





# Emotion Transformer: Attention Model for Pose-Based Emotion Recognition

Pedro V. V. Paiva<sup>1,3</sup><sup>a</sup>, Josué J. G. Ramos<sup>2</sup><sup>b</sup>, Marina L. Gavrilova<sup>3</sup><sup>c</sup> and Marco A. G. Carvalho<sup>1</sup><sup>d</sup>

<sup>1</sup>*School of Technology, University of Campinas, Limeira, Brazil*

<sup>2</sup>*Cyber-Physical Systems Division, Renato Archer IT Center, Campinas, Brazil*

<sup>3</sup>*Department of Computer Science, University of Calgary, Calgary, Canada*

**Keywords:** Body Emotion Recognition, Affective Computing, Video and Image Processing, Gait Analysis, Attention-Based Design.


**Abstract:** Capturing humans' emotional states from images in real-world scenarios is a key problem in affective computing, which has various real-life applications. Emotion recognition methods can enhance video games to increase engagement, help students to keep motivated during e-learning sections, or make interaction more natural in social robotics. Body movements, a crucial component of non-verbal communication, remain less explored in the domain of emotion recognition, while face expression-based methods are widely investigated. Transformer networks have been successfully applied across several domains, bringing significant breakthroughs. Transformers' self-attention mechanism captures relationships between different features across different spatial locations, allowing contextual information extraction. In this work, we introduce *Emotion Transformer*, a self-attention architecture leveraging spatial configurations of body joints for Body Emotion Recognition. Our approach is based on the visual transformer linear projection function, allowing the conversion of 2D joint coordinates to a regular matrix representation. The matrix projection then feeds a regular transformer multi-head attention architecture. The developed method allows a more robust correlation between joint movements with time to recognize emotions using contextual information learning. We present an evaluation benchmark for acted emotional sequences extracted from movie scenes using the BoLD dataset. The proposed methodology outperforms several state-of-the-art architectures, proving the effectiveness of the method.


## 1 INTRODUCTION


Humans can express a wide variety of information through communication channels, commonly defined as verbal and non-verbal (Burgoon et al., 2021). Numerous computer applications can benefit from mimicking the human ability to recognize the non-verbal state, a.k.a. affective state. Some examples of applications are surveillance, education, and health care (physical and/or emotional), among others. A particular field that can benefit from accurate affective state recognition is Socially Interactive Robotics (SIR). The goal of SIR is to establish a human-robot relationship that is closer to the human-human equivalent (Goodrich et al., 2008). Therefore, robots must


perceive, respect, and reproduce the two channels of communication (verbal and non-verbal), respecting the social rules while helping the human being. Non-verbal communication can be divided according to the mode of social interaction, the most relevant being: kinesics and chronemic (Jones, 2013). Kinesics is related to movements and some consider it as communicative as verbal communication. The chronemic, or temporal factor, allows to identify and understand the role of the rhythm of human communication.

Most of the methods available in the literature are restricted to emotions expressed using face, a subject that has been explored for decades (Noroozi et al., 2018; Luo et al., 2008). However, facial expressions are not the only emotional display in the human body. Humans can also infer others' emotional expressions from body movements, something recently explored by affective state recognition techniques (Avola et al., 2020; Noroozi et al., 2018; Bhatia et al., 2022). Body

<sup>a</sup> <https://orcid.org/0000-0002-3743-1985>

<sup>b</sup> <https://orcid.org/0000-0002-5815-2424>

<sup>c</sup> <https://orcid.org/0000-0002-5338-1834>

<sup>d</sup> <https://orcid.org/0000-0002-1941-6036>

affective state can be inferred by analyzing the coordinates of body joints over time. A frequent approach to collecting human poses consists of using motion capture systems (Menolotto et al., 2020). Depth-based sensors, like the Microsoft Kinect, came as an alternative to obtaining skeletal data (Rahman and Gavrilova, 2017). Tasks such as activity and gesture recognition can be performed efficiently using body skeletal data (Maret et al., 2018). With the advent of deep learning pose estimation algorithms, videos from simple RGB cameras are now able to generate body joints, allowing a great range of applications. Recent Body Emotion Recognition (BER) strategies have been using convolutional (Ilyas et al., 2021) and recurrent networks, like LSTM (Avola et al., 2020), to improve the accuracy of models.

