Diffusion-Based Unsupervised Pre-training for Automated Recognition of Vitality Forms

Noemi Canovi
Dep. of Information Engineering and
Computer Science
University of Trento
Trento, Italy
noecanovi@gmail.com

Federico Montagna
Dep. of Information Engineering and
Computer Science
University of Trento
Trento, Italy
federico.montagna@studenti.unitn.it

Radoslaw Niewiadomski
Department of Informatics,
Bioengineering, Robotics and Systems
Engineering
University of Genova
Genoa, Italy
radoslaw.niewiadomski@unige.it

Alessandra Sciutti
CONTACT Unit
Istituto Italiano di Tecnologia
Genoa, Italy
alessandra.sciutti@iit.it

Giuseppe Di Cesare
Department of Medicine and Surgery
University of Parma,
CONTACT Unit
Istituto Italiano di Tecnologia
Italy
giuseppe.dicesare@unipr.it

Cigdem Beyan
Department of Computer Science
University of Verona
Verona, Italy
cigdem.beyan@univr.it

ABSTRACT

Social communication involves interpreting nonverbal behaviors, detecting and anticipating others' actions and intentions. Actions convey not only the goal and motor intention but also the form, i.e., variations in action execution. These variations, termed vitality forms, communicate attitudes during interactions, such as being gentle, calm, vigorous, and rude. Automatic vitality form recognition may have several applications in social robotics, social skills training, and therapy, yet it remains a rarely studied topic. This paper introduces an unsupervised pre-training approach that utilizes 2D-body key point trajectories as input and employs diffusion models to derive more effective features for representing these trajectories. The features learned from the diffusion model's encoder are utilized to train a multilayer perceptron for vitality form recognition. Experimental analysis showcases the superior performance of the proposed method not only across various videos but also for action classes not encountered during training.

CCS CONCEPTS

• **Human-centered computing → Collaborative and social computing**.

KEYWORDS

Vitality forms, nonverbal communication, unsupervised pre-training, diffusion models, autoencoders, gestures, actions, trajectory

AVI 2024, June 03–07, 2024, Arenzano, Genoa, Italy
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1764-2/24/06
https://doi.org/10.1145/3656650.3656689

ACM Reference Format:
Noemi Canovi, Federico Montagna, Radoslaw Niewiadomski, Alessandra Sciutti, Giuseppe Di Cesare, and Cigdem Beyan. 2024. Diffusion-Based Unsupervised Pre-training for Automated Recognition of Vitality Forms. In *International Conference on Advanced Visual Interfaces 2024 (AVI 2024), June 03–07, 2024, Arenzano, Genoa, Italy.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3656650.3656689

1 INTRODUCTION

The vitality forms, introduced by Daniel Stern [46] are the forms of actions that convey information about the performer's attitude. They can be manifested through movements (e.g., goal-oriented actions, communicative gestures) [12], voice [11], or touch [28]. Although extensively studied in psychology and neuroscience, [13, 45], the potential benefits of studying vitality forms extend across diverse research domains. For instance, incorporating the recognition of vitality forms in artificial agents, such as humanoid robots or virtual agents, would enhance their understanding of human interaction partners, e.g., in the context of human-agent politeness [19, 32]. Beyond social robotics, models for recognizing vitality forms can find applications in social skills training (e.g., simulating job interviews, presentation skills, leadership) or as part of training programs dedicated to neurodivergent individuals with reduced abilities to perceive and communicate vitality forms [40].

Unfortunately, there is a scarcity of *publicly available* datasets containing relevant ground-truth data for automatic recognition of vitality forms. Recently, Niewiadomski et al. [33] introduced a dataset collected during various daily-life actions or gestures, capturing different vitality forms. Additionally, to illustrate that vitality forms are distinct from gesture velocity modulation, the authors included classes such as *neutral* (see its definition in Section 4.1), *fast*, and *slow*. The dataset underwent benchmarking using traditional machine learning methods like Support Vector Machines (SVM) and Random Forest, employing motion descriptors built by the data captured with a Motion Capture System (MoCap). The analysis ultimately demonstrates the possibility of the automatic differentiation of vitality forms from other classes, along with presenting differences among vitality forms in terms of kinematic features. On the other hand, considering the aforementioned real-life applications such as social robotics, social skills development, and so forth, it is worth performing the automatic recognition of vitality forms by using non-invasive sensors such as cameras.

In this paper, we tackle the challenge of automatically recognizing vitality forms from videos. Our approach involves analyzing the movements of an individual during action performance. The body movement is represented using body-key points (also referred to as skeletons or body pose), which are extracted for each frame of the video. We explore various methods, including the one based on hand-crafted features similar to those employed in [33]. Another approach is adopted by a recent study [3], where trajectories of different body key points sensed by MoCap technology are represented as images and distinguished using Convolutional Neural Networks (CNNs). Additionally, we evaluate the efficacy of unsupervised pre-training with autoencoders, following a similar approach to the recent studies: [14, 34, 35]. Notably, for the first time in this study, we propose a diffusion model-based unsupervised pre-training approach for more intricate movement modeling. To the best of our knowledge, such a pipeline has not been introduced, not only for vitality form recognition but also for overall movement analysis, encompassing areas such as social signal processing [5], body movement-based emotion recognition [26], or movement quality recognition [49].

The proposed method leverages diffusion models, a category of generative deep learning models that have demonstrated improvements across diverse computer vision and multimedia tasks. These tasks include content generation [25], denoising [41], and segmentation [16]. More notably, diffusion models have also exhibited effectiveness in discriminative tasks such as image classification [51], object detection [8], and anomaly detection in videos [47, 48]. The input to our diffusion model comprises the 2D trajectories of body key points extracted by a standalone body pose extractor over a specific time frame. The objective is to harness the noise sampling, corruption, and reconstruction characteristics inherent in diffusion models to learn more compact and effective feature representations. The latent features of the encoder structure of the proposed diffusion model are then utilized as input for a dense neural network, specifically a Multilayer Perceptron (MLP), to facilitate the training of this classifier and carry out the inference process.

The experimental analysis carried out on the solely available dataset [33] showcases the superior performance of the proposed diffusion-based model. Our method outperforms all others not only in accurately predicting vitality forms across various actions but also in demonstrating robustness in predictions for previously unseen action classes. Our findings indicate that unsupervised pre-training methods (i.e., autoencoders and diffusion models) exhibit greater promise in generalizing well to novel classes for predicting vitality forms and other categories, including soft, fast, and neutral. This outcome holds promise for implementing the proposed method in real-life applications.

