and maintains its prediction for that instance. On the other hand, high variability indicates that the model oscillates between high and low probabilities for the actual label, suggesting instability. High sample variability indicates ambiguity, suggesting the model may be unstable during training. The variability is formulated as

$$\hat{\sigma}_i = \sqrt{\frac{1}{E} \sum_{e=1}^{E} \left(p_{\theta^{(e)}}(y_i^* | x_i) - \hat{\mu}_i\right)^2} \tag{2}$$

Instances with high confidence and low variability are found in easy-to-learn regions. These examples show that the model has a high probability of making correct predictions and rarely encounters difficulties during training. These samples suggest that the model quickly learns characteristics that are typically representative and well-labeled. In contrast, instances with low confidence and variability are more challenging to learn and are found in hard-to-learn regions. Characteristics cause these difficulties that the mode struggles to identify. Low variability indicates a consistent model pattern failing to classify these instances accurately in this context. It might suggest that there are misannotated labels or confusing examples; these instances do not aid in the learning process.

High-variability instances are found in the ambiguous area because the model often struggles to classify these sam-

ples, leading to no convergence. In these cases, the instances can have multiple interpretations or be weakly contextualized. As these samples force the model to look for non-trivial patterns, the authors argue that they are necessary for more flexible models. Additionally, inconsistencies in annotation instructions or the existence of complex patterns can increase the number of samples in the ambiguity region. Despite that, these samples remain valuable when used strategically in various training techniques, such as data augmentation or curriculum learning.

Although the selected cartography approach was initially proposed for NLP tasks [Swayamdipta *et al*., 2020], dataset cartography has not yet been systematically explored for skeleton-based problems in affective computing. Given its potential for exploration, we extend its application to gait emotion recognition problems, leveraging its capacity to reveal instance-level learning difficulty, and propose a cross-domain approach that enables diagnosis of challenges that may not be visible in aggregate accuracy metrics.

3.2 Motion analysis

To analyze kinematic properties of the biomechanics of the movement, we characterized each body joint by the corresponding position vector $p(t)$ with temporal coordinates $(x(t), y(t), z(t))$. The velocity of the hand is the derivative of the temporal change in the position of the corresponding joint over time, given by

$$\vec{v}_{hand}(t) = \frac{d\vec{r}(t)}{dt}$$

and the scalar velocity is the magnitude of this vector, given by

$$|\vec{v}(t)| = \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2}$$

The hand movement is a direct consequence of the movement of the arm-forearm-hand chain, as it is the distal segment of this chain. According to Winter [2009], distal segments with high linear velocity are associated with an increase in angular proximal velocity. Thus, we measure arm movement as hand linear velocity, which also reflects arm kinematic energy and, consequently, the level of energy employed in hand movement during gait.

We measure walking speed by tracking the body's horizontal displacement over time. The movement of the root joint is considered to estimate this displacement. Mathematically

$$v(t) = \frac{d}{dt}\sqrt{\left(x_{\text{root}}(t)\right)^2 + \left(z_{\text{root}}(t)\right)^2}$$

Average speed is the ratio between the total displacement traveled in the horizontal plane and the total time:

$$v_{\text{avg}} = \frac{\sqrt{(x_T - x_0)^2 + (z_T - z_0)^2}}{\Delta t}$$

We employed the Euclidean distance between feet projected onto the horizontal plane as the distance between the contact on the left foot and the subsequent contact of the right foot (or vice versa) to measure the step length by the linear displacement in the plane of progression. Formally, this can be expressed as

$$Step_{length} = ||(x_{\text{right foot}}, y_{\text{right foot}}) - (x_{\text{left foot}}, y_{\text{left foot}})||$$

In biomechanics, two consecutive step lengths form a stride, the sum of the distances covered in each right and left step. The distance between the initial and subsequent contact of the same feet is known as the stride length [Winter, 2009], where:

$$Stride = Length_{right} + Length_{left}$$

We consider a symmetrical gait when right and left steps have the same length, in other words, when $Stride = 2{*}Step$. This analysis allows us to infer information about compensations in their mobility that may indicate differences between distinct emotions.

4 Experiments

This Section describes the setup for implementing the technique. As discussed by Zhou *et al*. [2020], deep learning models tend to learn instances with low hardness early due to stability during training. In contrast, instances with high hardness and, consequently, high instability tend to take longer to be learned. Given our main objective in analyzing how each sample behaves during the learning process, we introduced an ST-GCN-based model [Yan *et al*., 2018] for 300 epochs with unit batches and instance-loss updates on the Emotion-Gait Dataset [Bhattacharya *et al*., 2024] in three configurations: only samples from the ELMD; only the complementary Emotion-Gait samples; and the full Emotion-Gait dataset (ELMD and complementary samples). We opted for an extended training schedule (300 epochs) and a unit batch size to ensure stabilization of confidence and variability measures over time, which is important for dataset cartography, given that premature convergence could mask instability patterns, and to track per-sample behavior throughout training and preserve instance-level updates, respectively. Our experimental design prioritizes the analysis of training dynamics, shifting the evaluation perspective from aggregate performance to dataset diagnostic behavior.