For BER systems, the understanding of the time in which movements occur is as important as the joint positioning itself, as proved by time-aware methods. Previous research focused only on recurrent networks to learn long-range dependencies. Nevertheless, recurrence has been overcome by transformers (Lin et al., 2022) in most of the applications of its predecessor. Contextual information learning, a key aspect of transformer networks built-in in its self-attention mechanism, allows a more robust correlation between data and its position. However, solutions that rely exclusively on self-attention blocks have yet to be investigated for this task. This paper proposes encoding contextual body position information to improve emotion recognition performance.

Inspired by the transformer network successes in several areas, we propose a new model called *Emotion Transformer* using the same principles of the Vision Transformer architecture (Dosovitskiy et al., 2020). We first split body posture sequences into patches and provide the sequence of linear embeddings of these patches as an input to a self-attention architecture. We train the model in a supervised way to predict categorical emotions. A labeled video emotion dataset, extracted from movie scenes and containing occlusions and distance from camera variations, is used to evaluate the performance. Specifically, the present work has the following contributions:

- (i) A transformer network that utilizes a self-attention mechanism for identifying emotions from body movements is proposed.
- (ii) A context-aware methodology that efficiently extracts spatial and temporal features is introduced, taking advantage of long video sequences.
- (iii) A novel transformer-based deep learning architecture obtained a high precision on a challenging problem of identifying 26 emotional labels.

- (iv) The proposed method is more accurate than prior methods for emotion recognition on the in-the-wild benchmark BoLD dataset. Our method outperforms baseline methods by 28%-25% (mAP-mRA).

The remainder of this paper is organized as follows. Section 2 provides a brief review of related works, including the gaps in the area. In Section 3 the proposed approach is presented, including a description of the used dataset. Section 4 shows our initial results and comparison with different recent approaches. Finally, Section 5 concludes the paper and discusses some open issues for future works.

## 2 RELATED WORKS

Works in the literature have already identified the important relationship between spatiotemporal displacement and human emotions expressed through the body. (Yang et al., 2021), in addition to proposing a dataset, demonstrates that angles and distances of human body joints can serve as input for LSTM networks to identify emotions. In the case of (Yang et al., 2021), emotions were treated as simple actions of the body, which is reflected in a low accuracy of their recognition (average of 56.4%). In (Avola et al., 2020), the combination of three-dimensional poses, movement descriptors, and the use of temporal local features is able to reach the state of the art in detecting non-acted emotions (79.8%). It is worth mentioning that (Avola et al., 2020) validated its methods on a 3D human representation dataset, making its replication impossible without first solving the three-dimensional pose estimation problem. In (Shen et al., 2019), temporal information is associated with the representation of the body via 2D skeleton using optical flow. It is demonstrated in the evaluated dataset, created by the authors, that features extracted from the skeleton complement the optical flow information when merged.

Hybrid approaches that extract emotional expressions from both face and body have been also investigated. (Sun et al., 2018) proposes the combination of CNN and LSTM to identify emotions in video sequences where the face and upper body are visible. The method is able to find parts of the video where there are more spatiotemporal information, called “words” and video “skeletons”. After separate training for the body and face, the estimators are combined in a hierarchical fusion. (Ilyas et al., 2021) and (Ly et al., 2018) use very similar strategies of combining CNN and LSTM with changes only in the form and method of merging the modalities. How-

