Signal Processing: Image Communication 143 (2026) 117489

Contents lists available at ScienceDirect

Signal Processing: Image Communication

journal homepage: www.elsevier.com/locate/image

# Markerless emotion recognition from full-body movements for Social XR

Michael Neri [a], Sara Baldoni [b,*], Marco Carli [c], Federica Battisti [b]

[a] Tampere University, Faculty of Information Technologies and Communication Sciences, Korkeakoulunkatu, 1, Tampere, 33720, Finland
[b] University of Padova, Department of Information Engineering, Via Gradenigo, 6/B, Padova, 35131, Italy
[c] Roma Tre University, Department of Industrial, Electronic, and Mechanical Engineering, Via Vito Volterra, 62, Rome, 00146, Italy

ARTICLE INFO

Dataset link: https://github.com/michaelneri/emotion-recognition-human-movements

*Keywords:*
Emotion recognition
Deep learning
Social XR
2D skeleton representation
Markerless

ABSTRACT

In this work, an emotion recognition system for enhancing social XR applications is presented. Although several techniques for emotion recognition have been proposed in the literature, they either require invasive and advanced equipment or exploit facial expressions, speech excerpts, physiological data, and text. In this contribution, on the contrary, an approach for markerless emotion classification through body language is designed. More specifically, human movements are analyzed over time by extracting the skeleton joints in videos acquired by consumer cameras. A normalization procedure has been introduced to provide a depth-independent skeleton representation without distorting the skeleton shape. The performance of the proposed method have been assessed using a dataset of videos recorded from multiple points of view. An ad-hoc learning-based emotion classifier has been trained to recognize four emotions (happiness, boredom, interest, and disgust) achieving an average accuracy of 72.5%. The pre-processed dataset, code, and demo with pre-trained models are available at https://github.com/michaelneri/emotion-recognition-human-movements.

## 1. Introduction

Emotion recognition consists in understanding the emotional state of a person. Emotions can be defined as *processes that involve massive, interrelated changes in several organismic subsystems occurring in response to an eliciting event of major significance to the individual* [1].

The way subjects express their emotional status is a relevant component of human communication [2]. This holds true for both direct human-to-human interaction and mediated communication [3]. In this direction, a relevant research field focuses on emotion recognition and conveyance in Extended Reality (XR)-based communication systems. Indeed, critical factors for the success of multi-user XR applications include the perceived social presence and the quality of interaction. As highlighted in [4], the introduction of realistic social cues, gestures, and communication mechanisms allows the users to engage in genuine social interactions, leading to higher levels of engagement and immersion. To enhance these perceptions, the ability to reflect the emotional status of the involved users in the virtual environment is of paramount importance. In [5], for instance, the authors underline that during XR remote conversations, either informative or emotional ones, the inclusion of emotion-related animations had a significant impact on social presence and communication quality. As highlighted in [6], regardless of whether participants are represented through 3D scans or avatars, it is essential to preserve human social signals, with particular attention to non-verbal communication cues [7]. Therefore, the capability to automatically recognize emotions could help reflecting them in the virtual user representation, providing an experience that closely resembles real-world iteration. In addition, automatic emotion recognition could enable the seamless adaptation of the XR content to better suit users' preferences and expectations, further enhancing the overall experience.

While humans unconsciously analyze verbal and nonverbal behaviors to understand a person's feelings, the design of a system that automatically recognizes emotions is still a challenging task. Most of the approaches proposed in the literature exploit physiological signals. In [8], for instance, Electroencephalogram (EEG), Heart Rate Variability (HRV), functional Near-InfraRed Spectroscopy (fNIRS), and Galvanic Skin Response (GSR) have been used to evaluate the emotional status of the users while viewing multimedia content in Virtual Reality (VR). However, the acquisition of physiological data requires the use of contact sensors, making these techniques invasive and annoying. To cope with this issue, an alternative approach consists of analyzing behavioral cues (eye and head/body movements) and audiovisual data (facial/vocal expressions) [9].

While a large number of studies on emotion classification have focused on facial expressions, their application to XR-mediated communication may be partially impaired by the usage of the Head-Mounted

* Corresponding author.
*E-mail addresses:* michael.neri@tuni.fi (M. Neri), sara.baldoni@unipd.it (S. Baldoni), marco.carli@uniroma3.it (M. Carli), federica.battisti@unipd.it (F. Battisti).

https://doi.org/10.1016/j.image.2026.117489
Received 28 October 2024; Received in revised form 7 July 2025; Accepted 9 January 2026
Available online 21 January 2026

Display (HMD) needed for the fruition of the immersive experience [6]. In addition, depending on the distance between the subject and the camera, facial features may be difficult due to a reduction in spatial resolution [10]. A possible solution consists in relying on different emotional channels. In fact, it is well known that human movements and posture can be associated with specific emotional states [11], and it has been demonstrated that the analysis of the movements of arms and upper body provide sufficient information to discriminate between happiness, sadness, anger, and fear [12]. Moreover, the lower part of the body can provide useful insights into the emotional state. As described in [13], if something attracts our attention, the feet will be oriented toward the object of interest; otherwise at least one of the feet will generally face outward. Furthermore, in applications such as social VR, full body tracking is usually employed so that every user's action is reflected in the virtual world. On the one hand, this feature is a key element for making this social VR truly immersive [7], and on the other it makes body-based emotion recognition particularly suitable for this type of applications.

With respect to other approaches, body-based emotion recognition has the following advantages: (i) subjects are allowed to be relatively far from the data capturing system, (ii) body language requires less consciousness, thus being less prone to manipulation and imitation, and (iii) data collection requires less cooperation from the subject so that the emotion recognition task does not interfere with his/her spontaneous behavior [3]. Despite this, emotion recognition from body language is yet a poorly investigated topic [14]. Following this cue, we design an emotion recognition system based only on body language.

To analyze human movements, motion capture systems can be employed [15]. However, these methods often require users to wear reflective markers, special T-shirts, or gloves, thus compromising freedom of movement.

For this reason, in this work, we aim to design a markerless emotion recognition system. In more detail, we propose an approach that relies on non-calibrated video streams acquired from multiple viewpoints for extracting the human skeleton. Then, the skeleton joints are processed over time to classify the emotion expressed by the user. The proposed method does not require the facial keypoints, but exploits a single skeleton joint for representing the head position, thus being compliant with the usage of HMDs. In addition, employing commercial RGB cameras, the hardware deployed for video acquisition is low-cost and lightweight, thus being suitable for its usage in several application scenarios. To the best of our knowledge, the proposed method for emotion recognition based exclusively on body language is the first approach that is jointly non-invasive, HMD-compliant and Commercial Off-the-Shelf (COTS).

The contributions of this paper can be summarized as follows:

- the definition of a human skeleton representation which does not change with the distance between the user and the camera;
- the introduction of a new approach for classifying emotions based exclusively on body movements;
- the extension of Human Emotion Recognition Observing the Evolution of the Skeleton (HEROES) [16] dataset employing multiple points of view for classifying four emotional states: happiness, interest, boredom, and disgust.

The dataset used for performance assessment, together with the developed code and demos with pre-trained models, are available at *https://github.com/michaelneri/emotion-recognition-human-movements*. The remainder of the paper is organized as follows. In Section 2 the related literature is reviewed. In Section 3 the proposed emotion model is presented and the collected dataset is described. In Section 4 the proposed approach is detailed and in Section 5 the results are presented and discussed. Finally, in Section 6 and in Section 7 a comparison with state-of-the-art approaches is provided and the conclusions are drawn.

## 2. Related works

### 2.1. Emotion datasets

Many datasets have been proposed in the field of affective computing for recognizing emotions from behavioral cues. They can be classified according to two main criteria: the type of recorded information and the usage of markers. The majority of datasets focus on facial expressions and upper body movements. A first example is represented by the Bimodal Face and Body Gesture Database (FABO) that includes 206 videos in which 10 emotions are expressed. In addition, one of the most used datasets is the Geneva Multimodal Emotion Portrayals (GEMEP) [17]. Differently from FABO, it includes a large number of videos (7000) in which 10 professional actors expressed 18 emotions. Moreover, the Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) [15] is an example of a marker-based emotion recognition dataset in which facial expressions, head movements, and hand gestures have been recorded.

Datasets in which the whole body is considered have generally been collected with the use of markers, providing 3D joint coordinates. Ma et al. recorded the movements of 30 non-professional actors while expressing 4 emotions [18]. Data have been acquired through a motion capture system composed of a suit equipped with retro-reflective markers and eight cameras. Similarly, the Emilya dataset consists of 8 emotions expressed by 12 actors and recorded through a motion capture system and four cameras [19,20]. In this direction, Zhang et al. presented a kinematic dataset composed of 1402 recordings acquired through a portable motion capture system [21]. The dataset has been acquired thanks to the participation of 22 semi-professional actors. Another video dataset is the EGBM dataset [22] which has been collected from 16 professional actors performing 7 emotions using a Kinect V2 sensor. Differently, the MPI dataset [23] has been recorded using the Xsens MVN motion capture system, encompassing 11 emotions from 8 actors. DMCD database [24] has been collected using the PhaseSpace Impulse X2 MoCap system and contains 12 emotions performed by 6 dancers.

Although marker-based systems enable accurate detection of users' movements, they also require the use of specific hardware that can hinder spontaneous emotion expression. In addition, these systems cannot be deployed in large-scale application scenarios due to the need for marker installation.

Available markerless emotion datasets that record the whole body are small in size. An example is represented by the HUMAINE database, [25], in which 8 emotions have been expressed by 10 people and recorded through an RGB camera. The dataset is composed of 50 video clips only. In addition, in [16], the HEROES dataset was presented. It is composed of 256 videos recorded from a frontal RGB camera. In this work, the HEROES dataset is extended by considering the videos acquired from three different viewpoints as will be detailed in Section 3.2.

### 2.2. Emotion classifiers

Different types of human cues can be employed to differentiate among emotional states. A summary of the approaches proposed in the literature is presented in Table 1 and a detailed analysis is provided in the following.