We set a schedule for the learning rate based on cosine cycles, where training begins with relatively high rates that favor broader exploration of the parameter space, then gradually reduces to lower values to stabilize convergence. We employed data-mapping cartography [Swayamdipta *et al*., 2020] to analyse the behavior of instances in our model, indentifying the visual distribution of samples in a confidence versus variability map.

We ran all experiments on an NVIDIA GeForce RTX 4090 GPU using PyTorch. We used Adam as the optimizer with a weight decay of 0.001 and an initial learning rate of 0.0001.

5 Results and Discussion

In this Section, we present and discuss the study's main findings.

5.1 Dataset Cartography Map analysis

We apply the approach proposed by Swayamdipta *et al*. [2020] for a data map cartography. Following the original approach, we do not impose fixed thresholds on confidence or variability, given the alignment between the dataset and the architecture, which varies in each case depending on how it is applied.

Concretely, we considered easy-to-learn instances as those in the upper band of confidence and the lower band of variability, hard-to-learn instances as those in the lower band of both confidence and variability, and ambiguous instances as those in the upper band of variability, regardless of confidence. These considerations led us to use these band-based thresholds solely as a pragmatic operationalization, which does not imply absolute semantic boundaries between regions.

ELMD. As we show in Figure 1, we can see a clear separation between the emotional classes on the cartography map. The easy-to-learn region mainly concentrates on Angry class samples, with high confidence, which means high correctness over time, and low variability, meaning steady prediction for those instances. In contrast, the hard-to-learn area primarily locates the Sad class samples with low confidence, indicating wrong classifications over time, and reduced variability, which means steady predictions. We attributed the behavior of these two classes to class imbalance, where there are more Angry than Sad samples, which affects the model's ability to extract consistent patterns. In the ambiguity region of the map, we observe the Happy and Neutral classes. The Happy class displays its samples more heterogeneously, showing cases with different levels of confidence and variability. Meanwhile, the Neutral class concentrates most of its instances around average values. This ambiguous behavior from Neutral arises from its overlap with other emotional categories. Although the Happy class encounters this overlap, this case manifests more diffusely.

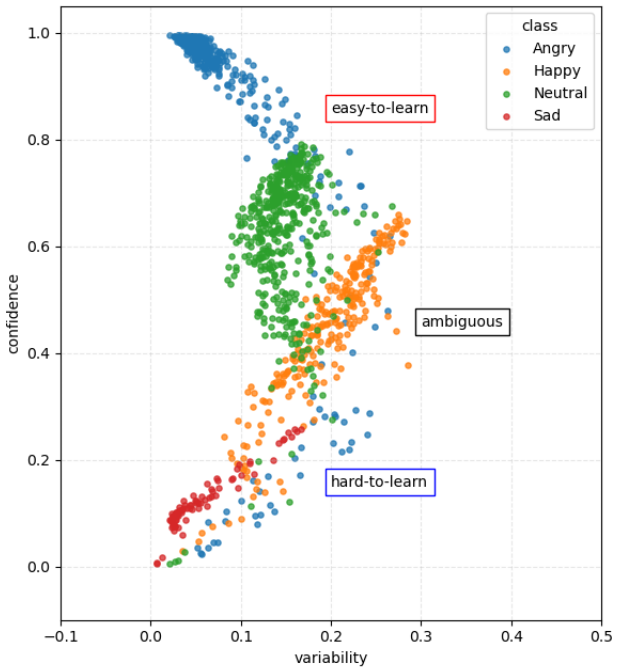

Figure 1. Cartographic map of ELMD data with distribution of confidence and variability values

It is essential to highlight the presence of inconsistent instances within classes that generally perform well. For example, in the case of some Angry samples, there are instances distributed across regions that are difficult to classify and ambiguous, reinforcing the hypothesis that not all manifestations of emotion follow a homogeneous pattern. Under certain conditions, the boundaries between classes can become less defined.

The histograms in Figure 2 suggest reliable classifications in a high concentration of samples with high confidence values in the confidence histogram. Moreover, it shows that most instances have variability values below 0.2. These results show that almost all samples have stability throughout the classification process. When analyzed alongside the data map, this result still suggests that a small proportion of instances are challenging (low confidence and high variability), and almost all of them are from the Sad class, further reinforcing this class as challenging. However, it is also important to note that specific subsets with significant instability contribute to the ambiguities observed in the results.