The remaining sections of the paper are organized as follows. Section 2 describes relevant works on automatic recognition of the vitality forms, expressive movement qualities, emotion recognition, and social signal processing from body movements. Section 3 describes the proposed method along with its implementation details. In Section 4.2, we provide details on the dataset used, experimental setups, and implementation details of the SOTA, and present the results with discussions. The paper concludes with a summary of the contributions of this study, along with elaborations about limitations and future work.

2 RELATED WORK

Despite numerous studies in psychology and neuroscience focusing on vitality forms [13, 45, 46], the literature on their automatic recognition is quite limited. A work on ***automatic recognition of vitality forms*** was presented in [33], which explores the spatiotemporal characteristics of various actions captured through a motion capture system for distinguishing vitality forms. The study extracts multiple hand-crafted features per body key point and employs traditional machine learning models for classification. The findings emphasize that recognizing vitality forms goes beyond features like velocity and acceleration, advocating for a more extensive feature set that models the spatiotemporal properties of body motion data. Our study distinguishes itself significantly from [33]. Firstly, we do not rely on motion capture data; instead, we present a methodology exclusively grounded in computer vision. Consequently, our methodology is non-intrusive, demonstrating the potential for integration, for instance, into a social robot to enhance its perceptual capabilities. Moreover, we opt not to rely on hand-crafted features. Instead, our approach leverages the trajectories of 2D-body key points, and the feature learning process is entirely unsupervised, employing a generative model known as the diffusion models. Through unsupervised pre-training, we extract effective features, facilitating proficient performance in the targeted task using a classifier.

The automatic classification of vitality forms, concerning methodology and data, is associated with topics such as the recognition of expressive movement qualities (e.g., [29–31]), emotion recognition from body movements (e.g., [2, 3, 9, 10, 21, 34, 37, 50]) and body pose-based social signal processing such as detection of emergent leaders [4], social role detection [20], and body language detection [1].

The advancements in the field of ***automatic recognition of expressive movement qualities*** are discussed in detail in a recent study [49]. As mentioned in [49], Laban Movement Analysis (LMA), developed by the choreographer Laban [24], is the most widely adopted movement system for formalizing movement qualities. For example, Hachimura et al. [17] detect poses corresponding to four Laban movement characteristics: Space, Weight, Shape, and Time. Four high-level features, each addressing one of them is defined and by observing the temporal changes in these feature values, they extract body movements associated with the different Laban characteristics. Similarly, Ran et al. [38] employ supervised learning to detect Laban qualities from Kinect data by introducing a comprehensive set of hand-crafted descriptors, including 100 features associated with Laban's qualities and an additional 6000 descriptors characterizing the skeleton data. More recently, Samadani et al. [42] propose a set of continuous measures of Laban Effort and Shape components in terms of low-level features such as position, kinetic energy, velocity, acceleration, and jerk, which are extracted from the motion capture data of hand and arm movements. Different from LMA, the other movement qualities automatically measured using hand-crafted features, and traditional machine learning methods such as SVMs are Fluidity [7], Impulsivity [31], Smoothness [27], Lightness and Fragility [30]. It is crucial to emphasize that the literature on the automatic recognition of expressive movement qualities is limited in terms of deep learning solutions, possibly due to the relatively smaller size of the collected datasets.

A recent survey [26] on ***emotion expression in human body posture and movement*** summarizes all the studies in this regard, highlighting the significance of this topic. For the analysis of emotional body gestures, existing methods commonly employ covariance matrices to capture spatial correlations among joints during human actions. Consequently, they leverage the geometric properties of the Riemannian manifold to extract features from these covariance matrices. For instance, Daoudi et al. [10] compute the Riemannian center of mass for each emotion using the training set and classified five emotions using the log-Euclidean Riemannian metric between the test data and class centers with a nearest-neighbor classifier. Kacem et al. [21] introduce a novel geometric measure and a pairwise proximity function Support Vector Machine (SVM) for emotion recognition based on the gesture covariance matrix. Instead, for temporal analysis, previous studies have extracted kinematic features to characterize emotional posture movements, including velocity, acceleration, force, fluency, height/vertical position, and so forth. For instance, Dael et al. [9] demonstrated the significance of such hand-crafted features as crucial posture features reflecting emotion. Piana et al. [37] utilize features such as contraction index, fluidity, and impulsiveness from posture movements, and employ an SVM classifier for automatic emotion recognition. Barliya et al. [2] suggested that happiness and anger are predominantly expressed through increased movement speed, arm swings, and cadence.

Recently, researchers have explored deep learning approaches to acquire discriminative emotional representations from body gestures in both spatial and temporal domains. For example, Beyan et al. [3] present an image representation of spatio-temporal skeleton data and subsequently employ a multiscale CNN structure to classify such images for emotional gesture recognition. Similarly, Wang et al. [50] encode the body skeleton data using gesture covariance matrices and also obtain 3D gesture images, similar to the method presented in [3]. These two representations are learned with a multiscale spatial network based on the Riemannian network architecture and a multiscale temporal network based on the CNN architecture, with fusion applied as the final step to recognize emotions. Rather than representing the trajectory of body skeleton data in the form of images, as done in [3, 50], Paoletti et al. [34] utilize the raw trajectories of full-body key points for a fixed time duration. These trajectories are input into a convolutional autoencoder for unsupervised pre-training. The latent features extracted from the autoencoder's encoder are then employed to train an MLP for emotion recognition. This method [26] demonstrates that the features learned through unsupervised pre-training are transferable across different datasets.

In this paper, we adapted the approach outlined in [3, 26] to assess its performance in comparison to our proposed method. Our method shares similarities with the approach presented in [26] as both involve unsupervised pre-training and use an MLP as the classifier. However, our method utilizes diffusion models, resulting in more effective features for vitality form recognition.

As mentioned earlier, ***body pose-based social signal processing*** is also relevant to the automatic recognition of vitality forms through movement analysis, as similar methodologies can be applied to both. For example, in the context of automatic leader detection and leadership style analysis, Feese et al. [15] utilize wearable motion sensors to extract nonverbal cues, including the number and average length of gestures and postures, which are indicative of behavioral mimicry. Beyan et al. [4] define nonverbal cues that represent the motion of 12 different body parts in terms of angles and extract the trace of each angle. Statistical measures of these traces, such as standard deviation, skewness, the number of zero crossings, and the number of mean crossings, are used to unsupervisedly train Deep Boltzmann Machines, providing more compact features for classification with SVMs in emergent leader detection. That study [4] is similar to our study in applying unsupervised feature learning, although the two generative models used in their work and ours are different.

In another study [43], the engagement of children playing with a robot was detected by analyzing posture and body motion. Joo et al. [20] present a relatively large collection of skeleton data during social interactions and gameplay, which can be used for role detection by extracting body movement-related features from the provided skeleton, once again highlighting the utility of body pose in interaction analysis.

