ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/355912602

Skeleton-Based Explainable Bodily Expressed Emotion Recognition Through Graph Convolutional Networks

Conference Paper · December 2021

DOI: 10.1109/FG52635.2021.9667052

CITATIONS
4

READS
235

4 authors, including:

Esam Ghaleb
University of Amsterdam
20 PUBLICATIONS **208** CITATIONS
SEE PROFILE

Stylianos Asteriadis
Maastricht University
86 PUBLICATIONS **1,373** CITATIONS
SEE PROFILE

Gerhard Weiss
Maastricht University
249 PUBLICATIONS **8,054** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

3D LIVE View project

ThinkSlim View project

All content following this page was uploaded by Esam Ghaleb on 10 December 2021.

The user has requested enhancement of the downloaded file.

Skeleton-Based Explainable Bodily Expressed Emotion Recognition Through Graph Convolutional Networks

Esam Ghaleb[1], André Mertens[2], Stylianos Asteriadis[1], and Gerhard Weiss[1]
Department of Data Science and Knowledge Engineering, Maastricht University, Maastricht, The Netherlands
[1]{esam.ghaleb, stelios.asteriadis, gerhard.weiss}@maastrichtuniversity.nl
[2]andre.mertens@student.maastrichtuniversity.nl

***Abstract*— Much of the focus on emotion recognition has gone into the face and voice as expressive channels, whereas bodily expressions of emotions are understudied. Moreover, current studies lack the explainability of computational features of body movements related to emotional expressions. Perceptual research on body parts' movements shows that features related to the arms' movements are correlated the most with human perception of emotions. In this paper, our research aims at presenting an explainable approach for bodily expressed emotion recognition. It utilizes the body joints of the human skeleton, representing them as a graph, which is used in Graph Convolutional Networks (GCNs). We improve the modelling of the GCNs by using spatial attention mechanisms based on body parts, i.e. arms, legs and torso. Our study presents a state-of-the-art explainable approach supported by experimental results on two challenging datasets. Evaluations show that the proposed methodology offers accurate performance and explainable decisions. The methodology demonstrates which body part contributes the most in its inference, showing the significance of arm movements in emotion recognition.**

I. Introduction

A growing body of studies demonstrated that variations of body movements convey specific information about people's emotions [1], [2]. For example, Dael et al. [2] showed that features of body movements such as the amount of movement, movement speed, force, fluency, size, and height/vertical position are strong determinants of potency and arousal. In addition, Dael et al. [1] found that several patterns of body movements occur when portraying emotions, which helps in emotion differentiation. These studies showed that body movements and gestures could be an integrated part of a unified nonverbal emotion communication framework since bodily expressions can modulate the conveyed information from the voice and face.

Recent advances in machine learning have brought tremendous improvements to the fields of HCI and affective computing. Usually, those improvements rely on a massive amount of annotated data while offering few insights into their predictions. Explainable Artificial Intelligence (XAI) aims to interpret AI methods to shed light on important aspects that drive models' decisions. Understanding computational models is of great interest in all scientific disciplines. For example, in natural sciences, where machine learning is heavily used, transparency, interpretability, and explainability of used models are essential to increase scientific discovery to have consistency with domain knowledge [3].

The literature on Emotion Recognition from Bodily Expressions (ERBE) points to the fact that there is no consistent quantization of body movement and its characteristics and relations to bodily expressed emotions. For example, a major challenge in the explainability of bodily expression recognition is the lack of movement coding systems, such as the one used for facial expressions, the Facial Action Coding System (FACS) [4]. For instance, there is not a direct correspondence between body movements and affective expressions [1], [2], [5]. Moreover, the relationship between these components is not transcultural and transcontextual, as they can be gender and age-specific as well as idiocentric [5]. Nonetheless, research suggests that there should be a mapping between descriptors and movement characteristics such as joints' positions and velocity. For example, perceptual studies in psychology point to the existence of movement features which contribute to emotion recognition. For instance, a study by De Meijer [5] found that the hands and arms movement are the most significant for distinguishing between affective states. In addition, trunk movement, degree of openness, force, and pace were found to be relevant cues. However, a systematic mapping between computational features of body movements and affective states is still lacking.

This paper aims to fill the gap between explaining how body parts' movements (and their computational features) are related to emotions expressed by subjects. In particular, our study addresses the following research questions: Which body parts' movements contribute the most to emotion recognition and how can body joints' spatio and temporal dynamics be exploited and explained for emotion recognition? We utilize sequences of body joints that represent the dynamics of the human body skeleton. The dynamics of human body skeletons convey significant information for activities, actions, and emotions. Sequences of body joints contain spatio-temporal patterns that can be exploited to capture body movements [6], [7]. Recently, Spatio-Temporal Graph Convolutional Networks (ST-GCNs) emerged as a state-of-the-art family of methods to capture body dynamics in recognizing actions and body gestures from skeletal data [6]. For these reasons, we propose an explainable approach, motivated by the success of GCNs to capture the spatio-temporal dynamics of body joints' movements. To summarize, in this study, our contributions are as follows:

978-1-6654-3176-7/21/$31.00 ©2021 IEEE

- We present an explainable approach for bodily expression recognition based on ST-GCNs.
- We improve the modelling and the representations of body joints, proposing a novel architecture that utilizes attention mechanisms on body parts, i.e., arms, legs, and torso.
- We conduct extensive experiments and evaluation, demonstrating the consistency and effectiveness of the proposed methodology to explain bodily expressions of emotions in two datasets (which are captured in different settings, contexts, and cultures).