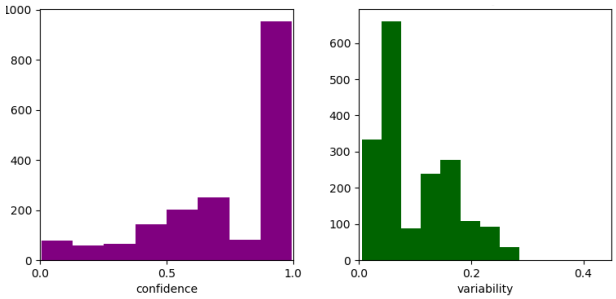

Figure 2. Histogram distributions of confidence and variability for ELMD dataset

Only Complementary data Contrary to the ELMD map, Figure 3 shows that the structural organization of this dataset displays greater homogeneity. The balanced representation among classes appears to mitigate dominance effects, as no single category occupies an exclusive area of the map. There is a high concentration of points in easy-to-learn and ambiguous regions, which suggests that the majority of instances have high correctness probability (high confidence) and variable inconsistencies in classification (variability).

As illustrated in Figure 4, the confidence and variability histograms clarify the tendency observed in the corresponding cartography map. The confidence histogram shows a high concentration of samples in ambiguous and easy-to-learn regions, with the majority having values above 0.7 and a variability of around 0.25.

Full Emotion-Gait. The scenario is entirely different when we compare the behavior of instances when the train joins both subsets (the full Emotion-gait). The cartographic map in Figure 5 shows an intense concentration of samples from different classes in the easy-to-learn area (high confidence and low variability), which suggests that the complementary data probably reduced the disparities observed when training only with ELMD data. The confidence and variability histograms in Figure 6 reinforce this pattern, because the new configurations bring a massive accumulation of instances with confidence close to 1.0, and a predominance of variability values below 0.2. Despite this predominance of easy instances,

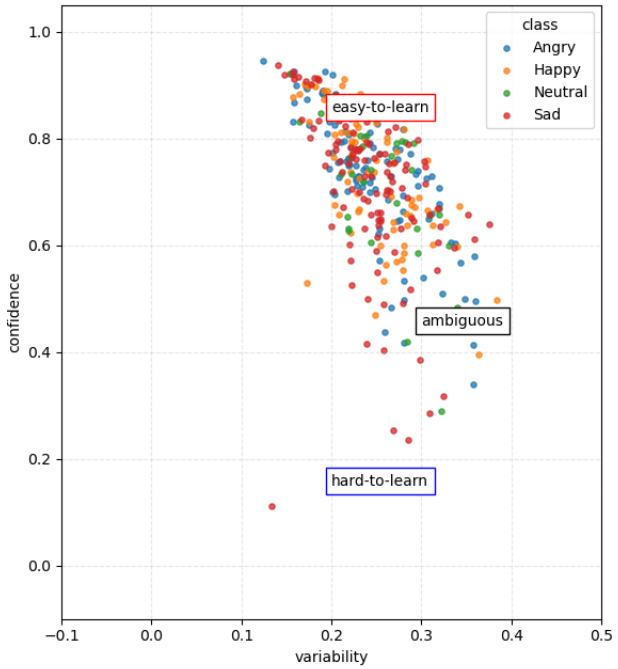

Figure 3. Cartographic map of complementary Emotion-Gait data with distribution of confidence and variability values

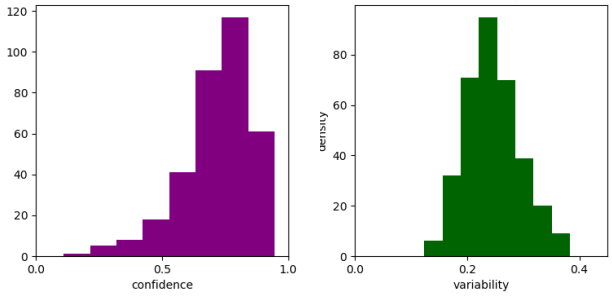

Figure 4. Histogram distributions of confidence and variability for complementary Emotion-Gait data

some data are still located in ambiguous regions, suggesting that these instances still have unstable responses, alternating between successes and failures, causing feature overlap between classes.

Additionally, we observe a high diversity with all classes in practically all map regions, lacking the clear separation observed in the ELMD set. This dilution may be related to the simple addition of data. Combining both datasets made the boundaries of previously delimited regions of easy, ambiguous, and hard learning softer.

In this context, the complete Emotion-gait set brought greater intra-class diversity and inter-class overlap, making it more difficult for the model to maintain well-defined groupings. In other words, the expansion of the dataset increased the variability of the samples and increased the space for confusion between emotions.

5.2 Sub-analysis of regions with distinct patterns

Following this evaluation, we can now focus on the core research question of this work: *what are examples of hard samples to learning for gait emotion recognition?* For this, we randomly selected five instances from each region of the map, as presented in Table 2, for a deeper exploratory analysis of their biomechanical behavior and its tendencies. We