3 METHODOLOGY

The proposed method comprises two stages. The first stage involves unsupervised pre-training, where features are learned without utilizing the ground-truth labels of the task at hand. The second stage involves recognition, encompassing the training of a classifier with the learned features and subsequent inference.

Specifically, our method relies on the trajectories of body key points extracted using a 2D body pose estimator. We have devised an approach that employs diffusion models for unsupervised pre-training, leveraging their reconstruction capability to potentially obtain more compact, less noisy, and transferable data representations. These representations are then fed as input to an MLP for Vitality Forms Recognition. It is crucial to emphasize that once the diffusion model is unsupervisedly trained, it is frozen and detached from the recognition part. An overview of the proposed method is presented in Fig. 1.

3.1 Preliminaries

Given a video of F frames, each containing a single person performing a certain action (also called gesture throughout this paper), we first extract K body points at each video frame. In case of missing body parts upon application of the 2D pose extractor, we apply spline interpolation that considers the timestamps as well. Spline interpolation was performed independently for each body part, emphasizing that it is not an interpolation between adjacent body parts but rather for the same type of body part using its detections over time.

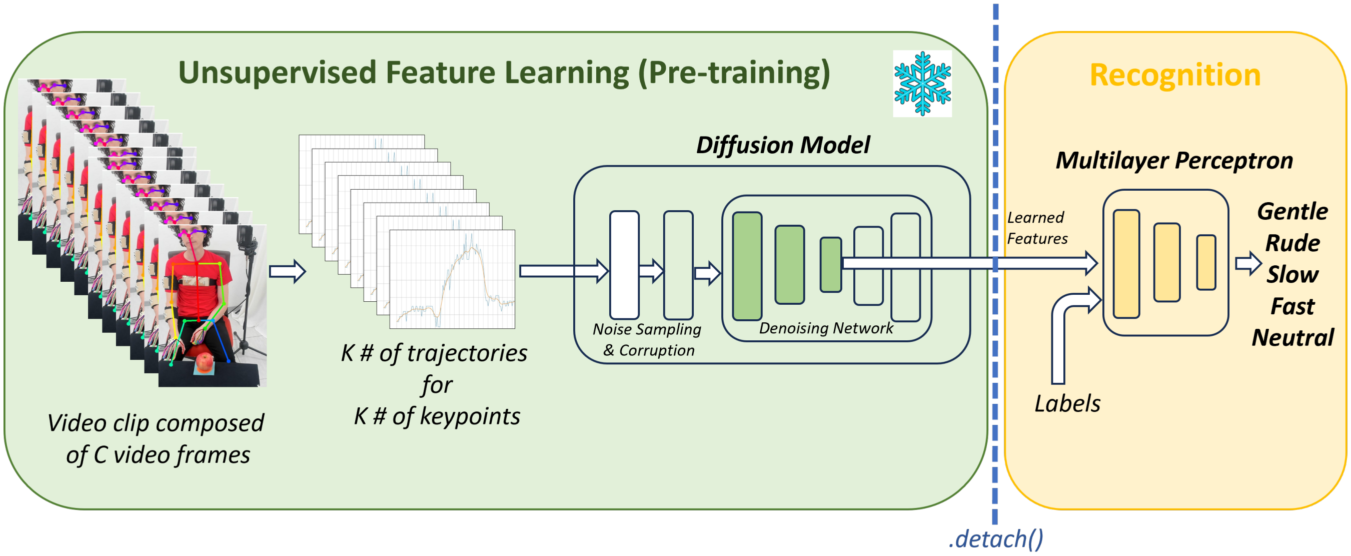
Unsupervised Feature Learning (Pre-training)
Video clip composed of C video frames
K # of trajectories for K # of keypoints
Diffusion Model
Noise Sampling & Corruption
Denoising Network
Learned Features
.detach()
Recognition
Multilayer Perceptron
Labels
Gentle
Rude
Slow
Fast
Neutral

Figure 1: The proposed method comprises two stages. The initial stage is *unsupervised feature learning (pre-training)*. The inputs for this stage include the trajectories of selected relevant body key points (refer to the text for details), obtained from the 2D pose estimator [6]. By employing the diffusion model, our objective is to generate a robust, compact, and transferable feature space. After completing the feature learning process, the learned features are extracted from the encoder of the diffusion model for use in the second stage, known as *recognition*. The diffusion model remains frozen and detached during the training and inference of the second stage. In the second stage, the learned features serve as input to a classifier (i.e., MLP), enhancing its capability to differentiate vitality forms (i.e., gentle and rude) among other classes such as slow, fast, and neutral. While the figure illustrates the pipeline for a video clip for simplicity, the actual task involves recognizing the entire video. This is accomplished by applying majority voting to the predictions made by the MLP for each video segment belonging to the same video.

We then divide this video into C equal segments (also called video clips), without performing any overlapping, such that each video segment contains C number of consecutive video frames, each is represented in terms of K body points. Such an approach allows us to obtain a fixed input space as well as augment the number of data fed to the diffusion models and the MLP. In case the last portion of the video is shorter than C frames, we replicate the corresponding body key points to arrive C number of data. Applying replications is indeed frequently applied in action recognition literature e.g., [34, 35]. It is important to clarify that an input of the diffusion model is a tensor of $C \times K \times 2$. Since the task involves recognizing the entire video, we additionally employ majority voting on the predicted classes after obtaining predictions for each corresponding video segment.

3.2 Diffusion Model

Diffusion models incorporate a stepwise addition of Gaussian noise ϵ_t with standard deviation σ_t to an input data point x_T sampled from a distribution $p_{data}(x)$ at each timestep $t \in [0, T]$. The resulting noised distribution $p(x, \sigma)$ transforms into an isotropic Gaussian, enabling efficient sampling of new data points $x_0 \sim \mathcal{N}(0, \sigma_{max}^2 \mathbf{I})$. These data points undergo gradual denoising with noise levels $\sigma_0 = \sigma_{max} > \sigma_1 > \cdots > \sigma_{T-1} > \sigma_T = 0$ to generate new samples. Diffusion models are trained by minimizing the expected L_2 error (also called Mean Squared Error (MSE)) between predicted and ground truth added noise [18], denoted as $\mathcal{L}_{simple} = |\epsilon_t - \hat{\epsilon}|_2$.

In this study, we adopt the diffusion formulation from [22] (called *k-diffusion* in the rest of this paper), motivated by its effectiveness to e.g., the DDPM [18] in other visual tasks. *k-diffusion* allows the network to predict either ϵ, x_0, or something in between based on the noise scale σ_t, which is also called σ-dependent skip connection. In this way, it mitigates the error amplification observed in DDPM [22].