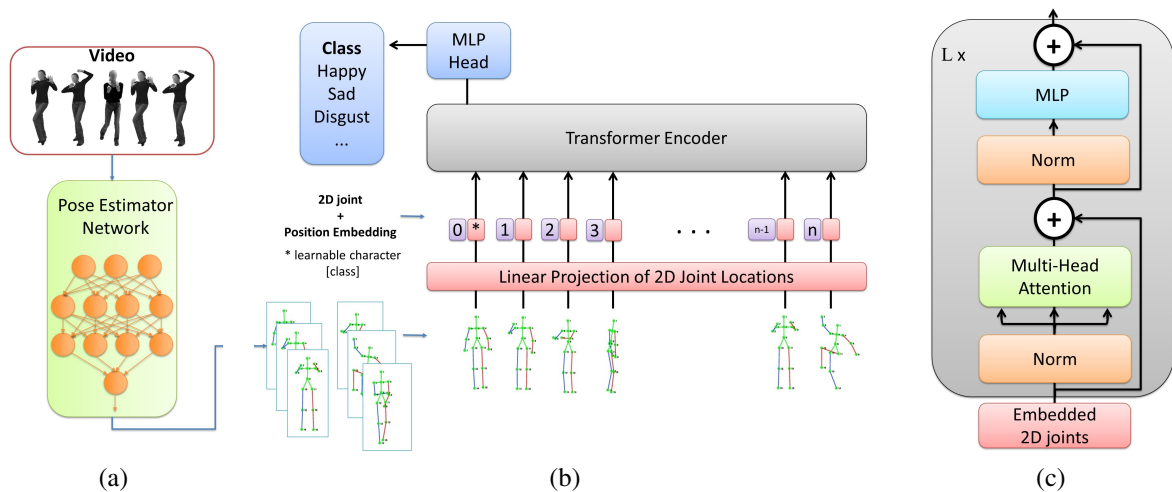


Figure 1: Proposed approach overview: (a) Pose estimation from RGB video, (b) *Emotion Transformer*, and (c) components of the Transformer Encoder.

ever, in both works, the level of precision achieved in the tested dataset is not representative of real-world situations since challenges such as discontinuity, occlusion, and variation in the distance between actor and sensor are not considered. Techniques that ignore the temporal relationship in the FABO dataset have their performance penalized, as in (Yan et al., 2018a), which even considering the hybrid approach reaches an accuracy of 62.60%. The problem is intensified when facial features are of low resolution or not visible, and the system performance drops even further.

Both pose-based and hybrid approaches founded are dealing with reliable joint positions obtained with motion capture or pose estimation in structure environments. Although those methods present a satisfactory performance, it is not representative of in-the-wild scenarios. A major flaw of all methods listed is the lack of evaluation in challenging data, when discontinuity, occlusion, and variation in the distance between human and sensor occur. Further proofs that demonstrate the applicability of BER methods in the real world are needed.

Recently, (Luo et al., 2020) proposed the Body Language Dataset (BoLD) a large-scale body emotion recognition video collection, created from movie scenes and labeled using rigorous procedures. BoLD is a in-the-wild human emotion dataset containing body language annotated and categorical and continuous emotional labels. The author also provides a wild range of baseline methods to evaluate the performance of machine/deep learning models. The first category of methods considers only joint information, previously extracted using pose estimation, and the second uses pixel and temporal information. The author evaluates Laban Movement Anal-

ysis (LMA) (Laban and Ullmann, 1971), a common and well-established way of documenting body movement through effort, shape, and space. With all LMA features combined, each skeleton sequence can be represented by feature vector computing joint relations. Spatial-Temporal Graph Convolutional Networks (ST-GCN) (Yan et al., 2018b), which automatically learn both the spatial and temporal patterns from data using graph convolution. For learning from pixel methods, two-stream models have been compared. A typical model of this type contains two convolutional neural networks taking static images and optical flow as input. One of the approaches uses ResNet (He et al., 2016) as a backbone, called TS-ResNet101. For Temporal Segment Networks (TSN) (Wang et al., 2016). For two-stream inflated 3D ConvNet (I3D) (Carreira and Zisserman, 2017), 3D convolution replaces 2D convolution in the original two-stream network. The authors report TSN as best performance for categorical and dimensional emotions, with a mean  $R^2$  of 0.095, a mean average precision of 17.02%, and a mean ROC AUC of 62.70%.

Taking into account the studies found in recent literature, the following points are evident: (i) the addition of temporal information to emotion prediction models can introduce rich features in the description of emotional states, (ii) simple temporal-aware classifiers largely used in literature cannot ensure high recognition performance, (iii) self-attention models, that allow rich contextual information extraction have not been investigated in the body emotion recognition field, and (iv) majority of current works are not able to take advantage of long video sequences as they cannot extract long-range dependencies.

To address those shortcomings, this paper presents

an emotion recognition method based on body information using Transformers, the state-of-the-art mechanics to process temporal information.