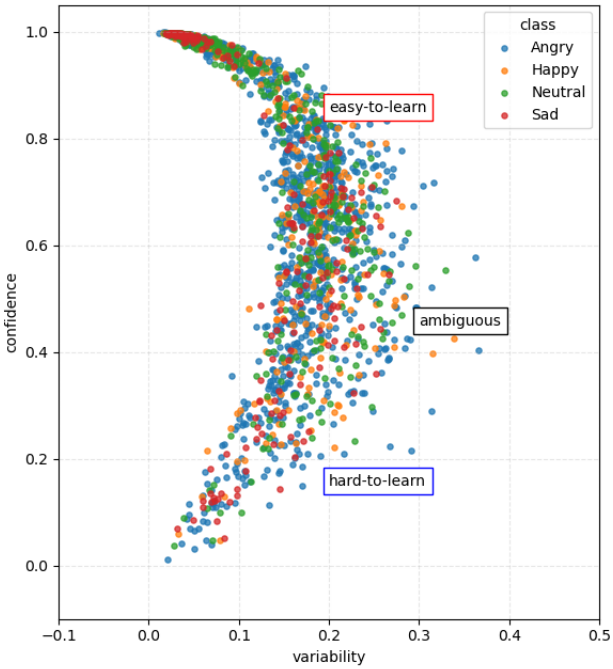

Figure 5. Cartographic map of full Emotion-Gait data with distribution of confidence and variability values

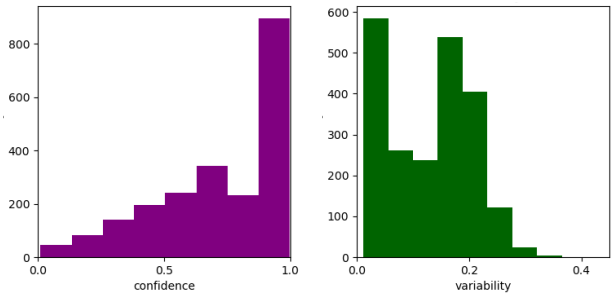

Figure 6. Histogram distributions of confidence and variability for the full Emotion-Gait dataset

evaluated kinematic parameters, which revealed significant differences between the hard, ambiguous, and easy sets. It is important to emphasize that the patterns described below suggest indicative trends, given the cartography map as a visual illustration rather than an intrinsic statistical value that validates differences across regions. We present the global summary of the Emotion-Gait Dataset in the Table 3, for general comparison with our main result.

Observing the plot of these gait samples in Figure 7, we can also notice recurring patterns across classes and regions of learning.

Easy-to-learn. The datapoints metrics show a distinct pattern compared to the other regions: the coexistence of low gait velocity with proportionally long steps and strides. Angry instances have the highest gait velocity combined with energetic arm moves; additionally, there is high asymmetry between left and right stride. Perhaps high arm movements for Happy, this class differentiates for gait velocity lower than Angry and symmetric left and right stride. Finally, the Neutral instance has lower gait velocity, symmetric strides, and symmetric hand movements. This selected region suggests a gait characterized by slow progression but with greater energy expended in arm movements. Observing the five selected instances of the easy-to-learn subset in Figure 7a, it is possible

Table 2. Quantitative analysis of biomechanical metrics evaluated by a subset of data selected by the region of interest

Area	Class	Gait (unit/s)	Step (unit)	L-stride (unit)	R-stride (unit)	L-hand (unit/s)	R-hand (unit/s)
Hard-to-learn	Neutral	0.29	0.26	0.05	0.11	0.76	0.62
	Sad	0.35	0.37	0.48	0.42	0.69	0.74
	Angry	0.58	0.38	0.31	0.27	0.89	0.81
	Neutral	0.41	0.24	0.26	0.35	0.59	0.58
	Sad	0.38	0.31	0.41	0.67	0.48	0.6
Ambiguous	Neutral	0.41	0.34	0.51	0.47	0.7	0.7
	Happy	0.57	0.61	1.42	1.18	0.92	1.05
	Happy	0.28	0.66	1.19	1.23	0.99	1.64
	Neutral	0.54	0.36	0.34	0.28	0.75	0.82
	Happy	0.23	0.55	0.66	0.87	0.86	1.19
Easy-to-learn	Angry	0.16	0.54	1.05	0.58	1.17	1.04
	Happy	0.14	0.49	1.33	1.33	1.53	1.36
	Happy	0.05	0.44	0.98	0.72	0.88	1.03
	Neutral	0.07	0.45	0.82	0.73	1.01	0.89
	Angry	0.31	0.5	0.58	0.97	1.38	1.21

to see apparent differences between the sequences. In terms of expressiveness, angry poses appear wider than happy ones, characterized by greater leg swing and larger arm amplitude. Neutral samples seem less energetic, which may be associated with a gait lacking emotional content.

Table 3. Global Summary of the Emotion-Gait Dataset

	Mean ± Standard Deviation
Gait (unit/s)	0.22 ± 0.16
Step (unit)	0.45 ± 0.20
L-stride (unit)	0.64 ± 0.49
R-stride (unit)	0.64 ± 0.48
L-hand (unit/s)	1.01 ± 0.47
R-hand (unit/s)	1.05 ± 0.48

Ambiguous. The selected datapoints show a different dynamic. The prevalence of Neutral and Happy instances in this region can be explained by the similarity with other classes (Sad and Angry) or the similarity between them. Like Angry samples in the hard-to-learn region, the Happy cases have longer strides and energetic arm movements. In contrast, Neutral samples maintain reduced strides and less upper limb oscillation, similar to Sad motion. On the other hand, the average gait velocity and step length of the datapoints in this region are higher than those observed in the hard-to-learn datapoints, which can be a feature of the difficulty in differentiating these two classes as an inherent characteristic. Figure 7b shows high intra- and inter-class differences. Neutral and happy samples include instances with strong body inclination, which may suggest that such characteristics make it difficult to associate these movements consistently with specific emotions. One of the happy cases, for example, shows less energetic body movement when compared to the same class in the easier cases; the same observation applies to neutral instances.