Our denoising network D_θ is defined as:

$$D_\theta(x; \sigma_t) = c_{skip}(\sigma_t)\, x + c_{out}(\sigma_t)\, G_\theta\big(c_{in}(\sigma_t)\, x;\ c_{noise}(\sigma - T)\big), \tag{1}$$

where G_θ functions as the network undergoing training, c_{skip} regulates the skip connection, $c_{in}(\cdot)$ and $c_{out}(\cdot)$ adjust input and output magnitudes, and $c_{noise}(\cdot)$ scales σ. Our denoising network D_θ assumes an encoder-decoder structure. The input is noised trajectories, composed of two channels x, y with a fixed length. In the encoder layers, the length of the trajectory is progressively reduced, and the channels are increased. In the decoder layers, the length and the channels return gradually to their original size. In the encoder and decoder parts of the model, the time step σ_t is integrated through transformation via Fourier embedding and FiLM layers [36]. As demonstrated in [47, 48], we exploit the flexibility that denoising does not have to commence from noise with variance

σ^2_{max}; instead, it can initiate at any arbitrary timestep $t \in (0, T]$. Consequently, following the implementation in [47], we sample $fea_t \sim C(fea, \sigma^2_t)$ and apply the reverse diffusion process to reconstruct fea_T.

In summary, the overall diffusion process for an input of the network x and its reconstructed counterpart x_r consists of:

(1) Noise sampling: $\epsilon \sim \mathcal{N}(0, \mathbf{I})$,
(2) Diffusion input corruption: $x_t = x + \epsilon * \sigma_t$ and
(3) Reconstruction of the data with k-diffusion:
$x_r = sampling(D(x_t, \sigma_t))$.

3.3 Multilayer Perceptron

After training the proposed diffusion model in an unsupervised pre-training fashion, where the data labels are not utilized, we adhere to the principles of representation learning literature. Therefore, the generative model is frozen (and detached) and exclusively employed to extract features for both training and testing data. These features are then used to train a classifier, which is an MLP consisting of four fully connected layers with PReLU as the non-linearity and Batch Normalization. The inference is also performed only by the trained MLP.

3.4 Implementation Details

Extraction of 2D Poses and Trajectory Construction. Given that the dataset under consideration in this study comprises a single individual at a time seated on a chair that maintains a fixed position and distance from the camera and considering that actions are executed solely with one hand (specifically, the right hand), we chose to employ body key point filtering. Therefore, we concentrate our analysis and model development on a specific subset of body key points. These are indexed in OpenPose API [6] as 0 (nose), 1 (chest), 2 (right shoulder), 3 (right elbow), 4 (right wrist), and 8 (hip) for body skeleton, and 0 (right wrist), 9, and 13 for the right hand where 9 and 13 are two points in the palm. It is worth noting that this key point selection aligns with the motion capture system attached to the participant's body who performs an action in the used dataset [33].

For each selected key point (in total 8), we construct a trajectory with the length of C video frames. Therefore, a single data point as an input of the diffusion model is $8 \times C \times 2$, representing the x and y image coordinates, respectively.

Unsupervised Feature Learning. Throughout training, *k-diffusion* model is exposed to the data comprising fixed-length trajectories. As the model aims to acquire the ability to reconstruct these trajectories without relying on labels, during this unsupervised pre-training stage following the relevant literature of emotion and action recognition, e.g., [14, 34], the entire dataset is utilized. The optimization of the network's parameters is performed using the Adam optimizer, in conjunction with an inverse decay learning rate scheduler. This scheduler starts with the learning rate set to a default value of zero. Subsequently, it incrementally raised the learning rate until it reached its maximum value during the final epoch of the training.

The noise plays a crucial role in the diffusion process, necessitating the configuration of its distribution and parameters based on the task and dataset. In our implementation, the noise is drawn from a *log-normal* distribution with parameters P_{mean} and P_{std}, denoting the mean and standard deviation of that distribution. These parameters are linked to the maximum and minimum σ values (σ_{max} and σ_{min}) through the formula:

$$\sigma_{max}, \sigma_{min} = e^{P_{mean} \pm 5 P_{std}}, \tag{2}$$

making the parameter search decrease to two values instead of four [47].

Diverse combinations of P_{mean} and P_{std} values, alongside standard hyper-parameters learning rate, batch size, and weight decay, underwent testing and exploration. Random generators were employed to assign these values, ensuring a thorough evaluation of the model's performance. The considered ranges for each parameter are given in Table 1.

parameter	min value	max value
learning rate	0.00002	0.002
batch size	64	8192
weight decay	0	0.59
P_{mean}	-4	1.8
P_{std}	0.65	1.68

Table 1: Parameters range values which were tested during training of our diffusion model.

As mentioned before the reverse process of a diffusion model does not have to commence from the maximum noise level with variance σ^2_{max}. It can initiate from any noise level, at an arbitrary step $t \in [0, T]$, where a value close to zero implies a noised x_t closer to $\sigma^2_{max} = \sigma^2_0$, and a value closer to T indicates a noised x_t approaching the original data distribution. Depending on the value of t, the network can yield various loss values and reconstructed trajectories. Additionally, the same network can generate distinct loss values and trajectories in multiple runs of the evaluation process. Therefore, to ensure a fair comparison between models and at each time step t, the noise is initialized once and then kept fixed. As t increases and the noise decreases, the error tends to take lower values, and the reconstructed trajectories more closely resemble the original ones. We utilized 10-time steps similar to [47], marked by diverse t values, to act as inputs for the subsequent classification model, facilitating a thorough examination of the influence of various diffusion stages on the classification task.

Finally, given our objective, which is to enhance the classification model by inputting not raw trajectories but rather the learned features derived from our diffusion model, we extract learned features of size 512 after the encoder segment of the denoising network. Note that the size of the learned features is much smaller, indicating that we achieve a more compact data representation.

Recognition. The MLP network's input consists of the learned features extracted from the diffusion model, with a size of 512. The inputs of the four layers systematically shrink, with the last layer having the same number of neurons as the class label. In other words, the defined MLP comprises layers with the following inputs: 512, 256, 128, 64, and a, where a represents the number of classes, depending on the experiments. The MLP was trained using Cross-Entropy Loss, employing Adam as the optimizer. The learning rate was dynamically adjusted by the scheduler until the network's performance becomes a plateaue. Various parameter combinations were explored, including learning rates of 0.00005 and 0.0005, batch sizes of 8, 16, 32, and 64, weight decays of 0, 0.0001, and 0.001, and dropout rates of 0.1 and 0.5.