#### 2.2.1. Physiological data

Physiological cues are a well-known indicator of human emotional states. Among the possible signals, the EEG is widely adopted. In [26], the authors computed the first and second-order Differential Entropy (DE) features of the EEG signal and provided them as input to a fully linear autoencoder. Then, an ensemble of classifiers (Decision Tree, k-nearest Neighbor, and Random Forest) was used with a soft-voting strategy to provide robust labels. Similarly, in [27], the authors used a

**Table 1**
State-of-the-art approaches for emotion recognition.

| SoA approach | Parameters | | | | Method features | | |
|---|---|---|---|---|---|---|---|
| | Physiological | Audio/Text | Body movements | Facial expressions | Non-invasive | HMD-compliant | COTS |
| Quing et al. [26] | Yes | No | No | No | No | Yes | No |
| Li et al. [27] | Yes | No | No | No | No | Yes | No |
| Zhang et al. [28] | Yes | No | No | No | No | Yes | No |
| Zhang et al. [29] | Yes | No | No | No | No | Yes | No |
| Zheng et al. [30] | Yes | No | No | No | No | Yes | No |
| Zhang et al. [31] | Yes | No | No | No | No | Yes | No |
| Zhang et al. [32] | No | Yes | No | No | Yes | No | Yes |
| Zhang et al. [33] | Yes | No | No | Yes | No | No | No |
| Guo et al. [34] | Yes | No | No | Yes | No | No | No |
| Ayari et al. [35] | No | Yes | No | Yes | Yes | No | Yes |
| Wang et al. [36] | No | Yes | No | Yes | Yes | No | Yes |
| Shi et al. [37] | No | Yes | Yes (3D) | No | Yes | No | No |
| Filntisis et al. [38] | No | No | Yes (2D) | Yes | Yes | No | Yes |
| Ahmed et al. [12] | No | No | Yes (3D) | No | Yes | Yes | No |
| Zhang et al. [10] | No | No | Yes (3D) | No | No | Yes | No |
| Wang et al. [39] | No | No | Yes (3D) | No | No | Yes | No |
| Chen et al. [40] | No | No | Yes (3D) | No | No | Yes | No |
| Zhai et al. [41] | No | No | Yes (3D) | No | No | Yes | No |
| Yumeng et al. [42] | No | No | Yes (3D) | No | No | Yes | No |
| Oğuz et al. [43] | No | No | Yes (3D) | No | No | Yes | No |
| Beyan et al. [44] | No | No | Yes (3D) | No | No | Yes | No |
| Ghaleb et al. [45] | **No** | **No** | **Yes (2D)** | **No** | **Yes** | **Yes** | **Yes** |
| **Proposed method** | **No** | **No** | **Yes (2D)** | **No** | **Yes** | **Yes** | **Yes** |

DE approach to characterize each EEG channel. The classification has been carried out by applying binary Support Vector Machines (SVMs) to each emotion pair, using a one-to-one voting strategy. In addition, in [28], a two-step spatio-temporal emotion recognition framework for the EEG signal has been proposed. More specifically, the authors introduced a self-attention network to consider both short-term continuity and long-term similarity of emotions, and they exploited the spatial correlation to identify relevant EEG channels.

Differently, in [29], the authors employed multiple physiological signals, i.e., Electromyography (EMG), GSR, and Respiratory rate (RES). In more detail, a multimodal classifier has been defined by introducing a regularized deep fusion of kernel machines. Low-level features have been extracted and given as input to Multi Layer Perceptrons (MLPs). Recently, in [31] the authors collected data from an armband with built-in EMG and Inertial Measurement Unit (IMU) sensors for sign language emotion recognition. The classification was carried out by a Graph Convolutional Network (GCN) that employed a U-shaped distribution model to better predict emotions.

The main drawback of using only physiological data is the need for invasive and costly measurement equipment. In addition, instead of recognizing a specific emotion, the mentioned approaches focused on predicting the pleasantness and intensity of the expressed emotional state.

### 2.2.2. Multimodal data

Another option to accomplish emotion recognition is to merge multiple signals. The combination of different data sources allows to overcome the weaknesses of the single modality, thus realizing a more robust emotion recognition system. Physiological signals have been used together with audio/visual data and behavioral cues. In [30], the authors proposed a generative approach using Restricted Boltzmann Machines (RBMs), modeling both EEG and eye tracking features. In addition, temporal features of the emotion evolution have been exploited in [33,34]. More specifically, recurrent layers with Convolutional Neural Network (CNN) have been employed to capture both spatial and temporal characteristics of multimodal data.

Since the acquisition of physiological signals results in the usage of invasive equipment and reduced freedom of movement, several state-of-the-art approaches disregarded this type of cues.

For instance, Shi et al. proposed to extract audio signals and transcriptions and to combine them with 2D human gestures. The result of the combination was provided as input of an attention-based model [37]. In [36] an emotion recognition system was developed based on facial expressions and speech features. Specifically, the combination of CNNs and recurrent layers on video data with a weighted decision fusion achieved enhanced performance compared to unimodal recognition methods.

Moreover, in [35], the authors proposed a hybrid deep learning architecture that exploits audio-visual data for extracting text, audio, facial expression, and other features (such as age, gender, and culture). In addition, they analyzed the surrounding scene to provide context to the emotion classifier. Zhang et al. [32] proposed a multiscale emotion recognition model from speech excerpts involving a CNN and a Recurrent Neural Network (RNN) on Mel spectrogram. In [38], a deep learning approach that exploits facial expressions and body movements has been proposed. The authors adopted the OpenPose [46] markerless motion analysis to extract the skeleton joints and analyze the user's movements.

The multimodal approach, however, has some disadvantages. First, when multiple sensors are required, the cost and complexity of the overall framework increases. In addition, approaches employing facial expressions and speech information are not HMD-compliant and may raise privacy issues. In addition, although facial information is an effective cue for emotion recognition, boundary conditions (such as illumination, distance, or partial occlusions) may preclude its identification [14].

### 2.2.3. Body language

As highlighted in [14], although body movements can be useful for understanding the fundamental mechanisms underlying emotion processing, only few studies on emotion recognition through body language have been performed.

In order to recognize emotions through body movements, one possible approach is to encode body language using features. For instance, in [47] the Laban Movement Analysis (LMA) is employed to model body language. LMA identifies four components of body movements: body, effort, shape, and space. The body component encodes physical body features such as the distance between joints, their connection, and inter-dependencies. The effort component captures the movement dynamics such as velocity and acceleration. The shape component represents static body features and spatial occupancy, while the space component describes the movement of the body in relation to the surrounding environment.

In a similar direction, in [12], a study on the relation between emotions and motion features has been performed, and a Machine Learning (ML)-based model has been employed for emotion classification. In addition, Zhang et al. proposed to exploit the temporal

characteristics of the 3D skeleton joints by applying RNNs [10]. In [44] an on-purpose CNN is designed to model 3D-position coordinates as images to perform emotion recognition, introducing "insecurity" in the set of target emotions. Toward this direction, in [39] the authors proposed an approach that captures both coarse-grained and fine-grained affective information from full-body motion. In more detail, they employ a multiscale spatio-temporal model to decode the complex mapping between emotions and body movements, exploiting a FCN on multi-scale temporal and spatial features. In [40] a combination of spatial and temporal modeling through GCNs of 3D skeleton coordinates. Then, the authors proposed an Interframe Shift Encode (ISE) module to enhance feature representation between time frames. In [42] a GCN was employed to combine 3D joint positions with hand-crafted temporal and spatial features to predict the emotion from an RGB video. Similarly, in [41] an emotion recognition model with two GCNs was developed to process 3D skeletons. In more detail, spatial–temporal convolutions were employed to model both postures and movements of users. In this direction, in [43] the Joint Neighborhood Distance (JND) was introduced in order to analyze joints and their immediate neighbors with Deep Neural Network (DNN).

All the aforementioned approaches process the 3D skeleton joints either exploiting depth cameras, marker-based acquisition systems, or employing computationally expensive DNN for 3D pose estimation.

Differently from the state-of-the-art, in this work, we propose a markerless emotion recognition system based on the analysis of the spatio-temporal evolution of 2D skeleton joints extracted from videos acquired by a commercial RGB camera. Being markerless, the system is not invasive and does not hinder the natural user movements. In addition, the proposed approach is completely HMD-compliant. Furthermore, the acquisition equipment is low-cost and lightweight and, therefore, can be employed in a wide set of application scenarios.

In the same direction, Ghaleb et al. proposed an emotion recognition system that extracts 2D joint coordinates from RGB videos and uses a GCN to predict the emotions [45]. With respect to [45] we significantly reduce the number of parameters to learn, thus proposing a more lightweight approach. An open-source project including a pre-processed dataset, code, and demo with pre-trained models is available at https://github.com/michaelneri/emotion-recognition-human-movements.

### 2.3. Emotion modeling

The typical approach for emotion recognition consists in employing the categorical model which splits the emotional sphere into six basic emotions: happiness, sadness, fear, anger, disgust, and surprise [48]. However, this model can only partially describe the affective states a person may express [49,50]. In [51] the authors investigated the relative importance of basic and non-basic emotions during interactions with computer interfaces showing that non-basic emotions such as boredom and engagement, were significantly prevalent with respect to the basic ones. The study highlighted that the focus on the six standard emotions allows to only partially represent the spectrum of human affective states. Moreover, as pointed out in [50], the affective states of interest, calm, and boredom, can be mapped to proto-social responses of humans.

As an alternative, dimensional emotion models have been proposed. In this case, emotions are categorized based on different features. A well-established model is the Geneva Emotion Wheel (GEW) introduced for the first time in [52]. It represents the emotional space with three dimensions: valence, i.e., positive versus negative evaluation of the emotion, control, i.e., the amount of power to exert control on the emotion, and arousal, i.e., the intensity of the emotion. However, since it is difficult to reproduce the emotion intensity, the GEW authors decided not to associate separate labels to different intensities, but to indicate only the emotion family as a whole [52].

Due to the mentioned inadequacy of basic emotions in characterizing the interaction through computer interfaces [51], in this work we selected the GEW dimensional approach as emotion model.