The remaining of the paper is organized as follows. Section II gives an overview of literature on bodily expressed emotion recognition, XAI, and GCNs. Section III explains the proposed methodology's components: ST-GCNs, spatial attention mechanisms on body parts, and the explaining method, namely, Class Activiation Maps (CAMs). Section IV presents experimental evaluations and results on two datasets, Green Stimuli [8], [9] and Kinematic Dataset of Actors Expressing Emotions (KDAEE) datasets [10]. Finally, Section V concludes the work and suggests future directions.

II. Related Work

A. Body Joint-Based Emotion Recognition

There has been many studies that focused on using body movements, posture, and gestures for the recognition of emotion expressions. A survey by Noroozi et al. [11] listed the usage of body joints as a major direction of modeling the human body for automatic Emotion Recognition through Bodily Expressions (ERBE). Traditionally, there have been many approaches which rely on designing handcrafted features to model human movement. For example, geometric features and movement features related to velocity, acceleration, and motion protocols have been used in predictive models for ERBE [12], [13].

Recently, deep learning models such as Long-Short-Term Memory (LSTM) and Convolutional Neural Network (CNNs) dominated the field of ERBE. For example, in 2019, Want et al. [14] proposed an end-to-end DNN model based on LSTMs and attention mechanisms, using angles and energy hand-crafted features extracted from a sequence of body joints. The model applied two attention mechanisms. The first one is spatial, on body parts, while the second is temporal across time windows. The model showed significant improvements when using attention mechanisms for recognising pain-related experiences such as fear, anxiety, and avoidance. Nonetheless, LSTMs and CNNs are not suitable for modelling the spatio-temporal dynamics of body movements based on the skeleton, which is embedded through graph data rather than grid-like data such as image sequences. Recently, Yan et al. [6] proposed Spatio-Temporal Graph Convolutional Networks (ST-GCNs) to overcome this limitation, giving a significant performance in domains such as action and activity recognition. In this study, we adopt ST-GCNs for emotion recognition.

B. Explainable AI

Recently, there have been many interpretation techniques for DNNs' features. In a broader aspect, in XAI for DNNs, there are two main approaches [15]: model-transparent and model-agnostic approaches. Model transparent approaches, such as Layer-wise Relevance Propagation (LPR) [16], Class Activation Mapping (CAM) (and its variants) [17], and saliency maps [18], [19] highlight the input features based on models' activation maps and weights. However, model agnostic methods such as Local Interpretable Model Agnostic Explanations (LIME) [20] approximate the relationship between the input data and the decision but treat the model as a black box. In our study, we use a model transparent approach, namely, CAM.

Zhou et al. [17] proposed CAM by replacing the fully connected layers in CNNs with a layer called Global Average Pooling (GAP). GAP averages the feature maps of the last convolutional layers and outputs them as a weighted vector. A weighted sum of this vector is fed to the final softmax loss layer. The score of the predicted class is projected back to the previous convolutional layer to generate the CAMs. Hence, the important regions can be highlighted using GAPs, giving a localization map highlighting the important regions in the input image for the classification process.

C. XAI for Bodily Expressed Emotion Recognition

In XAI for emotion recognition, most of the existing studies focus on the explainability of facial expressions, which covary with discrete emotions. To the best of our knowledge, no studies tackle the challenging problem of explaining DNNs' features for bodily expressed emotions. Recently, at the First International Workshop on Bodily Expressed Emotion Understanding (BEEU), leading experts in the field discussed the challenges and future directions of ERBE research [21]. They mentioned the explainability and interpretability of the developed models as a crucial direction for ERBE.

Nonetheless, a major challenge in explaining bodily expression recognition is the lack of movement coding systems, such as the one used for facial expressions FACS [1]. Hence, the literature points to the fact that there is not a consistent quantization of body movement and its characteristics and relations to bodily expressed emotions [1], [5]. Few attempts have been made, represented by adopting Laban Movement Analysis (LMA) and Body Action Coding System (BACS) [4]. These systems aim at building a mapping between descriptors and movement characteristics such as joints' positions and velocity [22]. Nonetheless, these methods have a few drawbacks. For instance, a drawback for the LMA system is its need for excessive attention for microanalysis and special training for adopting Laban Framework. Therefore, these challenges hamper their adaptation within computational methods in Affective Computing.

D. Graph Convolutional Networks

Recently, Graph Neural Networks (GNNs) have been proposed as models to overcome the challenges posed by

the fact that graphs do not have a spatial structure, and the locality among their vertices is not preserved. GNNs are powerful methods to model data generated via non-euclidean domains as they capture their internal dependence and learn efficient representations. GNNs have achieved great success across many domains, including health-records [23].

Graph Convolutional Networks (GCNs) are a family of GNNs inspired by the success of traditional CNNs. GCNs generalize the convolution operation (the template matching) of the CNNs into GNNs. There are two main types of GCNs [23]. The first type of approaches is based on spectral convolution, where graph convolution is performed in the frequency domain. Spectral convolution utilizes the eigenvalues and eigenvectors of the graph Laplace matrices, a computationally expensive process. ST-GCNs follow the second type, which is based on spatial convolution. The latter approaches apply template matching on the graph vertexes and their neighbours, which are extracted and normalized based on manually designed rules.

III. Methods

In this section, we explain the proposed methodology's components, namely, Spatio-Temporal Graph Convolutional Networks (ST-GCNs) and spatial attention mechanisms based on body parts. We also present the adaptation of Class Activation Maps (CAMs), the explanation method.