Hard-to-learn. Compared to other regions, our analysis shows that hard-to-learn samples frequently exhibited reduced step length and stride values within the selected data points in this region. Furthermore, we identified significant datapoints in these regions where the step/stride is below the expected average of 2. These characteristics suggest shorter steps that do not complete a full stride cycle, which may contribute to a biomechanical pattern of restrained gait. Furthermore, we observed slow, symmetrical arm movements, suggesting that fewer sweeping arm movements may contribute to class differentiation in this approach. Among the classes, the Angry emotion stands out for higher average gait velocity and energetic arm movement. Figure 7c reveals an important pattern common across all classes of hard-to-learn selected subsets. It is possible to observe moments without movement. This characteristic can suggest an important point in gait analysis: *which features should we consider when there is no movement, especially given that we are performing a spatio-temporal analysis?* In fact, learning quality strongly depends on temporal variation. When samples contain static parts, the extraction of dynamic features may be hindered, leading to a loss of relevant patterns. Since temporal convolution operates across changes across successive frames, the absence of variation may contribute to poor temporal representations, suggesting that there is an association between the absence of biomechanical characteristic variation and spatio-temporal cues interfering with training dynamics.

Moreover, redundancy introduced by static segments may lead the model to associate immobility with intrinsic aspects of a class, thereby making it harder to differentiate between emotions. Consequently, there is a higher probability of overfitting to irrelevant patterns. In light of this, the hard samples may be associated with many instances that contain non-representative frames.

5.2.1 Classes comparison

Within the evaluated dataset and model setting, the analysed samples suggest that higher-arousal emotions (e.g., Angry/Happy) tend to exhibit higher arm motion energy, while Neutral/Sad samples tend to be more restrained. Also, we observe that the highest step/stride ratios are more frequent in ambiguous and easy-to-learn samples, suggesting that shorter steps render a complex gait more challenging to master. Finally, restrained body movements in hard-to-learn samples, compared to expansive and relaxed movements from easy samples. Conversely, ambiguous cases exhibit significant metric variance, suggesting that the lack of a specific pattern

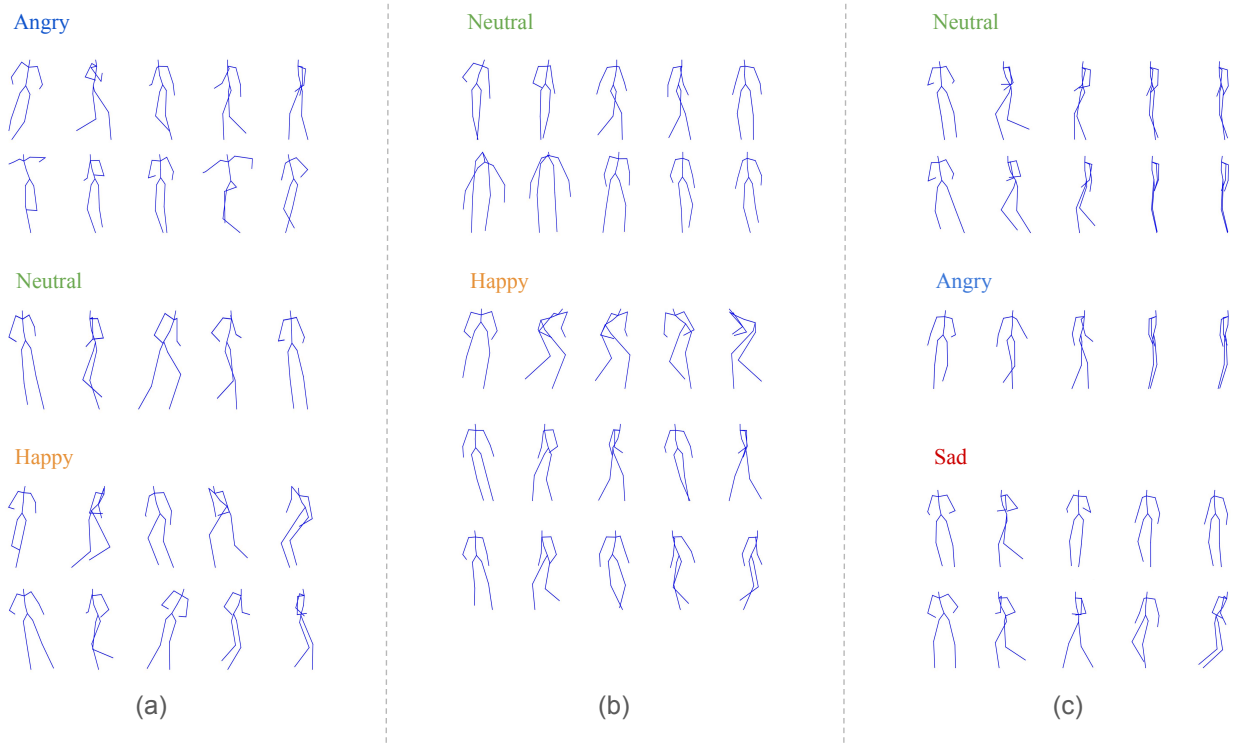

Figure 7. Subset of data selected by region: (a) easy-to-learn, (b) ambiguous, and (c) hard-to-learn