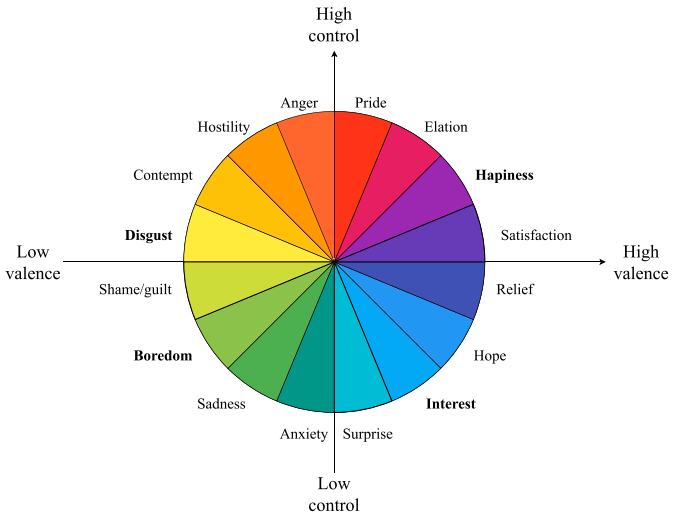



**Fig. 1.** Employed emotion dimensional model. Selected emotions are shown in bold.

## 3. Experimental settings

In this section, the set of considered emotions are presented and the extended version of HEROES [16] dataset is described with the adopted augmentation procedure and the related statistics.

### 3.1. Selected emotion model

To test the proposed emotion recognition system, we selected the GEW emotion model and considered four emotions that span the defined dimensional model. More in detail, we selected interest (characterized by high valence, low control), happiness (associated with high valence, high control), disgust (representing low valence, high control), and boredom (linked to low valence, low control). We report the emotion dimensional model in Fig. 1.

The selected emotions influence body language in distinct ways [16]. For instance, when a person is happy, the body is often in an upright position with open arms and open or parallel legs. In addition, their movements are fast and expansive. In contrast, when experiencing disgust, people often shift their trunk orientation from side to side, while keeping their hands close to the body (e.g., covering the mouth or neck). Interest is usually associated with slow forward movements, with the body fully oriented toward the object of interest. Finally, boredom is usually characterized by a bowing posture, slow side to side head rotations, and slow non-expansive movements. Based on these considerations, body movements encode the considered set of emotions in a meaningful way, making them suitable for automatic recognition.

It is useful to notice that the adopted emotion set has been selected to test all the four categories of the GEW while considering emotions that could be experienced during both engaging and annoying social XR interactions. However, the proposed emotion recognition system can be applied to other scenarios changing the emotion set and the corresponding dataset.

### 3.2. Dataset: Extended-HEROES

The original HEROES dataset [16] contains 256 videos of 16 non-professional actors expressing four emotions: happiness, disgust, interest, and boredom. This dataset is suitable for on-the-wild emotion recognition since it has the following features:

- there is no time synchronization among cameras;
- the cameras are not calibrated;
- the acquisitions have been performed with varying illumination conditions;

**Table 2**
Extended HEROES statistics.

| | | | | |
|---|---|---|---|---|
| **General Statistics** | | | | |
| Number of folds | 5 | | | |
| Number of people | 16 | | | |
| Number of total videos | 2676 (892 per camera) | | | |
| **Emotions Statistics** | | | | |
| | **Happiness** | **Boredom** | **Interest** | **Disgust** |
| Number of videos | 621 | 662 | 746 | 647 |
| Max number of frames | 224 | 249 | 246 | 226 |
| Min number of frames | 44 | 68 | 68 | 89 |
| Mean number of frames | 132.26 | 153.66 | 159.75 | 151.42 |
| Standard deviation of number of frames | 33.35 | 35.32 | 40.32 | 36.62 |

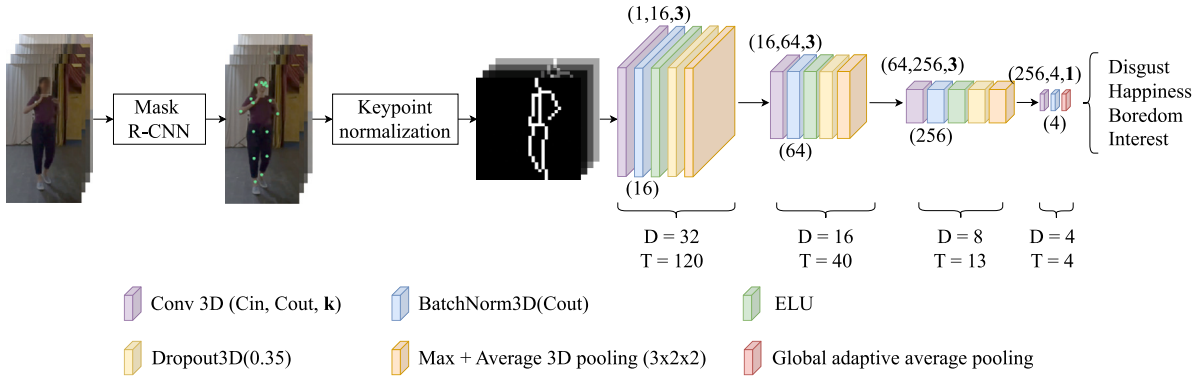



**Fig. 2.** The proposed method.

- the acquisition is markerless;
- the acquisition hardware is made of commercial, low-cost, and lightweight RGB cameras.

Dataset acquisition has been performed by providing the actors a description of two scenarios: a museum context and a conversational situation. The participants acted each emotion five times for each scenario. The choice of a scenario-based dataset collection has been performed according to [17,53]. Actors were asked to behave as they would in real life situations. Each actor was free to select the starting emotion and to move within the available space to express the emotion spontaneously, without any time constraint. The only behavioral instruction provided to the actors concerned the starting position, and the request of starting and concluding the performance with a natural pose (i.e., without signs of muscular activation).

Although the videos have been recorded through multiple cameras, in [16] only the central view has been exploited. In this work, the videos recorded from 3 points of view are considered. The three sets of videos have been synchronized in post-processing through a correlation-based audio alignment procedure. Each video recorded by the central camera is associated with an emotion label. Thanks to the synchronization, the labels are automatically propagated to the corresponding videos acquired by the lateral cameras.

During the labeling process, segments of each video containing hand gestures and head, torso, and foot movements that identify emotional states have been extracted. To this aim, the guidelines provided in [54] concerning the body movements associated to each emotion have been followed. The extracted video segments last between 2 and 10 s and the size of the Extended-HEROES dataset is approximately ten times larger than its initial release, thus allowing the usage of learning-based algorithms. The features of the Extended-HEROES dataset are shown in Table 2. It is worth noticing that the dataset is slightly unbalanced in the number of videos for each emotion (i.e., 621 videos for happiness, 662 videos for boredom, 746 videos for interest, and 647 for disgust). This characteristic is common to emotion recognition datasets in literature and it does not affect the performance assessment of the proposed emotion recognition system.

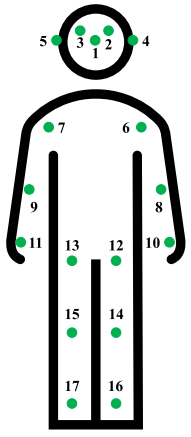
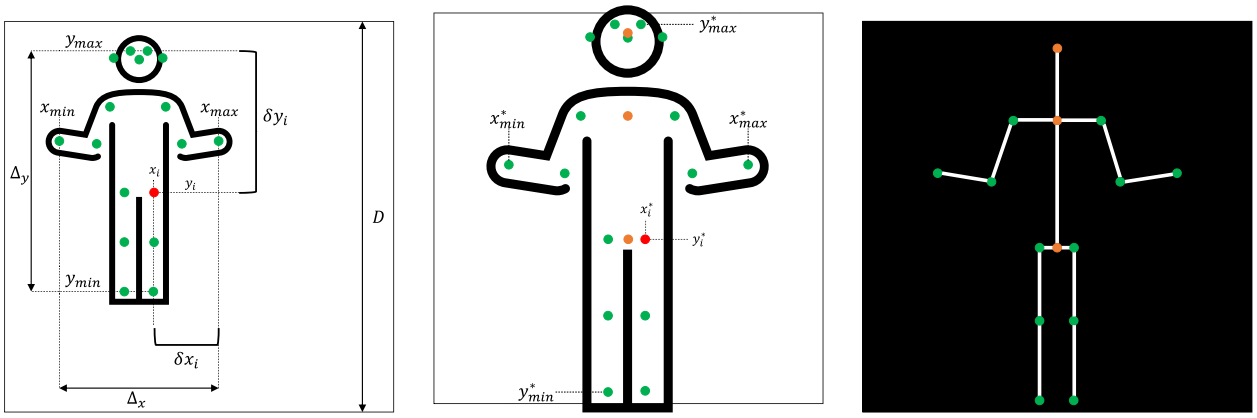



**Fig. 3.** Extracted skeleton joints.

## 4. Proposed approach

Our emotion recognition system takes as input a sequence of $T$ frames which are processed using the state-of-the-art Mask R-CNN [55] to extract skeleton keypoints. Then, the keypoints are normalized as detailed in Section 4.2 to obtain the proposed skeleton representation with fixed dimensions ($D \times D$). The resulting set of $T$ masks is provided as input to a 3D FCN as described in Section 4.3. The proposed network architecture is composed of four blocks. The first three blocks perform 3D convolutions and batch normalization, followed by the activation function, dropout and pooling. The last block extracts a vector that contains the logits associated with the targeted emotions. An overview of the proposed approach is shown in Fig. 2 and its components are described in details in the following.

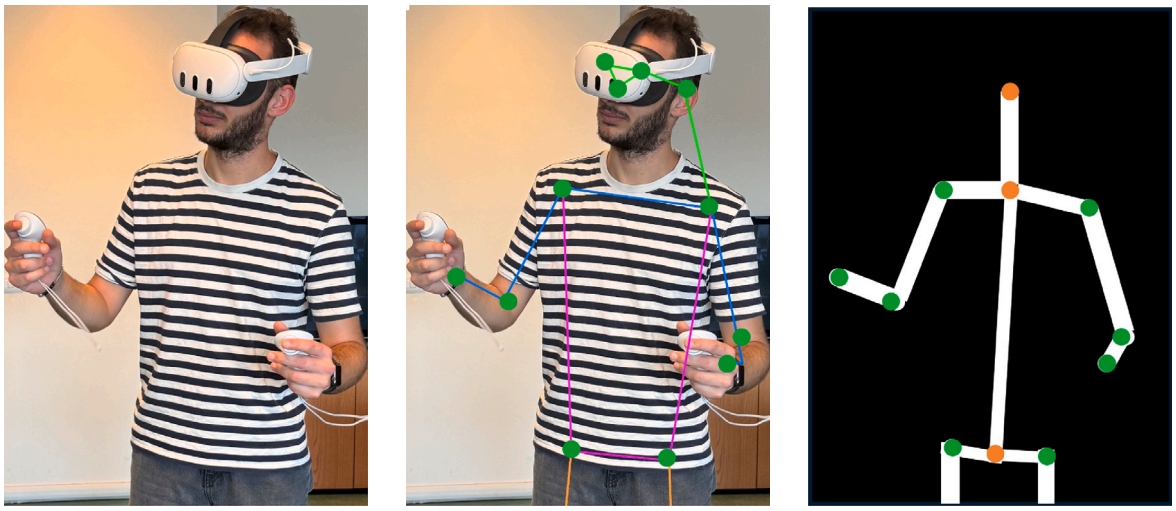



(a) Keypoints before normalization.