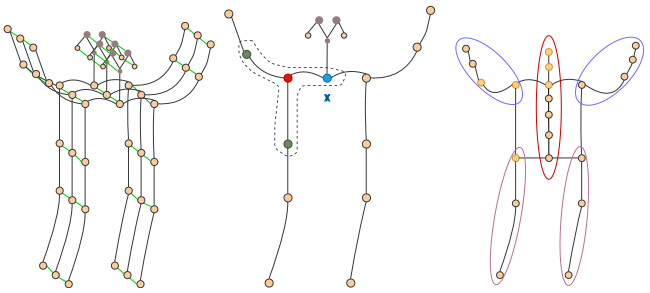
A. Motivation

Perceptual studies in psychology suggest that there exist general movement features which contribute to emotion recognition. For example, De Meijer [5] categorized body movements into the following classes: torso (stretching and bowing), right and left arms (which can include opening and closing), and gait movements characterized by the legs. Motivated by this categorization, we employ ST-GCNs on body joints' sequences to capture the dynamics of skeletal movement. ST-GCNs are suitable models to tackle this challenge. Additionally, they are enhanced by spatial attention mechanisms on body parts to improve their representations and facilitate the explainability of the computed features.

B. Body Skeleton Data and Graph

Recently, pose estimation methods have matured, resulting in accurate estimation of human pose and localization of the 2D and 3D coordinates of body joints [24]. For Green Stimuli Dataset [8], [9], we extract the sequence of body joints using OpenPose [24], which offers an accurate real-time tracking of 2D positions of body joints, and a confidence value for the estimated positions. Hence, each skeleton is represented with 18 body joints (as shown in Fig. 1b). KDAEE dataset was captured using the MoCap system. The dataset provides 21 anatomic nodes (as shown in Fig. 1c), with x, y, and z coordinates.

Subsequently, body motion can be represented with a sequence of 2D or 3D coordinates of the body joints. Instead of computing handcrafted features on the sequence of body joints, we employ ST-GCNs [6], which is used to represent



(a) ST Graph (b) Green Skeleton (c) KDAEE Skeleton

Fig. 1: Graph construction in ST-GCNs. (a) The ST graph consists of body joint (a vertex in the graph), spatial edges (dark lines), and temporal edges (light green lines). (b) a skeleton showing three subsets of the spatial configuration of the graph partitioning strategy. (b) also shows the skeletons of Green Stimuli datasets. (c) illustrates the partitioning of body parts using a skeleton of the KDAEE dataset.

the dynamics of the human body embedded within joint sequences. In this manner, ST-GCNs extract hierarchical representations of skeleton sequences. As suggested in [6], we construct undirected spatio-temporal graphs, consisting of V vertices (joints) and E edges, $G = (V, E)$. As shown in Fig. 1a, there are two types of edges in ST-GCN, spatial edges that adhere to the natural connectivity of joints and temporal edges that connect the same joints across time windows.

In the graph shown in Fig. 1a, each body joint (vertex) has a three-dimensional vector, i.e. 2D joint positions and their confidence or the 3D joint positions. A video clip of a skeleton sequence can be represented with a tensor as follows: $X \in \mathbb{R}^{C \times T \times V}$, where C is the data point of a vertex (x, y, z or confidence), T is the number of frames in a video sequence, and V is the total number of body joints (e.g., 18 body joints). For each joint, we augmented their relative positions (r_i) with respect to the centre joint of the skeleton (i.e., the central spine), in addition to the 2D and the 3D positions of the joints. The augmentation is calculated as follows: $r_i = X[:,:,i] - X[:,:,c]$, where $X[:,:,c]$ indicates the position of the central joint. Hence, the joints' coordinates and relative positions are concatenated, resulting in $6D$-vectors representing the body joints.

C. Spatio-Temporal GCNs

Given the Spatio-Temporal (ST) data defined above, ST graph convolution is applied across multiple layers using a predefined graph. GCNs aims at generalizing the convolution operation to GNNs where input features are represented on a spatial graph V. Specifically, in ST-GCNs, a feature map at a frame t can be defined as follows: f_{in}: $V \in \mathbb{R}^c$, which has a c-dimensional vector for each node in the graph, e.g. 6 in the input graph of our study. In the spatial dimension, a graph convolution on a node (v_i) is defined as follows [6]:

$$f_{out}(v_i) = \sum_{v_j \in B_i} \frac{1}{Z_{ij}} f_{in}(v_j).w(l_i(v_j)) \quad (1)$$

where B_i refers to the sampling area for the convolution around v_i, which is defined by the 1-distance neighbour nodes (v_j) of the target (root) node (v_i), Z_{ij} is a normalizing term equal to the cardinality of the corresponding sampling subset, f_{in} is a feature map at a frame t, w is a learnable kernel similar to the conventional convolutional operation, which is a weighting function on the input vector. Finally, l_i is a mapping function to assign a unique weighting function (w) for each vertex, given the defined topology of the spatio-temporal graph of the human skeleton.

A key step in GCNs is to define the sampling function B_i. Note that the sampling function (B_i) varies for each node; however, the number of weighting vectors (w) is fixed. l_i maps the weight vectors for each vertex. Specifically, the ST-GCNs proposed by Yan et al. [6] adopted a spatial configuration partitioning for mapping weights. In this partitioning, the spatial localization of the human skeleton is utilized, inspired by the fact that body motion occurs concentrically and eccentrically. As illustrated in Fig. 1b, the partitioning consists of three subsets: (1) a root node, (2) a centripetal group, which contains the neighbouring nodes of the root node that are closer to the gravity centre of the skeleton than the root node, (3) centrifugal nodes which are further from the gravity centre of the skeleton than the root node.