4 EXPERIMENTAL ANALYSIS & RESULTS

This section provides an overview of the dataset used, outlines the experimental setup, presents implementation details of the other methods adapted for comparison with the proposed method, and discusses the results.

4.1 Dataset

To assess and compare the effectiveness of the proposed method, experiments are conducted on the only available large-scale dataset for vitality forms recognition [33]. The dataset comprises various actions executed with two vitality forms: *gentle* and *rude*, along with the same actions performed at different speeds—*slow*, *fast*, and *neutral*. In this context, *neutral* denotes the functional aspect of the movement, such as passing an object, without involving any affective or communicative intention.

The choice of two specific vitality forms is based on previous fMRI studies showing neural responses when performing movements with these two vitality forms [13] and the actions are common ones that can occur during human-human and/or human-robot interaction. The involved actions are 1) grasping an object, 2) transferring an object to an agent, 3) raising a thumb, 4) pointing to another agent seated in front of the gesture maker, 5) placing an object down, 6) indicating a specific point on the table surface and 7) signaling the need for quiet or silence by lifting a finger to the mouth. Each action is performed by two professional theater actors (declare themselves as male and female) while the actions are performed in three directions: i) with 0 degrees derivation, ii) 45 degrees towards the right side of the performer, and iii) 45 degrees towards the left side of the performer.

The videos of the actions vary in length. In total, the dataset comprises 518 videos for the male actor: 151 in the rude class, 155 in the gentle class, 71 in the slow class, 71 in the fast class, and 70 in the neutral class. Additionally, there are 504 videos for the female actor: 143 in the rude class, 147 in the gentle class, 71 in the slow class, 72 in the fast class, and 71 in the neutral class. Overall, we used 1022 action executions, comprised of 302196 images, in which 145472 images belong to the male actor and 156724 images belong to the female actor.

4.2 Experiments

Following the experimental setup outlined in [33], we conducted leave-one-out and leave-one-action-out cross-validations. It is crucial to note that, in this study, leave-one-out refers to using the data from a single video as a test set while the remaining data is utilized for training the MLP classifier. In other words, the cross-validation folds are not divided based on trajectories which represent a video segment but not the entire video. This approach ensures comparability with the existing study presented in [33]. The results are assessed in terms of accuracy (ACC) and F1-score (F1) following the prior art.

4.3 Methods Employed for Comparisons

The implementation details of the methods utilized in this study for comparison with the proposed method are summarized as follows.

(1) **Niewiadomski et al. [33]** use motion capture data to extract 22 features composed of kinematics features such as velocity, acceleration, jerk, arc length, and curvature. For the classification of the five aforementioned classes, they use machine learning approaches such as SVMs with Radial Basis and polynomial kernels, k-NN, MLP, and Random Forest.

(2) Motivated by the performance of the features used in [33], we present a computer vision-based solution to extract semantically the same features (i.e., angles and velocities) from the selected key points (see Sec. 3.4 for the list of selected key points) obtained by using OpenPose [6]. Applying the formula from [4], we extracted angles between the following body key points: a) nose, neck, and middle hip, b) neck, right shoulder, and the right elbow, c) right shoulder, right elbow, and right wrist, d) right elbow, the midpoint between the right wrist from the body skeleton and the wrist from the skeleton's right hand, and the central point calculated from the average between two points on the right hand (9 and 13) to be used as features. Furthermore, we incorporated velocities calculated from two consecutive frames for the key points: nose, neck, right shoulder, right elbow, right wrist, and the right-hand middle finger knuckle. Once the aforementioned features were obtained, we implemented a bag-of-words strategy [44] for each of them to represent a single video in terms of it. The second and third derivatives of the key points, namely accelerations and jerks, were also considered as features. However, after applying bag-of-words, we observed that such features are not contributing, therefore, we limited the feature space to angles and velocities for this computer vision-based approach. As the classifier, we used an SVM with RBF kernel. We refer to this method throughout this paper as **Bag-of-Words**.

(3) We followed the methodology of **Beyan et al. [3]**, which represents 3D MoCap data as 8-bit RGB images. After obtaining these images, we applied the described augmentation from that study. Subsequently, we utilized all original and augmented images to train a single-head CNN structure, incorporating only the full images (referred to as coarse-grained representation in [3]) as input. The last layer of the CNN is specifically configured for the classification of the 5 classes of our study.

(4) Several studies on action recognition and emotion recognition [14, 23, 34, 35] have demonstrated the effectiveness of the autoencoders in unsupervised pre-training such that the learned features are further used for the training and testing of a classifier. We adopted the autoencoder structure of [14, 34], while we merged it with the MLP structure of the proposed method. The input of the autoencoder is a trajectory of length C, composed of two channels. At each encoder layer, the length of the trajectory gets progressively shrunk in half, while the channels (64 and 128) are increased. In the decoder layers, the opposite happens. Encoder and decoder blocks have 2 layers, and the latent representation is naturally placed in between the two blocks. During training, the network weights are changed iteratively through the Adam optimizer. The learning rate is changed accordingly

to a scheduler: when the loss reaches a plateau, i.e. it does not decrease in a certain number of epochs, it gets reduced. We explored different combinations of the parameter values for the learning rate (0.0001, 0.001, and 0.1), batch size (32, 64, 128, 256, 512), and latent representation dimension (128, 256, 512, 2048). The autoencoder was trained using the Mean Squared Error (MSE) loss. We refer to this method throughout this paper as **Autoencoder**.

4.4 Results

The results for the leave-one-out and leave-one-action-out settings are presented in Tables 2 and 3, respectively. In both settings, our proposed method outperforms all others.

In the leave-one-out setting, the second-best performance is achieved by [3], while the autoencoder-based approach performs 2% worse than [3], even though the feature learning is entirely performed without using ground-truth labels. Overall, the trajectory-based approaches surpass the methods relying on handcrafted features, regardless of the data modality. Another computer vision approach, the *bag-of-words*, despite utilizing fewer features, performs comparably to the method proposed by [33]. Importantly, the *bag-of-words* approach, along with the proposed method and the method with autoencoder, is less intrusive than [3, 33], as it does not require data captured with wearable sensors.

Table 2: The best results in terms of accuracy (ACC) and F1-score for leave-one-out setting. MoCap and CV stand for motion capture system and computer vision, respectively. The best and second best results are in black and underlined, respectively.

Approach	Data Modality	Feature Learning	ACC	F1-score
Niewiadomski et al. [33]	MoCap	Supervised	87.4	87.3
Bag-of-Words	CV	Supervised	86.7	86.9
Beyan et al. [3]	MoCap	Supervised	91.0	90.9
Autoencoder	CV	Unsupervised	88.9	88.8
Proposed (Diffusion)	CV	Unsupervised	**92.2**	**92.1**

Table 3: The best results in terms of accuracy (ACC) and F1-score for leave-one-action-out setting. MoCap and CV stand for motion capture system and computer vision, respectively. The best and second best results are in black and underlined, respectively.