(b) Result of the normalization. Orange dots represent the virtual keypoints.

(c) Output skeleton mask.

**Fig. 4.** Mask construction process.



(a) Input blurred face.

(b) Estimated facial keypoints from Mask R-CNN.

(c) Output upper body skeleton mask.

**Fig. 5.** Construction of the upper body and facial components of the skeleton representation. Green bullets denote the joints estimated by the Mask R-CNN whereas the orange bullets correspond to virtual keypoints.

### 4.1. Keypoints extraction

As a first step, Mask R-CNN [55] has been employed to extract the skeleton keypoints from each frame. The algorithm provides as output 17 keypoints representing the skeleton joints. In more detail, the keypoints extracted by the original algorithm are: nose, left and right eye, left and right ear, left and right shoulder, left and right elbow, left and right wrist, left and right hip, left and right knee, left and right ankle. A visual representation of the extracted joints is provided in Fig. 3. It is useful to notice that Mask R-CNN is able to estimate keypoints in presence of occlusions thanks to its multi-stage architecture, RoI alignment, keypoint confidence scores, and the use of keypoint heatmaps [55]. Therefore, even in presence of an HMD, approximate facial keypoints will be provided. The proposed method exploits these rough keypoints to estimate the overall head position, so that the exact coordinates of each facial element is not needed.

### 4.2. Skeleton mask generation

Given the keypoints extracted by Mask R-CNN [55], we define a skeleton mask where the value is set to 1 for all points belonging to the skeleton, and 0 elsewhere. However, the keypoint coordinates provided by the Mask R-CNN belong to the image plane, as shown in Fig. 4(a). For this reason, the same physical displacement or movement may appear smaller or larger in pixel space depending on the user's distance from the camera. To address this, we introduce a normalization procedure that does not distort the skeleton shape while providing a depth-independent representation. The normalization process consists of re-scaling the skeleton so that the maximum body extension in pixels is set to the mask dimension $D$. Two scenarios should be handled:

(a) the largest body extension occurs along the vertical axis, $y$;
(b) the largest body extension occurs along the horizontal axis, $x$.

In scenario (a) the maximum extension before normalization, $\Delta$, is set to

$$\Delta = \Delta_y = y_{max} - y_{min}, \tag{1}$$

where $y_{max}$ and $y_{min}$ are the maximum and minimum vertical coordinates corresponding to a keypoint. Before normalization, the extension along the horizontal and vertical axes corresponding to the $i$th keypoint

located in $(x_i, y_i)$ are:

$$\begin{cases} \delta y_i = y_i - y_{min}, \\ \delta x_i = x_i - x_{min}. \end{cases} \tag{2}$$

Therefore, the following relations hold for the new coordinates of the $i$th keypoint:

$$\begin{cases} y_i^* = \frac{\delta y_i \times D}{\Delta}, \\ x_i^* = \frac{\delta x_i \times D}{\Delta} + D/2. \end{cases} \tag{3}$$

It is useful to notice that $x_i^*$ has been horizontally shifted by $D/2$ in order to center the skeleton with respect to the mask. When case (b) occurs, the same procedure is applied but $\Delta$ is set to $\Delta_x = x_{max} - x_{min}$, and the $y_i^*$ coordinate undergoes the shift. It is important to highlight that this procedure also allows to represent on the same scale users with very different physical characteristics such as the height.

After the keypoints have been normalized, three virtual keypoints are defined: head, breastbone, and pubic region. The head keypoint is obtained by averaging the keypoints belonging to the head (IDs 1, 2, 3, 4, and 5). In this way, the only information concerning the human face is a point whose position varies depending on the head orientation. The other two virtual keypoints are defined for reconstructing the skeleton shape. In more details, the breastbone keypoint is given by averaging the shoulder joints (IDs 6 and 7), and the pubic region keypoint is obtained as the mean of the pelvis joints (IDs 12 and 13). An example of all the keypoints after normalization is shown in Fig. 4(b).

After being scaled, the keypoints are arranged in a squared binary mask in which the considered joints are connected to obtain the skeleton. Real and virtual keypoints are connected as follows:

- the head is linked to the breastbone;
- the breastbone is linked to the shoulders and to the pubic region;
- the shoulders are linked to the elbows;
- the elbows are linked to the wrists;
- the pubic region is linked to the hips;
- the hips are linked to the knees;
- the knees are linked to the ankles.

All the pixels that do not belong to the skeleton are set to zero, as depicted in Fig. 4(c).

The overall mask construction procedure is depicted in Fig. 4, and the pseudocode of the process is reported in Algorithm 1.

It is useful to underline that the keypoint extraction procedure is fully HMD-compliant thanks to Mask R-CNN's ability to estimate the keypoints in case of occlusions. A proof of this assertion is provided in Fig. 5, where the skeleton joints are extracted for a person wearing an HMD. Moreover, since facial keypoints are not processed directly, but are employed to define the virtual head keypoint, the proposed model is robust against inaccurate facial keypoint localization.

**Algorithm 1** Skeleton mask construction for a single frame.

**Input:** Mask dimension $D$, keypoint coordinates
**Output:** $M$ binary skeleton mask

```
M ← squared zero matrix of dimensions D
x_min ← minimum keypoint coordinate along the horizontal direction
y_min ← minimum keypoint coordinate along the vertical direction
x_max ← maximum keypoint coordinate along the horizontal direction
y_max ← maximum keypoint coordinate along the vertical direction
Δ_x = x_max − x_min                          ▷ maximum extension along x axis
Δ_y = y_max − y_min                          ▷ maximum extension along y axis
if Δ_y > Δ_x then                            ▷ Scenario (a)
    Δ ← Δ_y
    for i ← keypoint_index in keypoint_indices do
        δx_i ← x_i − x_min
        δy_i ← y_i − y_min
        x_i^* ← δy_i × D / Δ + D/2           ▷ Shift to the center
        y_i^* ← δy_i × D / Δ
        M(x_i^*, y_i^*) ← 1
    end for
else                                         ▷ Scenario (b)
    Δ ← Δ_x
    for i ← keypoint_index in keypoint_indices do
        δx_i ← x_i − x_min
        δy_i ← y_i − y_min
        x_i^* ← δx_i × D / Δ
        y_i^* ← δy_i × D / Δ + D/2           ▷ Shift to the center
        M(x_i^*, y_i^*) ← 1
    end for
end if
h_x, h_y ← coordinates of head joint
M(h_x, h_y) ← 1
b_x, b_y ← coordinates of breastbone joint
M(b_x, b_y) ← 1
p_x, p_y ← coordinates of pubic region joint
M(p_x, p_y) ← 1
Set to 1 the pixels in M belonging to the skeleton links
```

### 4.3. Proposed emotion classifier

Let $X \in \mathbb{R}^{1\times T\times D\times D}$ be the skeleton representation obtained from Algorithm 1 for a video clip lasting $T$ frames. This tensor is fed to a 3D FCN $f_{FCN} : \mathbb{R}^{1\times T\times D\times D} \to \mathbb{R}^C$ which provides as output the emotion $\hat{\mathbf{y}}$ associated to one among $C$ classes. The 3D convolutions allow processing the spatiotemporal evolution of the skeleton while having a reduced computational complexity with respect to recurrent and attention-based network architectures. In this way, the system is lightweight and can be easily deployed on mobile devices.

The neural network is composed of 3 sequential convolutional blocks. In more details, each block consists of:

- a Conv3D($C_{in}, C_{out}, \mathbf{k}$) layer where $C_{in}$ and $C_{out}$ are the number of input and output channels (or filters), respectively, and $\mathbf{k} = [k_t, k_x, k_y]$ is the vector containing the kernel sizes for each dimension. This module is responsible for learning 3D filters that are able to track the spatio-temporal dynamics of the keypoints;
- a BatchNorm3D($C$) regularization layer that allows to normalize the output of each convolution operation;
- an ELU [56] activation function that performs the identity operation on positive inputs and applies an exponential non-linearity on negative inputs;
- a 3D Pooling($k_t \times k_x \times k_y$) layer which combines Max and Average pooling operations to reduce the computational complexity through the model.

Finally, a Global Average Pooling (GAP) layer is used as an aggregator for the 3D feature maps, obtaining the vector $\hat{\mathbf{y}}$ that contains the logits of each emotion.

The overall architecture of the 3D FCN is reported in Table 3.

### 4.4. Augmentation strategies

Deep neural networks rely on large-scale datasets to avoid overfitting. This phenomenon occurs when a model learns a function that perfectly fits the training set, providing bad generalization capabilities on unknown data. To tackle this issue, data augmentation strategies are usually employed for extracting additional information from the training dataset, increasing its size [57]. Hence, two augmentation strategies have been employed in the proposed architecture:

**Dropout:** a portion of $C_{out}$ channels of the 3D feature maps provided by the convolutional layers is randomly set to zero. In more details, a Dropout3D($p$) layer is applied, where $p$ is a

**Table 3**
Description of the proposed 3D FCN.

| |
|---|
| **Input**: Skeleton representation $X \in \mathbb{R}^{1\times T\times D\times D}$ |
| Conv3D(1, 16, 3, padding = same)<br>BatchNorm3D(16)<br>ELU<br>Dropout3D(0.35)<br>Max + Average 3D Pooling $(3\times 2\times 2)$ |
| Conv3D(16, 64, 3, padding = same)<br>BatchNorm3D(64)<br>ELU<br>Dropout3D(0.35)<br>Max + Average 3D Pooling $(3\times 2\times 2)$ |
| Conv3D(64, 256, 3, padding = same)<br>BatchNorm3D(256)<br>ELU<br>Dropout3D(0.35)<br>Max + Average 3D Pooling $(3\times 2\times 2)$ |
| Conv3D(256, $C$, 1, padding = same)<br>BatchNorm3D($C$) |
| Global Adaptive Average Pooling |
| **Output**: $\hat{\mathbf{y}}$ Predicted emotion $\in \mathbb{R}^C$ |

scalar that determines the dropout Bernoulli distribution. By using this module, the model tends to generate uncorrelated channels [58]. Moreover, this augmentation strategy can mimic the presence of occlusions or missed detections of the skeleton joints;

**Flip:** a horizontal flip is performed on the skeleton representation. This allows the model to learn that mirrored body postures and movements can be associated with the same emotion.