To apply the convolutional operation, ST-GCNs [6] adopted a similar approach as in Kip and Welling [25], where skeleton graph is represented by the adjacency matrix $\boldsymbol{A}$ and an identity matrix $\boldsymbol{I}$, hence, in a vectorized form, the equation in (1) can be re-written as follows:

$$\boldsymbol{f}_{out} = \sum_{k}^{K_v} \boldsymbol{A}_k \odot \boldsymbol{M}_k (\boldsymbol{f}_{in} \boldsymbol{W}_k) \tag{2}$$

where K_v denotes the kernel size of the spatial dimension, i.e. with the spatial configuration partitioning strategy, and K_v is set to 3. $\boldsymbol{A}_k = \hat{D}^{-\frac{1}{2}} A \hat{D}^{-\frac{1}{2}}$, $\boldsymbol{A}_k \in \mathbb{R}^{V \times V}$ is the adjacency matrix which is defined according to the partitioning strategy explained above, and $\hat{D}^{ii} = \sum_j A^{ij} + I^{ij}$. $\boldsymbol{W} \in \mathbb{R}^{C_{out} \times C_{in} \times 1 \times 1}$ is learnable matrix of 1×1 convolution (which corresponds to the weighting vector in (1)). $\boldsymbol{M} \in \mathbb{R}^{V \times V}$ is a learnable matrix which indicates the importance of each vertex. $\odot$ denotes the element-wise product between two matrices. For the temporal convolution, an $L \times 1$ convolutional layer is applied to learn representational features on the adjacent frames.

D. Spatial Attention

Body gestures and movements are usually performed by a collection of joints on each of the main body parts, i.e., the torso, arms, and legs. In our framework, following the work of Song et al. [7] and Want et al. [14], we applied spatial attention on body parts as follows:

$$\boldsymbol{f}_{part} = \boldsymbol{f}_{in}(p) \odot softmax(ReLU(pool(\boldsymbol{f}_{in})\boldsymbol{W})\boldsymbol{W}_p) \tag{3}$$

$$f_{out} = Concatenate(\{\boldsymbol{f}_p | p \in 1,2,\ldots,P\}) \tag{4}$$

where $\odot$ indicates element-wise multiplication, $pool()$ refers to the temporal average pooling and a body parts' joints pooling. $\boldsymbol{W}$ and $\boldsymbol{W}_p$ are learnable projection matrices where $\boldsymbol{W}$ is shared among all body parts, while $\boldsymbol{W}_p$ is specific for each body part. Focusing on those body parts helps the model in learning patterns related to bodily expressions. More importantly, attending to the movement of body parts facilitates the explanation of body parts' involvement in expressing a certain emotion.

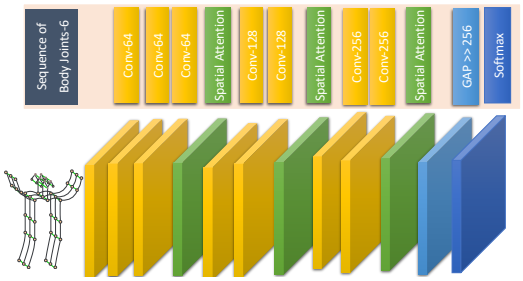


Fig. 2: The architecture of the proposed framework, consisting of seven spatio-temporal GCN layers, three layers for spatial attention mechanisms, GAP, a classification layer.

E. System Architecture and Training

Fig. 2 presents the topology of the proposed framework. It consists of 7 spatio-temporal convolutional blocks. The numbers of each block represent the number of input channels. Global Average Pooling (GAP) applies spatio-temporal pooling. Before the GAP and prediction layers, note that there is no average/mean spatial pooling over the joints to preserve the skeleton topology. This property is later utilized to compute joints' activations and explain body parts' movements in emotional expressions. We used Stochastic Gradient Descent (SGD) with a momentum of 0.9, and the optimization was run for 200 epochs. In our experimental evaluations, the learning rate was set to 0.1 and divided by 10 at the 50^{th} and 100^{th} epochs. The batch size was set to 64.

F. Explanation Method: Class Activation Maps

In our work, Class Activation Maps (CAMs) use GAP to highlight the discriminative body parts when inferring a specific emotion given a joint sequence. In particular, we can highlight the important regions using GAP. As explained in subsection III-B, the dimensionality of a given sequence, $f_c(t,v)$, at the final convolutional layer (i.e., the layer before GAP) is $f_c(t,v) \in \mathbb{R}^{C \times T \times V}$. GAP is performed across T and V, i.e., the spatial and temporal dimensions, preserving the number of feature channels: $F_c = \sum_{t,v} f_c(t,v)$. The resulting features ($F_c \in \mathbb{R}^C$) have a C-dimensional vector. This vector is used in the fully connected layer to produce the desired output (i.e., predictions of emotions at hand), using a weighted sum.