Approach	Data Modality	Feature Learning	ACC	F1-score
Niewiadomski et al. [33]	MoCap	Supervised	74.7	74.2
Bag-of-Words	CV	Supervised	65.4	65.3
Beyan et al. [3]	MoCap	Supervised	76.8	76.6
Autoencoder	CV	Unsupervised	78.4	78.2
Proposed (Diffusion)	CV	Unsupervised	**83.0**	**83.1**

In the leave-one-action-out setting (Table 3), the results for all methods are lower than those in Table 2. There are possible explanations for such a result. Firstly, the training data size in the leave-one-out setting is much larger than in the leave-one-action-out setting, and having a model trained on a larger dataset typically generalizes better, especially for deep models. On the other hand, this setting requires the prediction of 5 classes considered in this study to be made on a completely different class of action in the test. Therefore, one can claim that the prediction of these 5 classes is also dependent on the type of action class, especially for the methods [3, 33] and bag-of-words. Furthermore, the results show the effectiveness of unsupervised feature pre-training, indicating that the features extracted from autoencoders and diffusion models are still more generalizable (i.e., transferable), given that they perform better than others.

Fig. 2 demonstrates the confusion matrices representing the performance of the proposed method in both cross-validation settings. In the leave-one-out setting, the neutral class exhibits the most accurate classification, followed by rude, slow, gentle, and fast. Conversely, in the leave-one-action-out setting, the highest performance is achieved for fast and gentle, followed by rude, slow, and neutral classes. Notably, several instances across various classes are misclassified as rude. For example, there is a relatively higher likelihood of predicting fast as rude (12%), neutral as rude (12%), gentle as rude (11.7%), and slow as rude (11%).

5 CONCLUSIONS

Vitality forms communicate the attitudes and intentions behind actions. In this paper, we have explored the automatic recognition of two vitality forms—gentle and rude—expressed in various daily life actions and gestures. We have also considered the same set of actions and gestures performed neutrally, slowly, and quickly.

We have introduced a new method for automatically recognizing vitality forms through the analysis of body motion data. Using 2D-body pose skeleton data detected by computer vision, we employ diffusion models to reconstruct features without relying on labels. These features, which are compact, informative, and transferable, are then used to train a classifier to distinguish between vitality form classes. Additionally, we explore several other methodologies from related fields, such as social signal processing, action recognition, and emotion recognition, to assess their performance and compare them to our proposed approach. Experimental analyses are conducted with a single action video as a test set and with all videos of an action class as the test set, confirming the effectiveness of our method. Notably, the features learned with diffusion models are found to be informative and compact, with a smaller size compared to the input. The results, particularly in a leave-one-action-out setting, highlight the transferability of our proposed method.

The main contributions of this study can be summarized as follows:

- For the first time, we demonstrated that automatic recognition of vitality forms can be achieved within a fully non-invasive pipeline through video data processing.
- We benchmarked the targeted task by incorporating several state-of-the-art works in related fields (such as emotion recognition) and illustrated that computer vision-based methods can outperform MoCap data-based approaches.
- We introduced a diffusion model-based unsupervised pre-training approach, surpassing all other SOTA methods when tested on various setups to differentiate gentle, rude, slow, fast, and neutral expressions during the performance of various daily-life actions and gestures.

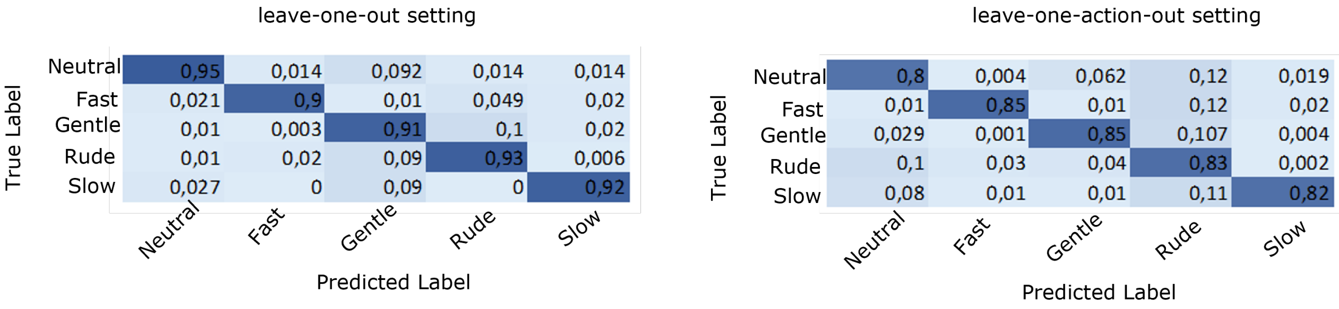

Figure 2: Confusion matrices illustrating the optimal performance of the proposed method align with those presented in Tables 2 and 3, respectively.

- By incorporating autoencoder-based unsupervised pre-training and proposing diffusion-based unsupervised pre-training, we demonstrated that such methods hold promise for effectiveness while exhibiting superior generalization across different action classes.

Limitations. We have only been able to assess the proposed method and our state-of-the-art implementations on a single dataset, as there is a scarcity of publicly available and annotated datasets for vitality forms. While the existing dataset is sufficiently large for effectively training deep models, encompassing a variety of action classes performed multiple times with inherent variations, it is necessary to conduct similar analyses on other action classes, involve more participants, and encompass additional vitality forms for a more comprehensive evaluation.

Future Work. We aim to study the automatic recognition of vitality forms from a multimodal perspective. We are interested in proposing effective unsupervised pre-training that yields informative features learned from different modalities (e.g., audio and video) to enhance the performance of vitality forms recognition. Furthermore, the proposed model will be integrated into a social robot to work in real-time interactions, especially in contexts where recognizing and expressing the appropriate attitudes is crucial. Examples of such scenarios include e.g., health care (robot receptionists in the hospitals [14]), and robot interviewers [39]. The other possible applications involve virtual reality (e.g., immersive virtual training environments to improve social skills), innovative forms of entertainment (e.g., video-games), security and surveillance (e.g., detection of aggressive behavior), therapy (e.g., support of autistic persons) and a large number of other multimodal interfaces allowing for natural human-like interaction such as virtual agents [32].

ACKNOWLEDGMENTS

Alessandra Sciutti is supported by a Starting Grant from the European Research Council (ERC) under the EU's Horizon 2020 research and innovation programme, G.A. No 804388, wHiSPER.