### 4.5. Loss function

The loss function used for training the model is the weighted cross entropy which minimizes the statistical difference between model prediction and ground truth. For each mini-batch with size $B$ the loss is computed as:

$$\mathcal{L}_{CE}(\hat{\mathbf{Y}}, \mathbf{y}) = -\mathbb{E}_B \left[ \sum_{c=1}^{C} w_c \log \frac{\exp \hat{Y}_c}{\exp \sum_{i=1}^{C} \hat{Y}_i} y_c \right], \tag{4}$$

where $\hat{\mathbf{Y}} \in \mathbb{R}^{B\times C}$ is the matrix whose rows contain the logits corresponding to the emotions for each video in the batch, and $\mathbf{y} \in \{0, \dots, C-1\}^B$ is the vector of the ground truth labels. In addition, $\mathbf{w} = [w_I, w_H, w_B, w_D]$ is the vector of class weights for Interest, Happiness, Boredom, and Disgust respectively. Weights can be set to account for differences in the complexity of the emotion recognition task for various emotional states.

## 5. Experimental validation

In this section, the performance assessment of the proposed emotion recognition system is provided. In more details, the metrics employed for the evaluation are described, and the experimental results are presented and discussed. In our implementation, the class weights of the loss function have been set to $\mathbf{w} = [1.0, 1.0, 1.0, 2.0]$. The weight corresponding to the disgust class, $w_D$, has been set higher with respect to the others since disgust is the most difficult emotion to be classified [59]. Thus, when disgust video clips are misclassified, the model is heavily penalized. In addition, the binary array size and the number of frames for each video have been set to $D = 32$ and $T = 120$, respectively. These parameters have been selected by a random-search optimization [60]. As a comparison, we employ two versions of ST-GCN-like [45] backbone which process 2D skeletons joints using GCNs. The rationale for using two versions of ST-GCN is that the original implementation contains 902k learnable parameters, which are approximately double with respect to our model (472k). To ensure a fairer comparison, we reduced the number of filters in the last block of the original ST-GCN architecture (from 256 to 128), decreasing the number of learnable parameters from 902k to 425k. We denote the original version as *O-ST-GCN* and we indicate the one with reduced parameters as *R-ST-GCN*.

### 5.1. Metrics

The performance of the proposed architecture has been evaluated by means of the accuracy metric. More specifically, let $TP$ be the number of true positives and $FP$ be the number of false positives, then the classification accuracy (imposing that the model always predicts a single label for each video) is evaluated as:

$$Acc = \frac{TP}{TP + FP}, \tag{5}$$

where a true positive occurs when the most probable predicted emotion for a video is the same as the ground truth, and a false positive corresponds to a misclassification. Moreover, the confusion matrix $C_F \in \mathbb{R}^{C\times C}$ has been employed since it provides information about how false positives are distributed.

To assess the performance of all approaches across all possible thresholds, we also compute the macro-average Area Under the Curve (AUC) of the Receiving-Operating Characteristic (ROC). Finally, 5 independent runs have been performed to compute the 95% mean confidence interval for both accuracy and AUC.

### 5.2. Results

To evaluate the impact of camera configuration on the emotion recognition task, five different setups have been defined as detailed in the following:

**Frontal camera:** only videos acquired by the frontal camera are used.

**Right camera:** only videos acquired by the right camera are used.

**Left camera:** only videos acquired by the left camera are used.

**Random cameras:** videos acquired by the three cameras are randomly provided as input to the classifier one at a time, ignoring the a-priori knowledge of the camera position;

**Camera fusion:** videos acquired by the three cameras are fed in parallel to three classifiers which are jointly optimized based on the fusion of the outputs of different branches.

The first three configurations were designed to evaluate whether one of the camera positions is more suitable for emotion recognition. Then, the impact of combining the different views was evaluated based on the fourth and fifth tests. In the random cameras setup the fusion is performed at data-level and the architecture design is the same as the one employed in the single-camera setup. Finally, the camera fusion setup has been introduced to exploit the a-priori knowledge on the camera position (i.e., central, right or left). To this aim, three FCNs, having the same structure as depicted in Table 3, have been employed. Each FCN focuses only on a single view and their output are fused after classification. More specifically, let $f_{FCN\,i} : \mathbb{R}^{1\times T\times D\times D} \to \mathbb{R}^C$ be the FCN that maps the skeleton representation to the predicted emotion $\hat{\mathbf{y}}_\mathbf{i}$ for the $i$th camera. To improve the generalization capabilities of the model, a soft-voting ensemble of the three FCNs has been employed for each video, that is:

$$\hat{\mathbf{y}} = \mathbb{E}\left[ \sum_{i=1}^{3} \hat{\mathbf{y}}_\mathbf{i} \right]. \tag{6}$$

**Table 4**
Accuracy with confidence intervals per fold for each model and scenario.

| Scenario | Model | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 |
|---|---|---|---|---|---|---|
| Frontal | O-ST-GCN | **0.671 ± 0.022** | **0.646 ± 0.012** | 0.656 ± 0.009 | 0.687 ± 0.008 | 0.665 ± 0.015 |
| | R-ST-GCN | 0.669 ± 0.018 | 0.643 ± 0.022 | **0.664 ± 0.012** | **0.691 ± 0.011** | 0.666 ± 0.015 |
| | Proposed | 0.659 ± 0.033 | 0.615 ± 0.036 | 0.638 ± 0.035 | 0.662 ± 0.037 | **0.691 ± 0.034** |
| Right | O-ST-GCN | 0.638 ± 0.027 | 0.631 ± 0.013 | **0.636 ± 0.013** | 0.622 ± 0.009 | 0.616 ± 0.010 |
| | R-ST-GCN | **0.647 ± 0.023** | **0.650 ± 0.015** | **0.636 ± 0.020** | 0.640 ± 0.011 | **0.625 ± 0.013** |
| | Proposed | 0.640 ± 0.020 | **0.650 ± 0.032** | **0.636 ± 0.061** | **0.663 ± 0.047** | 0.610 ± 0.044 |
| Left | O-ST-GCN | 0.586 ± 0.031 | 0.531 ± 0.015 | 0.601 ± 0.004 | 0.609 ± 0.018 | 0.595 ± 0.016 |
| | R-ST-GCN | 0.583 ± 0.024 | 0.546 ± 0.015 | 0.589 ± 0.010 | 0.609 ± 0.008 | 0.573 ± 0.011 |
| | Proposed | **0.605 ± 0.023** | **0.578 ± 0.050** | **0.633 ± 0.018** | **0.614 ± 0.030** | **0.616 ± 0.045** |
| Random | O-ST-GCN | 0.635 ± 0.015 | 0.613 ± 0.010 | 0.634 ± 0.016 | 0.647 ± 0.002 | 0.650 ± 0.015 |
| | R-ST-GCN | 0.646 ± 0.016 | 0.618 ± 0.006 | 0.644 ± 0.016 | 0.659 ± 0.012 | 0.653 ± 0.014 |
| | Proposed | **0.677 ± 0.010** | **0.669 ± 0.015** | **0.669 ± 0.023** | **0.690 ± 0.028** | **0.671 ± 0.013** |
| Fusion | O-ST-GCN | 0.688 ± 0.031 | 0.654 ± 0.009 | 0.680 ± 0.009 | 0.710 ± 0.012 | 0.667 ± 0.005 |
| | R-ST-GCN | 0.695 ± 0.029 | 0.659 ± 0.011 | 0.685 ± 0.009 | 0.719 ± 0.004 | 0.682 ± 0.019 |
| | Proposed | **0.734 ± 0.036** | **0.693 ± 0.020** | **0.702 ± 0.017** | **0.754 ± 0.012** | **0.742 ± 0.019** |

**Table 5**
AUC with confidence intervals per fold for each model and scenario.

| Scenario | Model | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 |
|---|---|---|---|---|---|---|
| Frontal | O-ST-GCN | 0.824 ± 0.019 | **0.810 ± 0.017** | 0.818 ± 0.013 | 0.808 ± 0.016 | 0.828 ± 0.022 |
| | R-ST-GCN | 0.819 ± 0.016 | 0.801 ± 0.010 | 0.821 ± 0.020 | 0.805 ± 0.012 | 0.827 ± 0.011 |
| | Proposed | **0.833 ± 0.022** | 0.807 ± 0.030 | **0.825 ± 0.022** | **0.820 ± 0.031** | **0.850 ± 0.022** |
| Right | O-ST-GCN | 0.804 ± 0.015 | 0.789 ± 0.012 | 0.800 ± 0.014 | 0.796 ± 0.011 | 0.807 ± 0.016 |
| | R-ST-GCN | 0.808 ± 0.011 | 0.803 ± 0.016 | 0.798 ± 0.020 | 0.806 ± 0.008 | **0.809 ± 0.011** |
| | Proposed | **0.809 ± 0.017** | **0.816 ± 0.030** | **0.799 ± 0.031** | **0.834 ± 0.037** | 0.792 ± 0.033 |
| Left | O-ST-GCN | 0.770 ± 0.021 | 0.737 ± 0.016 | 0.779 ± 0.010 | 0.785 ± 0.020 | 0.788 ± 0.007 |
| | R-ST-GCN | 0.773 ± 0.018 | 0.744 ± 0.013 | 0.777 ± 0.015 | **0.788 ± 0.016** | 0.786 ± 0.018 |
| | Proposed | **0.798 ± 0.019** | **0.783 ± 0.048** | **0.819 ± 0.017** | 0.782 ± 0.036 | **0.816 ± 0.019** |
| Random | O-ST-GCN | 0.820 ± 0.008 | 0.811 ± 0.006 | 0.813 ± 0.020 | 0.820 ± 0.008 | 0.829 ± 0.012 |
| | R-ST-GCN | 0.831 ± 0.011 | 0.816 ± 0.006 | 0.823 ± 0.012 | 0.835 ± 0.008 | 0.837 ± 0.011 |
| | Proposed | **0.861 ± 0.006** | **0.857 ± 0.006** | **0.857 ± 0.013** | **0.856 ± 0.014** | **0.868 ± 0.004** |
| Fusion | O-ST-GCN | 0.850 ± 0.026 | 0.823 ± 0.013 | 0.856 ± 0.003 | 0.841 ± 0.013 | 0.837 ± 0.024 |
| | R-ST-GCN | 0.859 ± 0.029 | 0.820 ± 0.009 | 0.865 ± 0.013 | 0.849 ± 0.016 | 0.860 ± 0.013 |
| | Proposed | **0.890 ± 0.022** | **0.868 ± 0.008** | **0.883 ± 0.009** | **0.879 ± 0.008** | **0.893 ± 0.008** |