Subsequently, the input to the softmax layer (the classification layer) for an emotion S_e is $\sum_c w_c^e F_c$, where w_c^e is the emotion score which is known as the weight score for the corresponding emotion e and the activation map c. Hence, the emotion score is plugged into the feature map at the final convolutional layer to obtain a discriminative localization

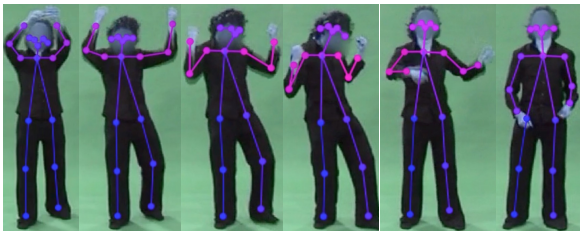


Fig. 3: Illustrations of body parts' activations in emotional expressions. Note the high activation when the body gestures are expressive, demonstrating the ability of the model to capture the joints' movements.

map as follows:

$$M_e(t,v) = \sum_c w_c^e f_c(t,v) \tag{5}$$

where $M_e(t,v)$ is the localization map, indicating the importance of the activations directly for the spatio-temporal grid, i.e., the corresponding body joint (v) at the time window t. Hence, these maps are a weighted linear sum of the corresponding body joints' activation maps that show the discriminative joints leading to the classification of the emotion in the body joints' sequence.

Figure 3 shows activation maps for body joints in a sequence. Note that the proposed method, utilizing CAM, highlights the activated body joints in terms of body parts (a darker shade of red means a higher activation). In our study, thanks to the employment of body attention mechanisms, we were able to obtain consistent CAMs for each body part. Specifically, our explanation method directly uses the localization maps ($M_e(t,v)$) to calculated body parts' activation, linking them to emotion expression and recognition. Mathematically, we perform the following average pooling on body parts:

$$P_i^m = \frac{1}{N_{v_{P_i}}} \sum_{t, v \in v_{P_i}} M_e(t,v) \tag{6}$$

where P_i^m is the activation value for the body part, P_i. P_i includes body joints of the corresponding body part ($v \in v_{P_i}$). $N_{v_{P_i}}$ is equal to the number of body joints in the body part P_i. Thus, we pool the activations obtained through the equation (5), over the joints' sequence, providing a single value that represents the activation.

IV. Evaluations

We present comprehensive experimental evaluations on the proposed approach and its explainability on two datasets, namely, Kinematic Dataset of Actors Expressing Emotions (KDAEE) [10] and Green Stimuli (GreSti) Dataset [8], [9].

1) Green Stimuli Dataset: Green Stimuli dataset is a bodily expression rich dataset collected in a study to examine bodily expressions, and human perception of emotion display [8], [9]. It consists of 871 video clips, where each video has a duration of 2 seconds. It was collected using RGB cameras. It is a gender-balanced dataset with 17 males and 17 females. Subjects were coached to express affective expressions in a naturalistic way. Actors mainly have a European cultural background. They performed bodily expressions of emotions which were varied as to their subjective characteristics. In particular, subjects expressed emotional body movements in six discrete basic Ekmanian emotions [26]: anger, disgust, fear, happiness, sadness, and surprise. In addition, subjects performed neutral body expressions such as coughing, walking, standing, or pulling the nose. The faces of the subjects were blurred to study emotions independent of facial expressions. Currently, Green Stimuli dataset is not publicly available due to the sensitivity of the collected data and the restrictive data sharing policies in the European Union. However, we intend to release portions of the data, such as body joints, for the research community.

Dataset	Method	A	D	F	H	N	Sa	Su	Folds Average
KADEE	Baseline	67.5	56.5	64.9	70.7	82.9	53.4	50.2	62.6
	SA	70.1	57.8	68.4	74.5	82.9	55.0	52.7	65.0
GreSti	Baseline	65.6	58.1	63.5	67.8	77.4	73.5	63.8	67.5
	SA	63.1	55.6	72.7	63.3	81.8	79.8	64.5	69.2
	I3D [27]	65.0	61.1	57.9	78.9	71.4	84.2	63.2	68.9

TABLE I: Ablation study on the performance (in terms of accuracy %) of the proposed method across two datasets' emotion types, with and without Spatial Attention (SA). Labels are referred to as follows: Anger (A), Disgust (D), Fear (F), Happiness (H), Neutral (N), Sadness (Sa), and Surprise (Su).

2) KDAEE: KDAEE is a kinematic dataset which has a total of 1402 recordings, gathered using motion capture technology. It is the largest kinematic dataset of bodily expressed emotions, capturing the movement of the whole body. The dataset was created to study discrete emotions (i.e., anger, disgust, happiness, fear, sad, and surprise) and neutral states, from bodily cues. The dataset collection was performed by 22 subjects (50% females). Actors have an Asian cultural background. For the aim of the data collection, subjects completed two types of movement: spontaneous (based on actors' understanding of emotion expression) and scenario-based movements (using predefined scenarios created by the dataset developers). Actors performed the movements successively to display the discrete emotions. The dataset provides only raw kinematic data, consisting of the positions and rotation of the targeted 72 anatomical nodes. In our study, we used the positions of the main 21 anatomical nodes, as shown in Fig. 1c.

3) Evaluation Protocols: Both datasets were divided into 10-folds for cross validation. In each fold, one part of the data (90%) is used for training, and the remaining part is used for testing. The reported results in this section are the average accuracies of the ten folds.

A. Ablation Study

We introduce an ablation study to demonstrate the effectiveness of the proposed framework and the employed spatial attention mechanisms. In particular, we provide the results of the proposed framework using the following components:

- A comparison with I3D [27], a state-of-the-art 3D-CNN model for video classification. In this comparison, we provide results only for Green Stimuli, which contains

(a) KDAEE (b) Green Stimuli

Fig. 4: The t-SNE embeddings of the two datasets features which are obtained from the seventh convolutional layer. Note that clusters are well structured, where colours represent body parts. Blue, green, and orange colours represent arms, legs, and torso, respectively.