REFERENCES

[1] Michal Balazia, Philipp Müller, Ákos Levente Tánczos, August von Liechtenstein, and Francois Bremond. 2022. Bodily behaviors in social interaction: Novel annotations and state-of-the-art evaluation. In *Proceedings of the 30th ACM International Conference on Multimedia*. 70–79.

[2] Avi Barliya, Lars Omlor, Martin A Giese, Alain Berthoz, and Tamar Flash. 2013. Expression of emotion in the kinematics of locomotion. *Experimental brain research* 225 (2013), 159–176.

[3] Cigdem Beyan, Sukumar Karumuri, Gualtiero Volpe, Antonio Camurri, and Radoslaw Niewiadomski. 2021. Modeling Multiple Temporal Scales of Full-body Movements for Emotion Classification. *IEEE Transactions on Affective Computing* (2021), 1–1. https://doi.org/10.1109/TAFFC.2021.3095425

[4] Cigdem Beyan, Vasiliki-Maria Katsageorgiou, and Vittorio Murino. 2017. Moving as a leader: Detecting emergent leadership in small groups using body pose. In *Proceedings of the 25th ACM international conference on Multimedia*. 1425–1433.

[5] Cigdem Beyan, Alessandro Vinciarelli, and Alessio Del Bue. 2023. Co-Located Human–Human Interaction Analysis Using Nonverbal Cues: A Survey. *Comput. Surveys* 56, 5 (2023), 1–41.

[6] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019).

[7] George Caridakis, Amaryllis Raouzaiou, Elisabetta Bevacqua, Maurizio Mancini, Kostas Karpouzis, Lori Malatesta, and Catherine Pelachaud. 2007. Virtual agent multimodal mimicry of humans. *Language Resources and Evaluation* 41, 3 (2007), 367–388.

[8] S. Chen, P. Sun, Y. Song, and P. Luo. 2022. Diffusiondet: Diffusion model for object detection. *arXiv preprint:2211.09788* (2022).

[9] Nele Dael, Martijn Goudbeek, and Klaus R Scherer. 2013. Perceived gesture dynamics in nonverbal expression of emotion. *Perception* 42, 6 (2013), 642–657.

[10] Mohamed Daoudi, Stefano Berretti, Pietro Pala, Yvonne Delevoye, and Alberto Del Bimbo. 2017. Emotion recognition by body movement representation on the manifold of symmetric positive definite matrices. In *Image Analysis and Processing-ICIAP 2017: 19th International Conference, Catania, Italy, September 11-15, 2017, Proceedings, Part I 19*. Springer, 550–560.

[11] G Di Cesare, V Cuccio, M Marchi, A Sciutti, and G Rizzolatti. 2021. Communicative And Affective Components in Processing Auditory Vitality Forms: An fMRI Study. *Cerebral Cortex* 32, 5 (08 2021), 909–918. https://doi.org/10.1093/cercor/bhab255

[12] Giuseppe Di Cesare, Elisa De Stefani, Maurizio Gentilucci, and Doriana De Marco. 2017. Vitality Forms Expressed by Others Modulate Our Own Motor Response: A Kinematic Study. *Frontiers in Human Neuroscience* 11 (2017). https://doi.org/10.3389/fnhum.2017.00565

[13] Giuseppe Di Cesare, Cinzia Di Dio, Magali J. Rochat, Corrado Sinigaglia, Nadia Bruschweiler-Stern, Daniel N. Stern, and Giacomo Rizzolatti. 2013. The neural correlates of 'vitality form' recognition: an fMRI study: This work is dedicated to Daniel Stern, whose immeasurable contribution to science has inspired our research. *Social Cognitive and Affective Neuroscience* 9, 7 (2013), 951–960. https://doi.org/10.1093/scan/nst068

[14] Moreno D'incà, Cigdem Beyan, Radoslaw Niewiadomski, Simone Barattin, and Nicu Sebe. 2023. Unleashing the Transferability Power of Unsupervised Pre-Training for Emotion Recognition in Masked and Unmasked Facial Images. *IEEE Access* 11 (2023), 90876–90890. https://doi.org/10.1109/ACCESS.2023.3308047

[15] Sebastian Feese, Bert Arnrich, Gerhard Tröster, Bertolt Meyer, and Klaus Jonas. 2012. Quantifying Behavioral Mimicry by Automatic Detection of Nonverbal Cues from Body Motion. In *International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*. 520–525. https://doi.org/10.1109/SocialCom-PASSAT.2012.48

[16] Zhangxuan Gu, Haoxing Chen, Zhuoer Xu, and Lan et al. 2022. DiffusionInst: Diffusion Model for Instance Segmentation. *arXiv preprint:2212.02773* (2022).

[17] Kozaburo Hachimura, Katsumi Takashina, and Mitsu Yoshimura. 2005. Analysis and evaluation of dancing movement based on LMA. In *Robot and Human Interactive Communication, 2005. ROMAN 2005. IEEE International Workshop on*. IEEE, 294–299.

[18] J. Ho, A. Jain, and P. Abbeel. 2020. Denoising diffusion probabilistic models. *NeurIPS* 33 (2020), 6840–6851.

[19] Ohad Inbar and Joachim Meyer. 2019. Politeness Counts: Perceptions of Peacekeeping Robots. *IEEE Transactions on Human-Machine Systems* 49, 3 (2019), 232–240. https://doi.org/10.1109/THMS.2019.2900337

[20] Hanbyul Joo, Tomas Simon, Mina Cikara, and Yaser Sheikh. 2019. Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10873–10883.

[21] Anis Kacem, Mohamed Daoudi, Boulbaba Ben Amor, Stefano Berretti, and Juan Carlos Alvarez-Paiva. 2018. A novel geometric framework on gram matrix trajectories for human behavior understanding. *IEEE transactions on pattern analysis and machine intelligence* 42, 1 (2018), 1–14.

[22] T. Karras, M. Aittala, Timo Aila, and Samuli Laine. 2022. Elucidating the Design Space of Diffusion-Based Generative Models. In *NeurIPS*.

[23] Panagiotis Koromilas and Theodoros Giannakopoulos. 2021. Unsupervised multimodal language representations using convolutional autoencoders. *arXiv preprint arXiv:2110.03007* (2021).

[24] Rudolf Laban and Frederick Charles Lawrence. 1947. *Effort*. Macdonald & Evans.

[25] Haohe Liu, Zehua Chen, and Yi et al. Yuan. 2023. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. *arXiv preprint arXiv:2301.12503* (2023).