Accordingly, the loss function has been adapted for optimizing each branch by applying the weighted cross-entropy loss $\mathcal{L}_{CE}$ in Eq. (4) to each FCN output. The final loss is computed as follows for each batch:

$$\mathcal{L}(\underbrace{\hat{\mathbf{Y}}_1, \hat{\mathbf{Y}}_2, \hat{\mathbf{Y}}_3, \hat{\mathbf{Y}}}_{\text{Predictions}}, \mathbf{y}) = \sum_{i=1}^{3} \mathcal{L}_{CE}(\hat{\mathbf{Y}}_i, \mathbf{y}) + \mathcal{L}_{CE}(\hat{\mathbf{Y}}, \mathbf{y}), \tag{7}$$

where $\mathcal{L}_{CE}(\hat{\mathbf{Y}}_i, \mathbf{y})$ optimizes each branch separately, whereas $\mathcal{L}_{CE}(\hat{\mathbf{Y}}, \mathbf{y})$ accounts for the agreement between the FCNs.

For each setup, the performance has been evaluated as follows:

- faces have been blurred in each video to simulate the absence of the facial information as it occurs when wearing an HMD. In more details, a median filter has been applied to faces with kernel size of 35;
- the extraction of the skeleton keypoints has been carried out and the normalized mask has been computed.
- the neural networks have been trained and tested for each setup.
- The dataset has been randomly split into 5 folds for cross-validation, ensuring an even distribution of labels.

Moreover, the Extended-HEROES dataset has been split in five folds. By doing so, the model has been evaluated by feeding non-overlapping training and validation sets, thus measuring its generalization capabilities. The results are shown in Tables 4, 5, and 6, and in Fig. 6.

Specifically, Tables 4 and 5 report the accuracy and AUC, respectively, with confidence intervals across the five folds for O-ST-GCN, R-ST-GCN, and our approach under different camera scenarios. Across most individual camera scenarios (Frontal, Right, and Left), the proposed model achieves comparable or slightly superior performance to ST-GCN-like models in several folds. In the Random scenario, where camera viewpoints are randomly varied, the proposed model consistently achieves highest accuracy and AUC across all folds with respect to ST-GCN models, suggesting superior generalization to unseen or mixed viewpoints. The same trend can be observed for the Fusion scenario, which achieves the best performance. The comparison of the three camera positions in Table 6 shows that the model trained on videos captured by the central camera outperforms those trained on lateral views (Left and Right) in terms of both accuracy and AUC. In fact, as underlined in [59], it is easier to classify an emotion from body movements when the user faces the camera. In addition, it is possible that lateral views may not be able to record significant dynamics of the human body due to self-occlusions.

**Table 6**
Average and confidence interval of accuracy and AUC for each model and scenario.

| Scenario | Model | Average | |
|---|---|---|---|
| | | Acc | AUC |
| Frontal | O-ST-GCN | **0.671 ± 0.022** | 0.823 ± 0.019 |
| | R-ST-GCN | 0.669 ± 0.018 | 0.819 ± 0.016 |
| | Proposed | 0.659 ± 0.032 | **0.833 ± 0.022** |
| Right | O-ST-GCN | 0.638 ± 0.027 | 0.804 ± 0.014 |
| | R-ST-GCN | **0.647 ± 0.023** | 0.808 ± 0.011 |
| | Proposed | 0.639 ± 0.020 | **0.809 ± 0.017** |
| Left | O-ST-GCN | 0.586 ± 0.031 | 0.770 ± 0.020 |
| | R-ST-GCN | 0.583 ± 0.024 | 0.773 ± 0.018 |
| | Proposed | **0.605 ± 0.023** | **0.798 ± 0.019** |
| Random | O-ST-GCN | 0.635 ± 0.015 | 0.820 ± 0.007 |
| | R-ST-GCN | 0.646 ± 0.016 | 0.831 ± 0.011 |
| | Proposed | **0.678 ± 0.010** | **0.861 ± 0.006** |
| Fusion | O-ST-GCN | 0.688 ± 0.031 | 0.850 ± 0.026 |
| | R-ST-GCN | 0.694 ± 0.029 | 0.859 ± 0.029 |
| | Proposed | **0.734 ± 0.036** | **0.890 ± 0.022** |

Additionally, Table 6 shows that the contribution of multiple viewpoints is significant as both the Random scenario and the Fusion

**Table 7**
Ablation study.

| | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 | Average |
|---|---|---|---|---|---|---|
| No dropout | 0.662 | **0.703** | 0.725 | 0.735 | 0.744 | 0.714 |
| No horizontal flipping | 0.672 | 0.703 | **0.741** | 0.677 | 0.750 | 0.708 |
| No augmentations | 0.692 | 0.680 | 0.710 | 0.700 | **0.750** | 0.706 |
| Camera fusion | **0.734** | 0.693 | 0.702 | **0.754** | 0.742 | **0.734** |

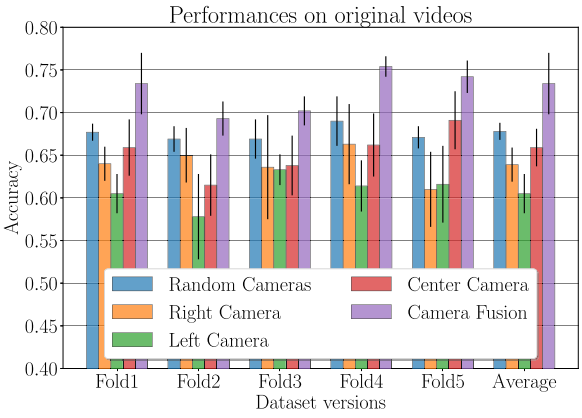



**Fig. 6.** Accuracy of the proposed model for each fold using different camera setups.


**Fig. 7.** Confusion matrix $C_F$ of the best model from Camera Fusion setup.

setup outperform the single-camera configurations. However, by exploiting the a-priori information on camera position, the Fusion scenario achieves the best results outperforming the frontal-view configuration by about 8.0% and the Random scenario by about 5% in terms of accuracy. Furthermore, the table underlines the superiority of the proposed approach with respect to ST-GCN-like architectures.

The confusion matrix (see Fig. 7) shows that the performance of the proposed approach is consistent with state-of-the-art findings, as the worst results are obtained for the disgust stimulus [59]. As shown in Fig. 7, disgust is generally recognized as boredom, predicting the correct emotion valence. Moreover, the confusion matrix highlights that there is a partial ambiguity between interest and boredom. This phenomenon is probably due to similarities between the body dynamics associated to the two emotional states, as reported in [16].

### 5.3. Ablation study

To evaluate the benefits of the adopted augmentation strategies presented in Section 4.4, the following tests have been performed on the camera fusion setup:

- only the Dropout3D(0.35) layer is employed;
- only the horizontal flipping is applied to the skeleton representation;
- no augmentation strategy is used.

The results are provided in Table 7. As expected, removing augmentations leads to a performance degradation (approximately 2.0%). Moreover, from the results, the horizontal flipping strategy impacts more than the dropout layer. This augmentation, in fact, increases the variability of the movements, helping the model to generalize the human body language.

## 6. Discussion

We compared the proposed method with state-of-the-art approaches for emotion classification that exploit body movements. We report in Table 8 the performance in terms of average accuracy and emotion-wise accuracy, considering both 3D and 2D skeleton representations. It is useful to underline that for multimodal methods, we reported the results achieved using body movements only.

As shown in Table 8, the average accuracy of the methods based on 3D skeletons is usually better than the proposed approach. This result can be due to several factors. First, the additional information provided by the third dimension represents a key contribution. In fact, for other tasks such as human action recognition, a comparison between the usage of 3D and 2D skeletons has been performed, and it has been demonstrated that the usage of 2D skeleton representations may impair the classification results [64].

This is confirmed by the comparison between the proposed approach and the method presented in [38,45] which, to the best of our knowledge, are the only papers that employ 2D skeleton representations. In this case, the proposed approach shows significantly higher average accuracies.

However, it is worth noting that 3D representations require advanced motion capture systems with markers or depth cameras for data acquisition, together with an increased computational burden for emotion prediction. On the contrary, the data collection process for 2D keypoints is less cumbersome and more cost-effective compared to 3D skeletal data [64]. Therefore, the proposed model's ability to operate effectively with readily available 2D data not only simplifies the data acquisition process but also extends the potential for deploying emotion recognition technologies across a multitude of real-life scenarios. This is particularly crucial for resource-constrained environments or applications where processing speed is a priority.

In addition, the use of a 2D representation of the skeleton significantly reduces the computational complexity associated with emotion recognition tasks compared with the adoption of a 3D skeleton [64]. This reduction in execution time is fundamental for real-world applications where real-time feedback is essential, such as interactive gaming or online education.

Moreover, it is useful to notice that most of the state-of-the-art approaches that show better performance with respect to the proposed method rely on a smaller dataset. This may imply a poor ability to generalize to larger data samples.