RGB video recordings. Nonetheless, the developers of KDAEE provide only raw kinematic data obtained from the MoCap system. Hence, for KDAEE, we only provide the classification results based on skeleton data.

- A baseline framework which does not use spatial attention mechanisms.
- The proposed framework, which uses spatial attention mechanisms, utilizing the joints of specific body parts as explained in section III-D.

Table I presents the performance on the two datasets. As shown in the table, the framework's performance (in terms of accuracy) benefited significantly from the notion of spatial attention. We notice that attention mechanisms based on body parts enhanced the system performance by at least 1.7% in both datasets. This improvement is also present among all emotion types in the KDAEE dataset and most of the classes of the Green Stimuli dataset. More importantly, the spatial attention gives the proposed method a stronger representation of body parts since our analysis focuses on body parts to explain bodily movements in emotional expressions.

Finally, previous studies have pointed out that skeleton-based models (e.g., GCNs) can be inferior to those models based on RGB-data for tasks such as activity and action recognition [6]. Nonetheless, the results of Green Stimuli in Table I show that the proposed framework (based on skeleton data) obtains comparable accuracies to the ones obtained by RGB-images based I3D model. Our approach slightly outperformed I3D results by 0.3%. Moreover, an essential advantage of the employed approach (i.e., GCNs) is the natural correspondence of the body parts' joints and their movement over time with the GCNs nodes. This is an important advantage over methods such as I3D, where RGB image sequences are employed. In the latter case, the tracking of the body joints is challenging in terms of explainability.

B. Explainability Study

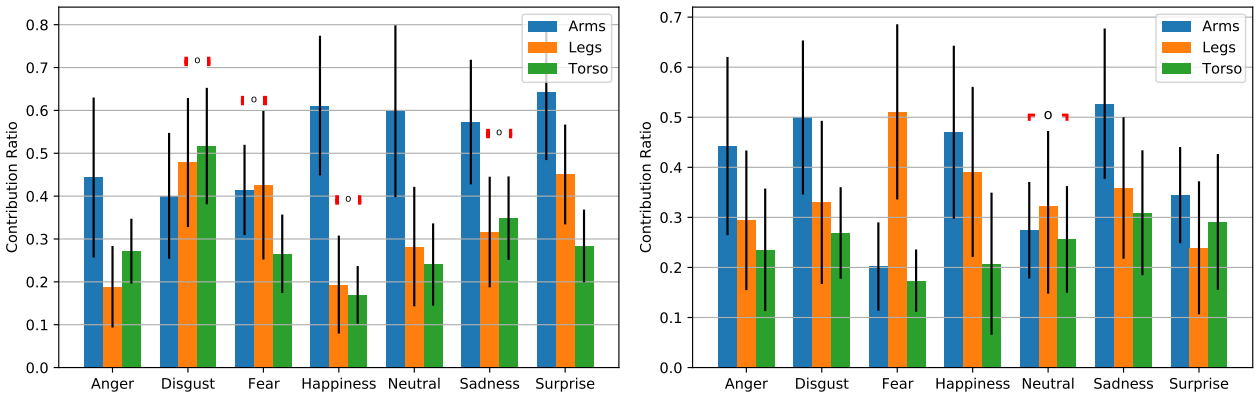
In section III-F, we presented the adaptation of Class Activation Maps (CAMs) as an explanation method for the proposed approach. An important question is whether the explanations of this method are correct. A simple way to verify the explanation is to use ground truths of the importance of the body parts in expressing emotions. Nonetheless, such ground truths are not provided in the selected datasets since the body parts' movements are not annotated regarding their importance in expressing emotions. Hooker et al. [28] suggested alternative procedures, where important features (nodes in the case of GCNs) are deleted while keeping the rest of the features. In case the explanation method identifies the discriminative features correctly, deleting those features causes a decline in the performance of the identification task. In our work, we followed this strategy, performing systematic body parts' occlusion. We replace parts' joints with Gaussian noise and subsequently re-run the classification task (evaluation without retraining) for assessment.

As explained in Sections III and III-D, motivated by studies coming from the field of psychology [1], [2], [5], we decided to categorize body movements into the following parts: torso, right and left arms, and legs for capturing gait movements. Combinations of the body parts are not considered since our study focuses on explaining the main body parts independently. In addition, particular movements such as arms' opening or closing were not annotated in the video clips of both datasets. Hence, we provide a high-level explanation for the focus of the approach when inferring emotions, aiming to answer the research question: which body parts are influential in the decision of computational models (i.e., GCNs) for emotion recognition. Hence, our evaluations present the study findings as follows:

- Employing CAMs to calculate the contribution of body parts' movement in emotion recognition.
- Verifying these findings in terms of body parts' occlusion.

1) Visualization of Joints Embeddings: Prior to delving into the analytical results, we present qualitative results, demonstrating how the proposed topology produces embeddings of joints aware of body parts. Fig. 4 presents the embeddings of the seventh convolutional layer. We extracted the features from the whole sequences for each body joint and each sample. We used t-SNE (dimensionality reduction and visualization tool) to visualize the resulting embeddings. The figure clearly shows that the embeddings are clustered into three main classes, representing the three main body parts, namely, arms, legs, and torso. Interestingly, our networks were not trained explicitly to cluster those body parts; however, the employed body parts attention mechanisms enhanced the performance of the proposed method and improved the categorization of body parts' embeddings.