### 3 METHODOLOGY

In this section, we describe the proposed approach used to recognize emotion. It consists of three steps, which are illustrated in Figure 1. First, poses are extracted from videos and represented as skeleton joints. Then, a linear operation converts the 2D joint space into a map projection. Finally, a regular Transformers encoder, containing multi-headed attention layers is used to infer the temporal correlation between frames. The details of each of the steps are described in the remainder of this Section.

#### Body Pose Estimation

Human Pose Estimation (HPE) methods, especially based on deep learning, have made significant progress during the last years. They are used in many applications, such as human-computer interaction, sports analysis, healthcare, and so on (Song et al., 2021). HPE methods localize the spatial location of body keypoints of a person from a given image or video, generating a 2D skeleton representation (Cao et al., 2017). This is done by localizing body joints and grouping them into valid human pose configurations.

The first step of the proposed approach is to use the 2D poses estimated from HPE methods as an input for a transformer classifier (as seen in Figure 1(a)). This is done by processing a given video frame-by-frame to acquire human body landmarks. Any HPE can be used for this task, however, for this work the author choose OpenPose (Cao et al., 2017).

#### Emotion Transformer

Using already established pose estimation methods, for a video containing a person, a sequence of poses can be extracted. This sequence of poses can be used to learn the spatiotemporal relation of body movements and emotions. Unlike the traditional regression-based methods that are limited to short data sequences at the architecture level, our proposed approach is able to take advantage of long video sequences due to the self-attention mechanism. A regular Transformer (Vaswani et al., 2017) takes a sequence as an input and models its long-range dependencies with stacked multi-head self-attention layers

and feed-forward networks. For body emotion recognition, the input is a sequence of poses in the skeletal representation. Formally, for an input video  $V$ , it is first split into a fixed frame-size containing poses  $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N]$  where  $N$  is the number of frames where a pose is found. To maintain a consistent  $N$ , a copy padding frame is performed. The pose vector is mapped to a  $D$ -dimensional pose embedding  $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_N]$  with a linear layer. A linear embedding layer's goal is to map a discrete data sequence to a continuous vector representation. During training, a linear transformation is learned in terms of  $\mathbf{E} = \mathbf{W}^J \mathbf{x} = \mathbf{W}_{(k,\cdot)}^J := \mathbf{z}_{w_j}^J$ , for a regular matrix of weights  $\mathbf{W}$  of size  $N \times J$  ( $J$  is the number of joints). The embedding layer  $\mathbf{E}$  projects the input symbols  $\mathbf{x}$ , which are represented as one-hot vectors, onto a continuous space  $\mathbf{z}$ . Because it allows the model to process the input sequence as a continuous rather than a discrete sequence, the resulting continuous vector representation is more effective for learning.

This final tokenized sequence  $\mathbf{Z}$  is prefixed with an optional learned classification token  $z_{cls}$ , whose representation at the encoder's final layer serves as the final label representation. As the succeeding self-attention procedures in the transformer are permutation invariant, a learned positional embedding  $\mathbf{E}_{pos}$  is additionally added to the tokens to keep positional information. To summarize, the input to the first transformer block is:

$$\mathbf{Z} = [z_{cls}; \mathbf{z}_1; \mathbf{z}_2; \dots; \mathbf{z}_N] + \mathbf{E}_{pos} \quad (1)$$

with  $\mathbf{z} \in \mathbb{R}^D$  and  $\mathbf{E}_{pos} \in \mathbb{R}^{(N+1) \times D}$ .

This first part of the *Emotion Transformer* is illustrated in Figure 1(b). The backbone network of *Emotion Transformer* consists of  $L$  blocks, each of which consists of a multi-head self-attention layer (MSA) and a feed-forward network (FFN). A usual Transformer Encoder (see Figure 1(c)) consists of alternating layers of multiheaded self-attention and MLP blocks. In particular, single-head attention is computed as below:

$$\text{Attn}(Q, K, V) = \text{softmax} \left( \frac{QK^T}{\sqrt{d_k}} \right) V \quad (2)$$

where  $Q, K, V$  are query, key and value matrices respectively, and  $d_k$  is a scaling factor, in the same manner as in a regular Transformer. For more effective attention on different representation subspaces, multi-head self-attention concatenates the output from several single-head attentions and projects it with another parameter matrix:

$$\text{head}_{i,l} = \text{Attn} \left( \mathbf{Z}_l \mathbf{W}_{i,l}^Q, \mathbf{Z}_l \mathbf{W}_{i,l}^K, \mathbf{Z}_l \mathbf{W}_{i,l}^V \right) \quad (3)$$

$$\text{MSA}(\mathbf{Z}_l) = \text{Concat}(\text{head}_{1,l}, \dots, \text{head}_{H,l}) \mathbf{W}_l^O, \quad (4)$$

where  $\mathbf{W}_{i,l}^O, \mathbf{W}_{i,l}^K, \mathbf{W}_{i,l}^V, \mathbf{W}_{i,l}^Q$  are the parameter matrices in the  $i$ -th attention head of the  $l$ -th transformer block, and  $\mathbf{Z}_l$  denotes the input at the  $l$ -th block. The output from MSA is then fed into FFN, a two-layer MLP, and produces the output of the transformer block  $\mathbf{Z}_{l+}$ . Residual connections are also applied on both MSA and FFN as follows:

$$\mathbf{Z}'_l = \text{MSA}(\mathbf{Z}_l) + \mathbf{Z}_l, \quad \mathbf{Z}_{l+1} = \text{FFN}(\mathbf{Z}'_l) + \mathbf{Z}'_l \quad (5)$$

The final prediction is produced by a linear layer taking the class token from the last transformer block  $\mathbf{Z}_L^0$  as inputs.

## Dataset

The dataset used to train the proposed Emotional Transformer architecture is the BoLD (Body Language Dataset) (Luo et al., 2020). It contains 9,876 video clips of acted emotions, primarily through body movements. Each clip contains multiple characters, yielding a total of 13,239 annotations. The dataset has been annotated by crowdsourcing within a total of 26 emotional labels and continuous emotional dimensions. The cataloged emotions are the following: peace, affection, esteem, anticipation, engagement, confidence, happiness, pleasure, excitement, surprise, sympathy, doubt, disconnection, fatigue, embarrassment, yearning, disapproval, aversion, annoyance, anger, sensitivity, sadness, disquietment, fear, pain and suffering. The dataset is split into train/validation/test (80%, 10%, 10%, respectively) using stratified shuffling. The BoLD dataset has built-in OpenPose body format 2D joint positions, containing 18 keypoints for each labeled sample, as seen in Figure 2.

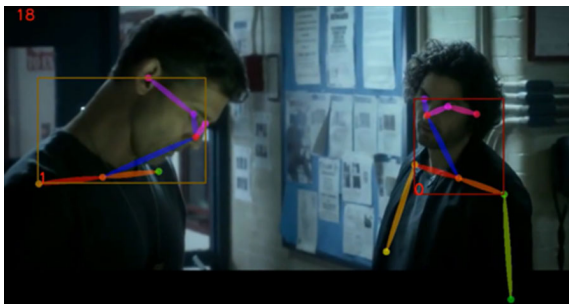


Figure 2: A frame in a video clip sample from BoLD. Body and facial landmarks were detected (indicated with the stick figure).

## 4 EXPERIMENTS AND RESULTS

This section describes the main experiments conducted to study the advantages of using a self-attention model for body emotion recognition. The BoLD dataset (Luo et al., 2020) is used to evaluate the proposed approach and compare the performance of different methods over 26 categories of emotion. Average Precision (AP, area under precision-recall curve) and area under the receiver operating characteristic curve (ROC AUC) are used to evaluate the classification performance. AP is the proportion of the positive samples. ROC AUC measures the ability of a classifier to distinguish between classes. The higher the value of a classifier, the better its ability to distinguish between positive and negative classes; a random baseline for that is 0.5. To compare the performance of different approaches, we report mean average precision (mAP), and mean ROC AUC (mRA). Formally, mAP is given by:

$$mAP = \frac{1}{N} \sum_{i=1}^N \sum_i (R_i - R_{i-1}) P_i, \quad (6)$$

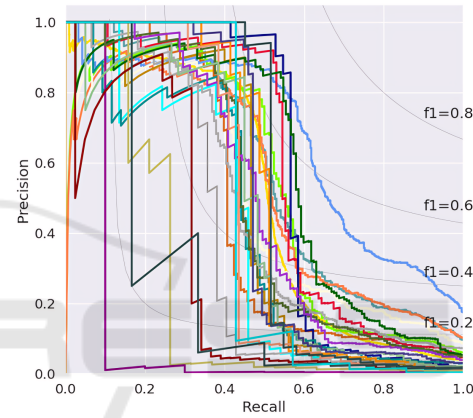
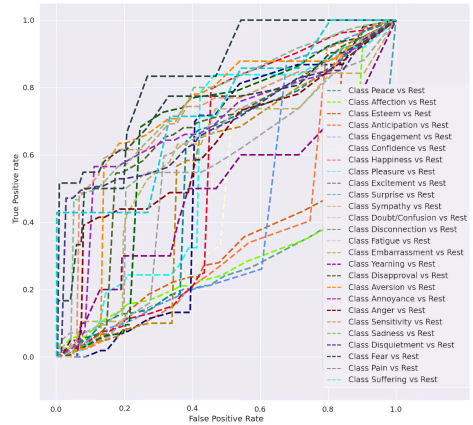
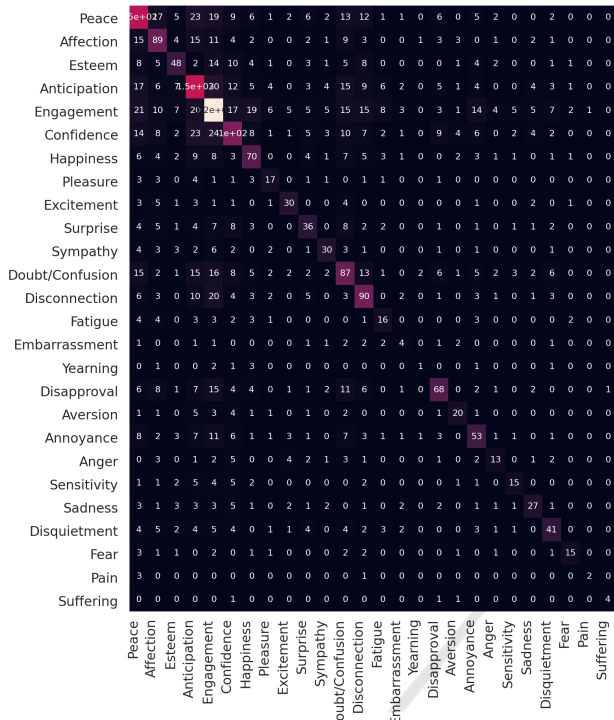
where  $N$  is the number of classes,  $R$  and  $P$  are *Recall* and *Precision* obtained from confusion matrix. For mRA, the method of (Hand and Till, 2001) is used to estimate AUC for multi-class. Let  $\hat{A}(i|j)$  indicate the probability of randomly drawn member of class  $j$  that belong to class  $i$  and  $\hat{A}(j|i)$  the opposite operation, mRA is given by:

$$\hat{A}(i, j) = \frac{1}{2} (\hat{A}(i|j) + \hat{A}(j|i)) \implies mRA = \frac{1}{n} \sum_{i=1}^N AUC_i \quad (7a)$$

for  $c$  equal to the labels of pairs of classes.

### 4.1 Experimental Setup

In the following experiments, we employ the COCO OpenPose skeleton available on the BoLD dataset. The data samples contain a maximum number of 18 two-dimensional keypoints (joints) in regularized 120 frames. The training set is composed of 9222 training samples and 2864 for testing. The remaining 1153 instances are used as validation. We employ the TensorFlow framework to train the proposed network on a server with 380-GB RAM, 2x Intel Xeon(R) Silver 4210 CPU, and 4x Nvidia Quadro RTX 6000 GPU. The training takes approximately 4 hours.



(a) Confusion Matrix, (b) AUC ROC curve (mRA) and (c) mean Precision-Recall curve over classes (mAP). Colors indicate emotional categories.

Figure 3: Multi-class metrics over 26 emotion categories of BoLD obtained from *Emotion Transformer* prediction: (a) Confusion Matrix, (b) AUC ROC curve (mRA) and (c) mean Precision-Recall curve over classes (mAP). Colors indicate emotional categories.