Finally, the average accuracy is affected by the emotion set considered for the experiments. To provide a fair comparison, we reported in Table 8 the emotion-wise performances for the selected emotions. Concerning happiness, the performance of the proposed approach is in between the 3D skeleton methods, and clearly outperforms the 2D

**Table 8**
Comparison with state-of-the-art approaches. Dash symbol means no results are available for the specific emotion.

| Method | Dataset | # of samples | # of classes | Average *Acc* | Happiness | Interest | Boredom | Disgust |
|---|---|---|---|---|---|---|---|---|
| **3D skeleton approaches** | | | | | | | | |
| Shi et al. [37] | IEMOCAP [15] | 5492 | 4 | 0.66 | 0.58 | – | – | – |
| Ahmed et al. [12] | own dataset | 30 | 5 | 0.87 | 0.83 | – | – | – |
| Zhang et al. [10] | EGBM [22] | 560 | 7 | 0.77 | 0.75 | – | – | 0.67 |
| Wang et al. [39] | EGBM [22] | 560 | 7 | **0.96** | **0.98** | – | – | **0.96** |
| Wang et al. [39] | KDAE [21] | 1402 | 7 | **0.96** | 0.95 | – | – | 0.95 |
| Wang et al. [39] | Emilya [19,20] | 8206 | 8 | 0.94 | 0.93 | – | – | – |
| Wang et al. [39] | MPI [23] | 1447 | 11 | 0.90 | 0.92 | – | – | 0.84 |
| Wang et al. [39] | DMCD [24] | 108 | 12 | 0.86 | 0.85 | – | 0.75 | – |
| Beyan et al. [44] | own dataset | 61 | 4 | 0.95 | 0.95 | – | – | – |
| Beyan et al. [44] | DMCD [24] | 108 | 12 | 0.75 | 0.73 | – | 0.74 | – |
| Ghaleb et al. [45] | KDAE [21] | 1402 | 7 | 0.65 | 0.75 | – | – | 0.58 |
| Yumeng et al. [42] | Emotion-Gait [61] | 2177 | 4 | 0.85 | 0.80 | – | – | – |
| Chen et al. [40] | Emotion-Gait [61] | 2177 | 4 | 0.88 | 0.90 | – | – | – |
| Zhai et al. [41] | Emotion-Gait [61] | 2177 | 4 | 0.85 | 0.80 | – | – | – |
| Zhai et al. [41] | ELBM [62] | 3924 | 4 | 0.86 | 0.93 | – | – | – |
| Oğuz et al. [43] | PhysioNet [21] | 1402 | 7 | 0.91 | 0.88 | – | – | 0.92 |
| **2D skeleton approaches** | | | | | | | | |
| Filntisis et al. [38] | GEMEP [17] | 145 | 12 | 0.34 | 0.45 | 0.26 | – | – |
| Ghaleb et al. [45] | Green Stimuli [63] | 871 | 7 | 0.69 | 0.63 | – | – | 0.55 |
| Proposed approach | extended HEROES [16] | 892 | 4 | 0.73 | 0.85 | **0.73** | **0.87** | 0.31 |

**Table 9**
Analysis on computational complexity in terms of learnable parameters and inference times.

| Methods | Learnable parameters | Inference time (ms) |
|---|---|---|
| O-ST-GCN [45] | 902k | $1.79 \pm 0.42$ |
| R-ST-GCN [45] | **425**k | $1.75 \pm 0.48$ |
| Proposed approach | 472k | $\mathbf{0.76} \pm \mathbf{0.39}$ |

state-of-the-art methods. As for interest, only [38] included it in the evaluation, and the proposed approach shows better performances. In addition, the presented model provides a better accuracy with respect to the 3D skeleton approaches which included boredom in the emotion set. Finally, disgust is the worst-performing emotion. The significantly higher performance of 3D approaches with respect to 2D techniques may indicate that, to properly reveal disgust from body movements, the additional information provided by the third dimension is relevant. This is not surprising since, when a person is disgusted, he/she tends to move away from the source of disgust. Therefore, the depth cue could be extremely important in distinguishing this emotion from the others. Moreover, the increased performance of [45] with respect to the proposed method can be attributed to the higher complexity (3M parameters with respect to 425k).

To demonstrate the low-complexity characteristic of our approach, we analyzed the execution time and provided the number of learnable parameters for each model in the experimental results section in Table 9. In our setup, the proposed architecture runs on a NVIDIA RTX 4070, a commercial-off-the-shelf GPU for gaming, with an average inference time of **0.76** ms per video.

To conclude, this work offers a less complex method for emotion recognition in terms of acquisition, storage, and computing resources, providing a good balance between efficiency and accuracy. By doing so, it expands the possibilities for using emotion recognition technologies in real-world scenarios. The accuracy gap with respect to 3D-based techniques opens new paths for future research targeting emotion recognition exploiting body language in computational-constrained devices.

## 7. Conclusions

In this work, an approach for emotion recognition based on body language for social XR applications has been presented. In more detail, the proposed method relies on the extraction of body joints and processes their spatio-temporal evolution through an on-purpose designed deep learning architecture. In addition, to obtain a distance-independent skeleton representation, a normalization algorithm has been introduced. For performance assessment, the HEROES dataset has been extended by including the videos acquired by multiple cameras. Extensive tests have been carried out for validation purposes and an average accuracy of 73.4% has been achieved. It is worth noticing that these results have been obtained in a markerless challenging scenario, with a non-invasive and HMD-compliant approach that exploits low-cost and lightweight commercial RGB cameras. Future contributions will address the inclusion of additional emotional states relevant in social XR applications field such as confusion and frustration. These emotions will allow both to evaluate the level of engagement of the user and to reveal his/her need for assistance. Finally, an in-depth comparative study between 2D and 3D skeleton representations for emotion recognition will be carried out in the future.

## CRediT authorship contribution statement

**Michael Neri:** Writing – original draft, Software, Methodology, Conceptualization. **Sara Baldoni:** Writing – original draft, Methodology, Conceptualization. **Marco Carli:** Supervision, Conceptualization. **Federica Battisti:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

## Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Sara Baldoni reports financial support was provided by European Union. Federica Battisti reports financial support was provided by European Union. Marco Carli reports financial support was provided by European Union. Michael Neri reports financial support was provided by European Union. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This work was partially supported by the European Union under the Italian National Recovery and Resilience Plan (NRRP) Mission 4, Component 2, Investment 1.3, CUP C93C22005250001, partnership on “Telecommunications of the Future” (PE00000001 - program “RESTART”).

## Data availability

The code and dataset are publicly available at https://github.com/michaelneri/emotion-recognition-human-movements.

## References


[1] M. Hewstone, W. Stroebe, Introduction to Social Psychology: A European perspective, Oxford: Blackwell, 2000.

[2] M. Karg, A. Samadani, R. Gorbet, K. Kühnlenz, J. Hoey, D. Kulić, Body movements for affective expression: a survey of automatic recognition and generation, IEEE Trans. Affect. Comput. 4 (2013) 341–359, http://dx.doi.org/10.1109/T-AFFC.2013.29.

[3] S. Xu, J. Fang, X. Hu, E. Ngai, W. Wang, Y. Guo, V.C.M. Leung, Emotion recognition from gait analyses: current research and future directions, IEEE Trans. Comput. Soc. Syst. 11 (2024) 363–377, http://dx.doi.org/10.1109/TCSS.2022.3223251.

[4] D. Checa, B. Rodriguez-Garcia, H. Guillen-Sanz, I. Miguel-Alonso, A framework for developing multi-user immersive virtual reality learning environments, in: L.T. De Paolis, P. Arpaia, M. Sacco (Eds.), Extended Reality, Springer Nature Switzerland, Cham, 2023, pp. 89–103.

[5] S. Kang, Investigating avatar facial expressions and collaboration dynamics for social presence in avatar-mediated xr remote communication, in: 2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), 2024, pp. 1114–1115, http://dx.doi.org/10.1109/VRW62533.2024.00351.

[6] C.F. Purps, S. Janzer, M. Wölfel, Reconstructing facial expressions of HMD users for avatars in VR, in: M. Wölfel, J. Bernhardt, S. Thiel (Eds.), ArtsIT, Interactivity and Game Creation, Springer International Publishing, Cham, 2022, pp. 61–76.

[7] D. Maloney, G. Freeman, D.Y. Wohn, "Talking without a voice": Understanding non-verbal communication in social virtual reality, Proc. ACM Hum.-Comput. Interact. 4 (2020) http://dx.doi.org/10.1145/3415246.

[8] J. Marí n Morales, C. Llinares, J. Guixeres, M. Alcañiz, Emotion recognition in immersive virtual reality: from statistics to affective computing, Sensors 20 (2020) 5163.

[9] Y. Matsuda, D. Fedotov, Y. Takahashi, Y. Arakawa, K. Yasumoto, W. Minker, Emotour: estimating emotion and satisfaction of users based on behavioral cues and audiovisual data, Sensors 18 (2018) 3978.

[10] H. Zhang, P. Yi, R. Liu, D. Zhou, Emotion recognition from body movements with as-lstm, in: IEEE 7th International Conference on Virtual Reality (ICVR), 2021a, http://dx.doi.org/10.1109/ICVR51878.2021.9483833.

[11] M. de Meijer, The contribution of general features of body movement to the attribution of emotion, J. Nonverbal Behav. 13 (1989) 247–268.

[12] F. Ahmed, A.S.M.H. Bari, M.L. Gavrilova, Emotion recognition from body movement, IEEE Access 8 (2020) 11761–11781.

[13] A. Pease, B. Pease, The Definitive Book of Body Language, Peace International, 2004.

[14] L.L. Lott, F.B. Spengler, T. Stächele, B. Schiller, M. Heinrichs, Embody/emface as a new open tool to assess emotion recognition from body and face expressions, Sci. Rep. 12 (2022) 1–13.

[15] C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, S.S. Narayanan, Iemocap: interactive emotional dyadic motion capture database, Language Resources and Evaluation 42 (2008) 335–359.

[16] I. Mannocchi, K. Lamichhane, M. Carli, F. Battisti, Heroes: a video-based human emotion recognition database, in: 10th European Workshop on Visual Information Processing (EUVIP), 2022, pp. 1–6, http://dx.doi.org/10.1109/EUVIP53989.2022.9922723.

[17] D. Glowinski, A. Camurri, G. Volpe, N. Dael, K. Scherer, Technique for automatic emotion recognition by body gesture analysis, in: Computer Vision and Pattern Recognition Workshops, 2008.

[18] Y. Ma, H.M. Paterson, F.E. Pollick, A motion capture library for the study of identity, gender, and emotion perception from biological motion, Behav. Res. Methods 38 (2006) 134–141.

[19] N. Fourati, C. Pelachaud, Emilya: emotional body expression in daily actions database, in: International Conference on Language Resources and Evaluation, 2014.