2) Body Parts' Contributions: Fig. 5 presents the distributions of body parts' activations for each emotion type. For each video, three activation values for arms, legs, and torso joints are calculated using CAMs as described in Section III-F. The distributions shown in bar plots report the average activations (bars) and standard deviations (antennas), across emotions' recordings (video clips). Firstly, as shown in the figure, arms' activations are the highest among most classes, with exceptions in fear and disgust (in Green Stimuli) and fear and neutral state (in KDAEE). Secondly, the contributions of legs' joints are second in four classes of Green Stimuli (namely, disgust, happiness, neutral state, and surprise), and four classes of KDAEE datasets (namely, anger, disgust, happiness, and sadness).



(a) Green Stimuli

(b) KDAEE

Fig. 5: Bar plots illustrate body parts' contribution given body parts' activations in emotion classification of body joints' sequences. The letter "o" marks the two distributions that do not differ significantly ($p > 0.05$).

	Green Stimuli								KDAEE							
Emotions	None	L-A	R-A	L-L	R-L	R&L-A	R&L-L	T	None	L-A	R-A	L-L	R-L	R&L-A	R&L-L	T
Anger	63.1	59.3	49.3	58.5	55.0	36.3	49.7	57.4	70.1	44.4	35.5	48.8	57.7	4.7	32.8	50.4
Disgust	55.6	50.9	51.6	53.2	54.1	35.3	45.5	40.0	57.8	40.3	36.3	37.0	44.6	23.6	30.6	42.3
Fear	72.7	63.0	56.3	60.6	62.0	41.3	36.3	60.6	68.4	46.0	51.1	42.7	10.1	41.2	5.3	44.6
Happy	63.3	49.7	29.7	57.6	53.7	4.7	48.2	58.4	74.5	57.3	35.1	51.9	44.1	19.8	27.9	52.5
Neutral	81.8	66.4	59.5	79.9	79.5	38.5	72.5	73.1	82.9	46.7	58.3	61.5	60.8	21.8	45.4	60.3
Sad	79.8	62.7	60.4	72.6	74.7	35.2	66.7	58.9	55.0	36.1	34.9	37.2	42	28.5	33.1	42.7
Surprise	64.5	48.1	33.9	57.7	60.2	10.5	49.2	51.2	52.7	26.7	29.3	35.6	36.1	20.8	26.6	28.0
Avg-Acc	69.2	58.2	50.6	63.5	63.0	32.5	54.1	57.9	65.0	41.4	38.7	43.2	43.4	25.4	30.2	45.0

TABLE II: The classification accuracies of the proposed method when occlusion is applied on the following body parts: Left Arm (L-A), Right-Arm (R-A), Left Leg (L-L), Right-Leg (R-L), Torso (T). The results are reported for the two datasets, namely, KADEE and Green Stimuli, and for all emotion types.

Interestingly, the legs' joints are leading the contributions in the recognition of fear in both KDAEE and Green Stimuli datasets. Thirdly, torso joints are less important than arms and legs joints, leading the contributions only in recognizing disgust (in Green Stimuli). Besides, they are in the second place to recognize anger and sadness (in Green Stimuli), and surprise and neutral state (in KDAEE). The importance of arms' movements (which include hands' movements) in the expression and perception of emotions is in alignment with findings from studies in the field of psychology, which suggest that they are the most expressive body parts [1], [5].

Additionally, Welch's t-test was performed to check the similarities between the distributions of body parts' activations. If t-tests' p-value is greater than 0.05, the distributions were considered not significantly different from each other. In Fig. 5, the distributions that do not differ significantly are marked with the letter "o". In Green Stimuli, our evaluations show that the distributions of body parts' activations are significantly different from each other in the following classes: anger, neutral state, and surprise. However, in the cases of disgust, sadness, and happiness, the distributions of legs' and torso's contributions are not significantly different. In the case of fear, it was observed that the distributions of arms' and legs' activations do not differ significantly. In KDAEE, we notice that the distributions of body parts' activations are significantly different, except in neutral expressions where the activations of arms and torso do not differ significantly.

It is important to note that the difference between the contribution of the body parts among the same emotions across the two detests to be expected. The two datasets were captured in two different cultures, where Green Stimuli and KDAEE subjects come from European and Asian cultural backgrounds, respectively. Additionally, there is not a direct correspondence between movement and affective expressions. The relationship between these components is not transcultural and transcontextual. Bodily expressions can be gender and age-specific, as well as idiocentric [5].

3) Verification of Body Parts' Explanations: In this analysis, we aim at verifying the explanations mentioned above provided by the explanation approach for each body part and emotion. To do so, we control whether the occlusion of a body part that is considered discriminative will affect the classification performance or not.

Table II presents a complete picture of the results following the performed occlusion evaluations. On average, for both datasets, we observe that the occlusion of the arms decreases the classification performance the most, followed by legs and torso, respectively. The results show that occluding body parts, which were considered the most discriminative by CAMs, rapidly decreases the performance. On body part and class levels, the main observations are as follows:

- The occlusion of left or right arms has a different impact, where the right arm is more discriminative in recognition of most emotions than the left arm. A study by

Poyo Solanas et al. [9] found that limbs' movements and symmetry features contribute differently in emotion perception.