In terms of hyperparameters, we choose an exploratory approach to finding architecture configuration. During the evaluation, activation function, batch size, learning rate, and model size have been tested. The best settings found are the following: numbers of head  $H = 3$ , a linear projection with dimension  $D = 192$ , and a MLP dimension of 256 with  $L = 6$  layers. The total number of trainable parameters is equal to 2.7 million. The AdamW optimization algorithm (Loshchilov and Hutter, 2017) was found to be most suitable among other tested (RMSProp and AdaDelta) and was employed for training with a step drop of the learning rate equal to  $1 \times 10^{-4}$ , which achieved the best model performance.

### 4.2 Emotion Transformer on BoLD Dataset

We experiment on BoLD considering some baselines, common BER architectures, and our proposed model. We report the mAP and mRA to obtain statistically relevant results to evaluate model performances. The baselines chosen for the benchmark are the ones used in the BoLD dataset report, using pose information or RGB values as input. Among those, we compare the Laban Movement Analysis (LMA) (Laban and Ullmann, 1971) and Spatial-Temporal Graph Convolutional Networks (ST-GCN) (Yan et al., 2018b), for the pose-based input. For learning from pixel methods, we compare TS-ResNet101 (He et al., 2016), TSN (Wang et al., 2016), and I3D (Carreira and Zisserman, 2017). The details of the comparator implementation are described in (Luo et al., 2020).

The results of the experimentation are reported in

Table 1: Categorical emotion from gait recognition performance on the BoLD dataset test set.

Model	Classification (%)	
	mAP	mRA
<i>Learning from pixels</i>		
TS-ResNet101	17.04	62.29
I3D	15.37	61.24
TSN	17.02	62.70
<i>Learning from skeleton</i>		
ST-GCN	12.63	55.96
LMA	13.59	57.71
Proposed Approach	<b>42.23</b>	<b>81.63</b>

Table 1. The proposed approach strongly outperforms all of the other methods, with a mean average precision of 42.23%, and a mean ROC AUC of 81.63%. Figure 3 presents detailed metric comparisons over all categorical emotions in the BoLD dataset.

As seen in Figure 3 (b), the *Emotion Transformer* is able to has class separation property for most of the evaluated emotions ( $1.0 \geq ROCAUC > 0.5$ ). Some emotions, such as fatigue and yearning, reported  $ROCAUC < 0.5$  values, due to class imbalance and inherent similarities of movements. The emotions engagement, peace, confidence, happiness and anticipation received the most correct predictions as they were highly distinctive and had sufficient number of samples in the dataset, illustrated in the Figure 3 (a).

The proposed *Emotion Transformer* model demonstrates the potential of Transformer-based architectures. Moreover, robust features from temporal correlations extracted from long video sequences allow for a significant increase in performance over other methods.

## 5 CONCLUSIONS

Emotion recognition using body movement is an emerging area that can bring benefits for healthcare, e-learning, gaming, and social robotics. Furthermore, body movement is still a poorly explored modality, in comparison with facial expressions for emotion recognition. In this paper, we presented *Emotion Transformer*, a 2D body-pose-based self-attention transformer backbone for emotion recognition. This work is a proof of concept research that establishes contextual processing as an essential aspect to consider for body emotion recognition.

Our proposed emotion recognition system achieved significant improvement in a challenging in-the-wild emotion dataset, namely BoLD. Experiments demonstrated that our method obtains superior results, outperforming several baseline methods of

classification. For future work, the effect of the missing body joints and how to efficiently train attention models with missing values can be studied. Multi-modal approach of fusing emotion from the body with facial expressions can also be explored and unbalancing data handling as well.

## ACKNOWLEDGEMENTS

This work is supported in part by FAPESP (São Paulo Research Foundation), grant number 2020/07074-3, by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES)[88881.690185/2022-01] and by the Canada NSERC Discovery Grant on Machine Intelligence for Biometric Security, and by the Canada Defense and Security, IDEaS Innovation Network Establishment Grant AutoDefence. The authors are grateful to the Renato Archer IT Center, for its infrastructure support and MSc. Yajurv Bhatia for his insightful comments and suggestions.