[26] Mehdi-Antoine Mahfoudi, Alexandre Meyer, Thibaut Gaudin, Axel Buendia, and Saida Bouakaz. 2023. Emotion Expression in Human Body Posture and Movement: A Survey on Intelligible Motion Factors, Quantification and Validation. *IEEE Transactions on Affective Computing* 14, 4 (2023), 2697–2721. https://doi.org/10.1109/TAFFC.2022.3226252

[27] Barbara Mazzarino and Maurizio Mancini. 2009. The Need for Impulsivity & Smoothness - Improving HCI by Qualitatively Measuring New High-Level Human Motion Features. *SIGMAP 2009 - International Conference on Signal Processing and Multimedia Applications, Proceedings*, 62–67.

[28] Radoslaw Niewiadomski, Cigdem Beyan, and Alessandra Sciutti. 2023. Affect Recognition in Hand-Object Interaction Using Object-Sensed Tactile and Kinematic Data. *IEEE Transactions on Haptics* 16, 1 (2023), 112–117. https://doi.org/10.1109/TOH.2022.3230643

[29] R. Niewiadomski, M. Mancini, and S. Piana. 2013. Human and virtual agent expressive gesture quality analysis and synthesis. In *Coverbal Synchrony in Human-Machine Interaction*, M. Rojc and N. Campbell (Eds.). CRC Press, 269–292.

[30] Radoslaw Niewiadomski, Maurizio Mancini, Stefano Piana, Paolo Alborno, Gualtiero Volpe, and Antonio Camurri. 2017. Low-intrusive Recognition of Expressive Movement Qualities. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction* (Glasgow, UK) *(ICMI 2017)*. 230–237.

[31] Radoslaw Niewiadomski, Maurizio Mancini, Gualtiero Volpe, and Antonio Camurri. 2015. Automated Detection of Impulsive Movements in HCI. In *Proceedings of the 11th Biannual Conference on Italian SIGCHI Chapter* (Rome, Italy) *(CHItaly 2015)*. ACM, New York, NY, USA, 166–169. https://doi.org/10.1145/2808435.2808466

[32] Radoslaw Niewiadomski and Catherine Pelachaud. 2010. Affect expression in ECAs: Application to politeness displays. *Int. J. Hum.-Comput. Stud.* 68, 11 (Nov. 2010), 851–871. https://doi.org/10.1016/j.ijhcs.2010.07.004

[33] Radoslaw Niewiadomski, Amrita Suresh, Alessandra Sciutti, and Giuseppe Di Cesare. 2023. Vitality forms analysis and automatic recognition. *Authorea Preprints* (2023).

[34] Giancarlo Paoletti, Cigdem Beyan, and Alessio Del Bue. 2022. Graph Laplacian-Improved Convolutional Residual Autoencoder for Unsupervised Human Action and Emotion Recognition. *IEEE Access* 10 (2022), 131128–131143. https://doi.org/10.1109/ACCESS.2022.3229478

[35] Giancarlo Paoletti, Jacopo Cavazza, Cigdem Beyan, and Alessio Del Bue. 2021. Unsupervised Human Action Recognition with Skeletal Graph Laplacian and Self-Supervised Viewpoints Invariance. In *The 32nd British Machine Vision Conference (BMVC)*.

[36] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[37] Stefano Piana, Alessandra Staglianò, Francesca Odone, and Antonio Camurri. 2016. Adaptive Body Gesture Representation for Automatic Emotion Recognition. *ACM Trans. on Interactive Intelligent Systems* 6, 1 (2016), 6:1–6:31. https://doi.org/10.1145/2818740

[38] Bernstein Ran, Shafir Tal, Tsachor Rachelle, Studd Karen, and Schuster Assaf. 2015. Multitask Learning for Laban Movement Analysis. In *Proceedings of the 2nd International Workshop on Movement and Computing* (Vancouver, British Columbia, Canada) *(MOCO '15)*. 37–44.

[39] Maria Elena Lechuga Redondo, Radoslaw Niewiadomski, Francesco Rea, Sara Incao, Giulio Sandini, and Alessandra Sciutti. 2023. Comfortability Analysis Under a Human–Robot Interaction Perspective. *International Journal of Social Robotics* (2023). https://doi.org/10.1007/s12369-023-01026-9

[40] Magali J. Rochat, Vania Veroni, and Nadia Bruschweiler-Stern et al. 2013. Impaired vitality form recognition in autism. *Neuropsychologia* 51, 10 (2013), 1918–1924. https://doi.org/10.1016/j.neuropsychologia.2013.06.002

[41] Chitwan Saharia, William Chan, and Saurabh et al. Saxena. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487* (2022).

[42] A. Samadani, S. Burton, R. Gorbet, and D. Kulic. 2013. Laban Effort and Shape Analysis of Affective Hand and Arm Movements. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. 343–348. https://doi.org/10.1109/ACII.2013.63

[43] Jyotirmay Sanghvi, Ginevra Castellano, Iolanda Leite, André Pereira, Peter W. McOwan, and Ana Paiva. 2011. Automatic analysis of affective postures and body motion to detect engagement with a game companion. In *6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 305–311.

[44] Sivic and Zisserman. 2003. Video Google: A text retrieval approach to object matching in videos. In *Proceedings ninth IEEE international conference on computer vision*. IEEE, 1470–1477.

[45] Daniel N. Stern. 1999. Vitality contours: The temporal contour of feelings as a basic unit for constructing the infant's social experience.. In *Early social cognition: Understanding others in the first months of life*. 67–80.

[46] Daniel N. Stern. 2010. *Forms of vitality exploring dynamic experience in psychology, arts, psychotherapy, and development*. Oxford University Press.

[47] Anil Osman Tur, Nicola Dall'Asen, Cigdem Beyan, and Elisa Ricci. 2023. Exploring diffusion models for unsupervised video anomaly detection. In *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2540–2544.

[48] Anil Osman Tur, Nicola Dall'Asen, Cigdem Beyan, and Elisa Ricci. 2023. Unsupervised Video Anomaly Detection with Diffusion Models Conditioned on Compact Motion Representations. In *International Conference on Image Analysis and Processing*. Springer, 49–62.

[49] Giovanna Varni and Maurizio Mancini. 2020. *Movement Expressivity Analysis: From Theory to Computation*. 213–233. https://doi.org/10.1007//978-3-030-46732-6_11

[50] Tao Wang, Shuang Liu, Feng He, Weina Dai, Minghao Du, Yufeng Ke, and Dong Ming. 2023. Emotion Recognition From Full-Body Motion Using Multiscale Spatio-Temporal Network. *IEEE Transactions on Affective Computing* (2023), 1–15. https://doi.org/10.1109/TAFFC.2023.3305197

[51] Xingyi Yang and Xinchao Wang. 2023. Diffusion Model as Representation Learner. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 18938–18949.