- The occlusion of the left and right legs does not differ significantly compared to the left and right arms.
- Neutral state is the least affected expression when removing body torso and legs for both datasets. It is also the most accurately recognized class. A potential explanation is that these body parts are less activated in neutral expressions, and removing them decreases the accuracy less than emotional expressions.
- For fear, the occlusion of leg joints affects the recognition rate the most. The legs are the body part activated the most in the expressions of fear (as shown in Fig. 5).
- Removing any body part which is considered the most discriminative by the explanation method (and is significantly different from other parts) decreases the accuracy more than removing other parts. There are two exceptions: the occlusion of the torso for disgust (in Green Stimuli) and of the legs in the neutral state (in KDAEE). The reason behind these exceptions is that the body parts' activations do not significantly differ from each other completely ($p > 0.05$, as introduced in the previous Subsection); hence, the occlusion of one body part has a low impact.

This analysis demonstrates the applicability of the employed methods (the topology and the adapted explanation methods) to explain the computed features.

V. Conclusions

This study tackles a major challenge of the explainability of deep neural networks in bodily expressed emotion recognition using skeleton joints. It proposes a novel approach based on state-of-the-art methods, namely, Graph Convolutional Networks and spatial attention mechanisms. Our research provides comprehensive evaluations of the framework, demonstrating its robustness and effectiveness to explain body movements in emotion recognition. Our findings show that hands and arm movements are the most significant for emotion recognition. Future work should benefit from annotated body movements in expressing emotions to check if the proposed method agrees with such annotations.

References

[1] N. Dael, M. Mortillaro, and K. R. Scherer, "Emotion expression in body action and posture." *Emotion*, vol. 12, no. 5, p. 1085, 2012.

[2] N. Dael, M. Goudbeek, and K. R. Scherer, "Perceived gesture dynamics in nonverbal expression of emotion," *Perception*, vol. 42, no. 6, pp. 642–657, 2013.

[3] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke, "Explainable machine learning for scientific insights and discoveries," *IEEE Access*, vol. 8, pp. 42 200–42 216, 2020.

[4] E. M. Huis In't Veld, G. J. van Boxtel, and B. de Gelder, "The body action coding system ii: muscle activations during the perception and expression of emotion," *Frontiers in behavioral neuroscience*, vol. 8, p. 330, 2014.

[5] M. De Meijer, "The contribution of general features of body movement to the attribution of emotions," *Journal of Nonverbal behavior*, vol. 13, no. 4, pp. 247–268, 1989.

[6] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[7] Y.-F. Song, Z. Zhang, C. Shan, and L. Wang, "Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1625–1633.

[8] M. Kret, S. Pichon, J. Grèzes, and B. de Gelder, "Similarities and differences in perceiving threat from dynamic faces and bodies. an fmri study," *Neuroimage*, vol. 54, no. 2, pp. 1755–1762, 2011.

[9] M. Poyo Solanas, M. J. Vaessen, and B. de Gelder, "The role of computational and subjective features in emotional body expressions," *Scientific reports*, vol. 10, no. 1, pp. 1–13, 2020.

[10] M. Zhang, L. Yu, K. Zhang *et al.*, "Kinematic dataset of actors expressing emotions," *Scientific data*, vol. 7, no. 1, pp. 1–8, 2020.

[11] F. Noroozi, D. Kaminska, C. Corneanu, T. Sapinski, S. Escalera, and G. Anbarjafari, "Survey on emotional body gesture recognition," *IEEE transactions on affective computing*, 2018.

[12] A. Kleinsmith, N. Bianchi-Berthouze, and A. Steed, "Automatic recognition of non-acted affective postures," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 41, no. 4, pp. 1027–1038, 2011.

[13] Y. Luo, J. Ye, R. B. Adams *et al.*, "Arbee: Towards automated recognition of bodily expression of emotion in the wild," *International Journal of Computer Vision*, vol. 128, no. 1, pp. 1–25, 2020.

[14] C. Wang, M. Peng, T. A. Olugbade, N. D. Lane, A. C. D. C. Williams, and N. Bianchi-Berthouze, "Learning temporal and bodily attention in protective movement behavior detection," in *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, 2019, pp. 324–330.

[15] J. V. Jeyakumar, J. Noor *et al.*, "How can i explain this to you? an empirical study of deep neural network explanation methods," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[16] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, "Layer-wise relevance propagation: an overview," in *Explainable AI: interpreting, explaining and visualizing deep learning*. Springer, 2019, pp. 193–209.

[17] B. Zhou, A. Khosla *et al.*, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[18] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[19] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[20] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[21] R. O. Gilmore, K. L. Johnson, A. Lapedriza, X. Lu, and N. Troje, "Panel bodily expressed emotion understanding research: A multidisciplinary perspective," 2020.

[22] M. Karg, A.-A. Samadani, R. Gorbet, K. Kühnlenz, J. Hoey, and D. Kulić, "Body movements for affective expression: A survey of automatic recognition and generation," *IEEE Transactions on Affective Computing*, vol. 4, no. 4, pp. 341–359, 2013.

[23] Z. Wu, S. Pan, F. Chen *et al.*, "A comprehensive survey on graph neural networks," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 1, pp. 4–24, 2020.

[24] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: realtime multi-person 2d pose estimation using part affinity fields," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 1, pp. 172–186, 2019.

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

[26] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[27] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[28] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim, "A benchmark for interpretability methods in deep neural networks," *Advances in Neural Information Processing Systems*, vol. 32, pp. 9737–9748, 2019.

View publication stats