related to the emotion complicates the subtleties of emotion to be captured.

Neutral. This class has significant variability across different learning regions. While in the hard-to-learn region, these samples display moderate gait velocity with short steps, the ambiguous region presents higher speed and long strides, and the easy-to-learn region maintains a slow velocity but still shows long strides. These observations contribute to the idea of Neutral as a particularly challenging class to identify, as these diverse gait characteristics overlap considerably with patterns observed in other emotional classes. Furthermore, this observation amplifies a strong subject-specific bias: how an individual expresses a "neutral" gait highly depends on their baseline walking style. Consequently, establishing a personal baseline for each participant, collected across different emotional conditions, becomes essential. By anchoring the analysis in these individualized baselines, it is possible to interpret Neutral not as an absolute class, but as a relative state whose position in ambiguous regions of the map may reflect how easily it can be confused with other, similar gait expressions.

Happy. These cases show two different behaviors. There is a high variance between the gait velocity for collected samples in the ambiguous area, but always energetic arm movement and high stride (until 2.3). Compared to ambiguous cases, the same class has more consistency metrics for easy-to-learn regions, with slow gait velocity, but keeping high strides and energetic arm movement, suggesting that Happy gates are more associated with slow gait velocity, which is the slowest and has more symmetry than Angry cases.

Angry. Those instance in the hard-to-learn subset have high gait velocity, low step/stride rate, and subtle arm movement. On the other hand, easy-to-learn instances are slower, with high strides and energetic arm movements. This difference between instances may suggest that slow, expansive gait patterns make the sample more distinguishable, while the fast, restrained configuration increases classification complexity. This result suggests that the difference between the same emotion characteristics may arise from varying interpretations of the data annotation when an emotion is more restrained or energetic.

Sad. The samples from this class are found only in areas that are difficult to learn (this is also evident in cartographic maps). Generally, these samples exhibit asymmetric hand movements, while their gait velocity and step size show similarities with those from other classes in the same area. From a broader perspective, the presence of these samples within hard-to-learn region can be attributed to the limited number of instances. This scarcity presents a challenge for learning in this particular class, potentially leading to overlooked details that may be critical for differentiating these samples.

5.3 Limitations and Future works

Our work analysed dataset cartography as a visual tool for identifying class tendencies based on architectural difficulties in learning a sample during the training process. Although this may support the development of deep learning training strategies, this method has several limitations that must be acknowledged.

While the approach may provide early insights into the data and help identify inter-class trends that may disrupt training, its interpretations depend on diverse factors such as architecture, training dynamics, and dataset configuration. Cartographic regions such as easy, ambiguous, and hard are not predefined statistical populations, but rather emergent patterns

derived from model behavior during training. Therefore, cartography should be used as a visual support for deep learning decisions, rather than a definitive evidence of the structural properties of cartographic regions. As such, they should not be interpreted as inferentially validated categories. The biomechanical interpretations presented here are exploratory and dependent on the evaluated configuration. For future works, we may complement this approach with new dataset representation patterns that enable large-scale computation of biomechanical feature distributions and statistical tests to determine whether inter-class differences in kinematic descriptors are robust rather than artifacts of specific training dynamics.

Second, our findings are bound to the evaluated experimental settings with Emotion-Gait dataset and the ST-GCN architecture. Although ST-GCN is widely adopted for skeleton-based recognition tasks, different architectures, including transformer-based or alternative temporal aggregation models, may respond differently to similar motion patterns. Moreover, since the analysis was conducted on a single benchmark dataset, the generalizability of the identified difficulty patterns across datasets remains uncertain. In general, gait-based emotion recognition faces challenges similar to those in other areas of affective computing, such as inter-class overlap, subjective annotations, and dataset imbalance. However, it is also uniquely constrained by its spatio-temporal and biomechanical traits, which limit the generalization of findings beyond the evaluated dataset and model. Restrained movements, reduced step–stride relations, and static segments may hinder the extraction of meaningful temporal representations, but our conclusions are based on the specific ST-GCN configuration employed. Cross-dataset and multi-architecture evaluations represent important future steps to determine whether the observed learning instabilities are intrinsic to gait-based emotional expression or influenced by dataset-specific characteristics.