## REFERENCES

- Avola, D., Cinque, L., Fagioli, A., Foresti, G. L., and Marsarone, C. (2020). Deep temporal analysis for non-acted body affect recognition. *IEEE Transactions on Affective Computing*, pages 1366–1377.
- Bhatia, Y., Bari, A. H., Hsu, G.-S. J., and Gavrilova, M. (2022). Motion capture sensor-based emotion recognition using a bi-modular sequential neural network. *Sensors*, 22(1):403.
- Burgoon, J. K., Manusov, V., and Guerrero, L. K. (2021). *Nonverbal Communication*. Routledge.
- Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. (2017). Real-time multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pages 7291–7299.
- Carreira, J. and Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Goodrich, M. A., Schultz, A. C., et al. (2008). Human-robot interaction: a survey. *Foundations and Trends® in Human-Computer Interaction*, 1(3):203–275.
- Hand, D. J. and Till, R. J. (2001). A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine learning*, 45(2):171–186.

- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.
- Ilyas, C. M. A., Nunes, R., Nasrollahi, K., Rehm, M., and Moeslund, T. B. (2021). Deep emotion recognition through upper body movements and facial expression. In *VISIGRAPP (5: VISAPP)*, pages 669–679.
- Jones, R. (2013). *Communication in the real world: An introduction to communication studies*. The Saylor Foundation.
- Laban, R. and Ullmann, L. (1971). *The mastery of movement*. ERIC.
- Lin, T., Wang, Y., Liu, X., and Qiu, X. (2022). A survey of transformers. *AI Open*.
- Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Luo, Y., Gavrilova, M. L., and Wang, P. S. (2008). Facial metamorphosis using geometrical methods for biometric applications. *International Journal of Pattern Recognition and Artificial Intelligence*, 22(03):555–584.
- Luo, Y., Ye, J., Adams, R. B., Li, J., Newman, M. G., and Wang, J. Z. (2020). Arbee: Towards automated recognition of bodily expression of emotion in the wild. *International Journal of Computer Vision*, 128(1):1–25.
- Ly, S. T., Lee, G.-S., Kim, S.-H., and Yang, H.-J. (2018). Emotion recognition via body gesture: Deep learning model coupled with keyframe selection. In *Proceedings of the 2018 International Conference on Machine Learning and Machine Intelligence*, pages 27–31.
- Maret, Y., Oberson, D., and Gavrilova, M. (2018). Real-time embedded system for gesture recognition. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 30–34. IEEE.
- Menolotto, M., Komaris, D.-S., Tedesco, S., O’Flynn, B., and Walsh, M. (2020). Motion capture technology in industrial applications: A systematic review. *Sensors*, 20(19):5687.
- Noroozi, F., Corneanu, C. A., Kamińska, D., Sapiński, T., Escalera, S., and Anbarjafari, G. (2018). Survey on emotional body gesture recognition. *IEEE Transactions on Affective Computing*, 12(2):505–523.
- Rahman, M. W. and Gavrilova, M. L. (2017). Kinect gait skeletal joint feature-based person identification. In *2017 IEEE 16th International Conference on Cognitive Informatics & Cognitive Computing (ICCI\* CC)*, pages 423–430. IEEE.
- Shen, Z., Cheng, J., Hu, X., and Dong, Q. (2019). Emotion recognition based on multi-view body gestures. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 3317–3321. IEEE.
- Song, L., Yu, G., Yuan, J., and Liu, Z. (2021). Human pose estimation and its application to action recognition: A survey. *Journal of Visual Communication and Image Representation*, 76:103055.
- Sun, B., Cao, S., He, J., and Yu, L. (2018). Affect recognition from facial movements and body gestures by hierarchical deep spatio-temporal features and fusion strategy. *Neural Networks*, 105:36–51.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., and Gool, L. V. (2016). Temporal segment networks: Towards good practices for deep action recognition. In *European Conference on Computer Vision*, pages 20–36. Springer.
- Yan, J., Lu, G., Bai, X., Li, H., Sun, N., and Liang, R. (2018a). A novel supervised bimodal emotion recognition approach based on facial expression and body gesture. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, 101(11):2003–2006.
- Yan, S., Xiong, Y., and Lin, D. (2018b). Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*.
- Yang, Z., Kay, A., Li, Y., Cross, W., and Luo, J. (2021). Pose-based body language recognition for emotion and psychiatric symptom interpretation. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 294–301. IEEE.