[20] N. Fourati, C. Pelachaud, Perception of emotions and body movement in the emilya database, IEEE Trans. Affect. Comput. 9 (2018) 90–101, http://dx.doi.org/10.1109/TAFFC.2016.2591039.

[21] M. Zhang, L. Yu, K. Zhang, B. Du, B. Zhan, S. Chen, X. Jiang, S. Guo, J. Zhao, Y. Wang, Kinematic dataset of actors expressing emotions, Sci. Data 7 (2020) 1–8.

[22] T. Sapiński, D. Kamińska, A. Pelikant, C. Ozcinar, E. Avots, G. Anbarjafari, Multimodal database of emotional speech, video and gestures, in: Pattern Recognition and Information Forensics: ICPR 2018 International Workshops, CVAUI, IWCF, and MIPPSNA, Springer, 2019.

[23] E. Volkova, S. De La Rosa, H.H. Bülthoff, B. Mohler, The mpi emotional body expressions database for narrative scenarios, PloS One 9 (2014) e113647.

[24] Dance Motion Capture Database (DMCD), https://dancedb.eu, (Accessed 31 October 2023).

[25] G. Castellano, S.D. Villalba, A. Camurri, Recognising human emotions from body movement and gesture dynamics, in: Affective Computing and Intelligent Interaction, 2007.

[26] C. Qing, R. Qiao, X. Xu, Y. Cheng, Interpretable emotion recognition using EEG signals, IEEE Access 7 (2019) 94160–94170, http://dx.doi.org/10.1109/ACCESS.2019.2928691.

[27] J. Li, S. Qiu, Y. Shen, C. Liu, H. He, Multisource transfer learning for cross-subject eeg emotion recognition, IEEE Trans. Cybern. 50 (2020) 3281–3293, http://dx.doi.org/10.1109/TCYB.2019.2904052.

[28] Y. Zhang, H. Liu, D. Zhang, X. Chen, T. Qin, Q. Zheng, Eeg-based emotion recognition with emotion localization via hierarchical self-attention, IEEE Trans. Affect. Comput. 14 (2023) 2458–2469, http://dx.doi.org/10.1109/TAFFC.2022.3145623.

[29] X. Zhang, J. Liu, J. Shen, S. Li, K. Hou, B. Hu, J. Gao, T. Zhang, B. Hu, Emotion recognition from multimodal physiological signals using a regularized deep fusion of kernel machine, IEEE Trans. Cybern. 51 (2021b) 4386–4399, http://dx.doi.org/10.1109/TCYB.2020.2987575.

[30] W. Zheng, W. Liu, Y. Lu, B. Lu, A. Cichocki, Emotionmeter: a multimodal framework for recognizing human emotions, IEEE Trans. Cybern. 49 (2019) 1110–1122, http://dx.doi.org/10.1109/TCYB.2018.2797176.

[31] J. Zhang, Q. Wang, Q. Wang, U-shaped distribution guided sign language emotion recognition with semantic and movement features, IEEE Trans. Affect. Comput. (2024) 1–13, http://dx.doi.org/10.1109/TAFFC.2024.3409357.

[32] S. Zhang, X. Zhao, Q. Tian, Spontaneous speech emotion recognition using multiscale deep convolutional lstm, IEEE Trans. Affect. Comput. 13 (2022) 680–688, http://dx.doi.org/10.1109/TAFFC.2019.2947464.

[33] T. Zhang, W. Zheng, Z. Cui, Y. Zong, Y. Li, Spatial–temporal recurrent neural network for emotion recognition, IEEE Trans. Cybern. 49 (2019) 839–847, http://dx.doi.org/10.1109/TCYB.2017.2788081.

[34] J. Guo, R. Zhou, L. Zhao, B. Lu, Multimodal emotion recognition from eye image, eye movement and eeg using deep neural networks, in: EMBC, 2019, http://dx.doi.org/10.1109/EMBC.2019.8856563.

[35] N. Ayari, H. Abdelkawy, A. Chibani, Y. Amirat, Hybrid model-based emotion contextual recognition for cognitive assistance services, IEEE Trans. Cybern. 52 (2022) 3567–3576, http://dx.doi.org/10.1109/TCYB.2020.3013112.

[36] X. Wang, X. Chen, C. Cao, Human emotion recognition by optimally fusing facial expression and speech feature, :Signal Process., Image Commun. 84 (2020) 115831.

[37] J. Shi, C. Liu, C.T. Ishi, H. Ishiguro, 3D skeletal movement enhanced emotion recognition network, in: APSIPA ASC, 2020.

[38] P.P. Filntisis, N. Efthymiou, P. Koutras, G. Potamianos, P. Maragos, Fusing body posture with facial expressions for joint recognition of affect in child–robot interaction, IEEE Robot. Autom. Lett. 4 (2019) 4011–4018.

[39] T. Wang, S. Liu, F. He, W. Dai, M. Du, Y. Ke, D. Ming, Emotion recognition from full-body motion using multiscale spatio-temporal network, IEEE Trans. Affect. Comput. (2023) 1–15, http://dx.doi.org/10.1109/TAFFC.2023.3305197.

[40] C. Chen, X. Sun, Z. Tu, M. Wang, Ast-gcn: augmented spatial temporal graph convolutional neural network for gait emotion recognition, IEEE Trans. Circuits Syst. Video Technol. 34 (2024) 4581–4595, http://dx.doi.org/10.1109/TCSVT.2023.3341728.

[41] Y. Zhai, G. Jia, Y. Lai, J. Zhang, J. Yang, D. Tao, Looking into gait for perceiving emotions via bilateral posture and movement graph convolutional networks, IEEE Trans. Affect. Comput. 15 (2024) 1634–1648, http://dx.doi.org/10.1109/TAFFC.2024.3365694.

[42] Z. YuMeng, L. Zhen, L. TingTing, W. YuanYi, C. YanJie, Affective-pose gait: perceiving emotions from gaits with body pose and human affective prior knowledge, Multimedia Tools Appl. 83 (2024) 5327–5350.

[43] A. Oğuz, Ö. Ertuğrul, Emotion recognition by skeleton-based spatial and temporal analysis, Expert Systems with Applications 238 (2024) 121981.

[44] C. Beyan, S. Karumuri, G. Volpe, A. Camurri, R. Niewiadomski, Modeling multiple temporal scales of full-body movements for emotion classification, IEEE Trans. Affect. Comput. 14 (2023) 1070–1081, http://dx.doi.org/10.1109/TAFFC.2021.3095425.

[45] E. Ghaleb, A. Mertens, S. Asteriadis, G. Weiss, Skeleton-based explainable bodily expressed emotion recognition through graph convolutional networks, in: 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), 2021, http://dx.doi.org/10.1109/FG52635.2021.9667052.

[46] Z. Cao, G. Hidalgo, T. Simon, S. Wei, Y. Sheikh, Openpose: realtime multi-person 2d pose estimation using part affinity fields, IEEE Trans. Pattern Anal. Mach. Intell. 43 (2021) 172–186, http://dx.doi.org/10.1109/TPAMI.2019.2929257.

[47] Y. Luo, J. Ye, R.B. Adams, J. Li, M.G. Newman, J.Z. Wang, Arbee: towards automated recognition of bodily expression of emotion in the wild, Int. J. Comput. Vis. 128 (2020) 1–25.

[48] P. Ekman, Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique, Psychol. Bull. J. 115 (1994) 268–287.

[49] N. Sebe, I. Cohen, T. Gevers, T. Huang, Emotion recognition based on joint visual and audio cues, in: 18th International Conference on Pattern Recognition (ICPR'06), 2006, pp. 1136–1139, http://dx.doi.org/10.1109/ICPR.2006.489.

[50] C. Breazeal, Emotion and sociable humanoid robots, Int. J. Hum.-Comput. Stud. 59 (2003) 119–155.
[51] S. D'Mello, R.A. Calvo, Beyond the basic emotions: what should affective computing compute?, in: CHI '13 Extended Abstracts on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA, 2013, pp. 2287–2294, http://dx.doi.org/10.1145/2468356.2468751.
[52] K.R. Scherer, What are emotions? and how can they be measured?, Soc. Sci. Inf. 44 (2005) 695–729, http://dx.doi.org/10.1177/0539018405058216.
[53] H.G. Wallbott, K.R. Scherer, Cues and channels in emotion recognition, J. Pers. Soc. Psychol. 51 (1986) 690–699.
[54] H.G. Wallbott, Bodily expression of emotion, Eur. J. Soc. Psychol. 28 (1998) 879–896.
[55] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980–2988, http://dx.doi.org/10.1109/ICCV.2017.322.
[56] D. Clevert, T. Unterthiner, S. Hochreiter, Fast and accurate deep network learning by exponential linear units (elus), 2016, http://dx.doi.org/10.48550/ARXIV.1511.07289, https://arxiv.org/abs/1511.07289.
[57] C. Shorten, T. Khoshgoftaar, A survey on image data augmentation for deep learning, J. Big Data 6 (2019) http://dx.doi.org/10.1186/s40537-019-0197-0.
[58] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (2014) 1929–1958, http://jmlr.org/papers/v15/srivastava14a.html.
[59] M. Coulson, Attributing emotion to static body postures: recognition accuracy, confusions, and viewpoint dependence, J. Nonverbal Behav. 28 (2004) 117–139.
[60] J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization, J. Mach. Learn. Res. 13 (2012) 281–305, http://jmlr.org/papers/v13/bergstra12a.html.
[61] U. Bhattacharya, T. Mittal, R. Chandra, T. Randhavane, A. Bera, D. Manocha, Step: spatial temporal graph convolutional networks for emotion perception from gaits, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020a.
[62] U. Bhattacharya, C. Roncal, T. Mittal, R. Chandra, K. Kapsaskis, K. Gray, A. Bera, D. Manocha, Take an emotion walk: perceiving emotions from gaits using hierarchical attention pooling and affective mapping, in: European Conference on Computer Vision, Springer, 2020b.
[63] M. Poyo Solanas, M.J. Vaessen, B. de Gelder, The role of computational and subjective features in emotional body expressions, Sci. Rep. 10 (2020) 6202.
[64] P. Elias, J. Sedmidubsky, P. Zezula, Understanding the limits of 2D skeletons for action recognition, Multimedia Syst. 27 (2021) 547–561.