These results may inform the creation of new datasets and the use of existing ones for diagnostic pre-processing. Nevertheless, the proposed criteria for identifying prolonged immobility or low-variation sequences were not operationalized as automated filtering mechanisms in this study. Additional data characteristics, including class imbalance and high intra-class variation, can be diagnosed using the proposed approach, but their direct impact on downstream performance remains to be quantified. The identification of unstable, ambiguous, or potentially misannotated sequences suggests the need for manual inspection or reannotation; however, the effectiveness of such interventions should be empirically assessed in future work to determine whether they consistently improve model robustness.

An additional implication concerns the potential presence of annotation inconsistencies. Samples exhibiting biomechanical patterns that conflict with their affective labels may reveal subtle ambiguities detectable through training dynamics. However, the current dataset does not provide access to raw acquisition protocols or annotation procedures, limiting deeper validation. Future research based on controlled data collection may implement biomechanical consistency auditing frameworks grounded in domain literature, incorporating large-scale kinematic distribution analysis and clustering-based consistency checks.

Finally, our findings suggest that certain affective states may be inherently ambiguous when represented solely through skeletal motion, particularly in cases of restrained or low-variation gait patterns. Multimodal approaches integrating complementary cues, such as facial expressions or physiological signals, may mitigate these limitations by expanding the representational space. A promising direction involves cartography-informed multimodal fusion, in which regions of high uncertainty identified in unimodal training guide adaptive cross-modal integration. Comparative analyses between unimodal and multimodal training dynamics may clarify whether ambiguity detected in skeletal representations can be effectively reduced through multimodal learning.

6 Conclusion

In this study, combining computational cartography with biomechanical analysis provided an understanding of emotion recognition through gait analysis. For Emotion-Gait dataset combined with ST-GCN architecture, our results suggest that confidence–variability maps served as tool for distinguishing regions in easy, ambiguous, and hard learning. This approach allows for dependence onement of class-dependent tendencies and inherent structural limitations covered by the dataset.

The Emotion-gait dataset performed discrepancies in both of its subsets when tested with ST-GCN. The ELMD data highlighted strong class imbalance, where anger dominated the easy region and sadness remained persistently difficult to classify. Neutral and happy, concentrated in ambiguous zones, with high overlap between classes, and the instability of specific samples. When combined with complementary Emotion-Gait data, the complete Emotion-gait set reduced disparities and increased the prevalence of high-confidence cases. On the other hand, the joinment amplified inter-class overlap and changed the intra-class variability, which reduced the clear separation observed previously on ELMD data. In this case, expanding datasets by aggregation alone does not guarantee greater robustness, as suggested by the maintenance of heterogeneity between samples and remaining subject-specific central challenges.

The detailed sub-analysis of the available data reinforced the importance of biomechanical metrics in explaining why certain regions are systematically easier or harder to classify. The analysis suggests that restrained body movements, reduced step-stride relations, and static frames appear to be associated with hard-to-learn samples in the evaluated settings; still, expansive, slower, and more symmetric gait patterns better characterized easy-to-learn instances. For ambiguous samples, results support establishing subject-specific baselines to interpret neutrality as relative rather than absolute.

In conclusion, our findings indicate that gait properties such as step length, velocity, and arm kinematics emerge as candidate contextual descriptors associated with learning stability in skeleton-based emotion recognition. Rather than constituting definitive biomechanical markers, these features should be interpreted as exploratory indicators derived from training dynamics within the evaluated configuration. The used cartographic methodology should therefore be understood as a diagnostic pre-processing lens rather than a defini-

tive filtering mechanism. It may assist in identifying regions of instability, ambiguity, or potential annotation inconsistency, supporting more informed decisions regarding dataset curation and experimental design. Finally, our results suggest how inaccurately selected data that harms spatio-temporal features for learning may harm the architecture's performance, make it challenging to learn patterns, and impair its operation.

Future research should consider the applied methodology as a pre-processing step to plan alternatives to achieve robustness in others benchmarks and architectures. Also, this technique can serve as a basis for comparing different demographics and clarifying whether the context changes prevent the generalization of models.

Declarations

Acknowledgements

The authors would like to thank Voxar Labs at the Federal University of Pernambuco (UFPE) for providing institutional and technical support throughout this research. We also acknowledge the Coordination for the Improvement of Higher Education Personnel (CAPES) for financial support under Grant No. 88887.003983/2024-00.

Authors' Contributions

- **TS**. Conceptualization, Formal analysis, Methodology, Investigation, Software, Visualization, Writing - Original draft.
- **WLC**. Conceptualization, Project administration, Supervision, Writing - review and editing.
- **VT**. Funding acquisition, Supervision, Writing - review and editing.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The dataset used in this research is available at `https://go.umd.edu/emotion-gait`. Accessed on 26 April 2026. Please keep in mind that we are not the maintainers of this repository. The source code for the evaluation is available at `https://github.com/thifanysouza/emotion-cartography`. Accessed on 27 April 2026.

References

Bhattacharya, U., Mittal, T., Chandra, R., Randhavane, T., Bera, A., and Manocha, D. (2024). Step: Spatial temporal graph convolutional networks for emotion perception from gaits. DOI: https://doi.org/10.1609/aaai.v34i02.5490.

Habibie, I., Holden, D., Schwarz, J., Yearsley, J., and Komura, T. (2017). A recurrent variational autoencoder for human motion synthesis. In *Proceedings of the 28th British Machine Vision Conference (BMVC)*, London, United Kingdom. DOI: https://doi.org/10.5244/C.31.119.

Halovic, S. and Kroos, C. (2018). Not all is noticed: Kinematic cues of emotion-specific gait. *Human Movement Science*, 57:478–488. DOI: https://doi.org/10.1016/j.humov.2017.11.008.

Iwashita, Y., Kurazume, R., Ogawara, K., Tanaka, T., and Utsumi, A. (2013). Gait-based person identification robust to changes in appearance. *IEEE Transactions on Image Processing*, 22(6):2421–2431. DOI: https://doi.org/10.1109/TIP.2013.2246179.

Kipf, T. N. and Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR), 2017*.

Li, J., Dai, X., Yan, R., Tang, C., and Li, Y. (2025). Multi-anchor adaptive fusion and bi-focus attention for enhanced gait-based emotion recognition. *Scientific Reports*, 15:97922. DOI: https://doi.org/10.1038/s41598-025-97922-3.

Liao, R., Yu, S., An, W., and Huang, Y. (2020). A model-based gait recognition method with body pose and human prior knowledge. *Pattern Recognition*, 98:107069. DOI: https://doi.org/10.1016/j.patcog.2019.107069.

Lima, M. L., de Lima Costa, W., Martinez, E. T., and Teichrieb, V. (2024). St-gait++: Leveraging spatio-temporal convolutions for gait-based emotion recognition on videos. DOI: https://doi.org/10.48550/arXiv.2405.13903.

Lopez, L. D., Reschke, P. J., Knothe, J. M., and Walle, E. A. (2017). Postural communication of emotion: Perception of distinct poses of five discrete emotions. *Frontiers in Psychology, 8:710*. DOI: https://doi.org/10.3389/fpsyg.2017.00710.

Makihara, Y., Mannami, H., Tsuji, A., Hossain, M. A., Sugiura, K., Mori, A., and Yagi, Y. (2012). The ou-isir gait database comprising the treadmill dataset. *IPSJ Transactions on Computer Vision and Applications*, 4:53–62. DOI: https://doi.org/10.2197/ipsjtcva.4.53.

Montepare, J. M., Goldstein, S. B., and Clausen, A. (1987). The identification of emotions from gait information. *Journal of Nonverbal Behavior*, 11(1):33–42. DOI: https://doi.org/10.1007/BF00999605.

Reynolds, R. M., Novotny, E., Lee, J., Roth, D., and Bente, G. (2019). Ambiguous bodies: The role of displayed arousal in emotion [mis]perception. *Journal of Nonverbal Behavior*, 43:529–548. DOI: https://doi.org/10.1007/s10919-019-00312-3.

Roether, C. L., Omlor, L., Christensen, A., and Giese, M. A. (2009). Critical features for the perception of emotion from gait. *Journal of Vision, 9(6):15*. DOI: https://doi.org/10.1167/9.6.15.

Swayamdipta, S., Schwartz, R., Lourie, N., Wang, Y., Hajishirzi, H., Smith, N. A., and Choi, Y. (2020). Dataset cartography: Mapping and diagnosing datasets with training dynamics. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/2020.emnlp-main.746.

Valli, A. (2007). Notes on natural interaction. Available at: `https://www.cin.ufpe.br/~in1123/2017-1/leitura/Valli.pdf`. Accessed on 22 April 2026.

Winter, D. A. (2009). *Biomechanics and Motor Control of Human Movement*. John Wiley & Sons, Hoboken, New Jersey, 4th edition.

Yan, S., Xiong, Y., and Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32. DOI: https://doi.org/10.1609/aaai.v32i1.12328.

Yin, Y., Jing, L., Huang, F., Yang, G., and Wang,

Z. (2022). Msa-gcn:multiscale adaptive graph convolution network for gait emotion recognition. DOI: https://doi.org/10.48550/arXiv.2209.08988.

Zhai, Y., Jia, G., Lai, Y.-K., Zhang, J., Yang, J., and Tao, D. (2024). Looking into gait for perceiving emotions via bilateral posture and movement graph convolutional networks. *IEEE Transactions on Affective Computing*, 15(3):1634–1648. DOI: https://doi.org/10.1109/TAFFC.2024.3365694.

Zhou, J., Xiong, H., Lu, J., Lin, Z., and Feng, B. (2025). Cgtgait: Collaborative graph and transformer for gait emotion recognition. DOI: https://doi.org/10.48550/arXiv.2509.16623.

Zhou, T., Wang, S., and Bilmes, J. (2020). Curriculum learning by dynamic instance hardness. In *Advances in Neural Information Processing Systems*, volume 33, pages 8602–8613. Curran Associates, Inc.