Emotion Recognition from 3D Motion Capture Data using Deep CNNs

Haris Zacharatos
Department of Computer Science
University of Cyprus
Nicosia, Cyprus
haris@cellock.com

Christos Gatzoulis
School of ICT
Bahrain Polytechnic
Bahrain
christos.gatzoulis@polytechnic.bh

Panayiotis Charalambous
CYENS - CoE
University of Cyprus
Nicosia, Cyprus
p.charalambous@cyens.org.cy

Yiorgos Chrysanthou
CYENS - CoE
University of Cyprus
Nicosia, Cyprus
y.chrysanthou@cyens.org.cy

***Abstract*—Designing computer games requires a player-centered approach. Whilst following guidelines and functional requirement specifications is part of the process, observing and measuring qualities of the players experience is key in providing feedback to game designers. Moreover, it can also be used to create adaptive and personalized experiences for players. With the advancement of affective computing and gaming user interfaces, the opportunity to recognize the player's emotions becomes more feasible and each different modality can offer additional information as affect expression is less defined as compared to action selection. This paper explores the use of 3D skeleton motion data transformed to 2D images that encode pose and movement dynamics to represent annotated emotions. The 2D images are then used to train and test the Inception V3 CNN model on a binary classification emotion recognition between happy and sad emotions. Preliminary results in unseen test data indicate that the above transformation technique can capture emotional information. The paper also discusses future directions that may improve the effectiveness of the proposed method on a wider scale.**

***Index Terms*—emotion recognition from body movements, deep convolutional neural networks**

I. Introduction

As the gaming industry evolves, more sophisticated and natural user interfaces are being introduced, gradually replacing traditional controllers. Such interfaces make extensive use of modalities such as body motion gestures and voice which are nowdays becoming popular in gaming and virtual reality applications. The former, motion data, has the potential to achieve an added level of immersion through the physical embodiment of the player character in real time [1]. This makes it a very strong candidate for an emotion recognition modality, and has already been identified and is being explored by researchers [2] [3] [4] [9]. Deep Learning (DL) has been deployed in computer vision applications offering significantly improved results compared to traditional machine learning techniques. Particularly for human action recognition from motion data, Convolutional Neural Networks (CNNs) have been used extensively due to their high performance success on images or videos tasks [5].

This paper focuses on the classification of emotions from 3D body movements, which are transformed to 2D images, that encode posture and motion dynamics in pixel values. Those images are used as input to retrain the last layers of a pre-trained Deep CNN applying the popular methodology of transfer learning.

II. Related Work

Traditional machine learning techniques have been improving in terms of accuracy but rely on handcrafted features [2] [3] [4] [9]. The use of deep learning techniques to automatically extract effective features from multimodal information and classifications are new directions currently actively pursued by researchers, but several challenges remain in realising an end-to-end deep learning system. With the availability of large datasets, deep learning has become a state-of-the art solution to problems such as emotion recognition. Kim et al. [6] for example propose a CNN-based model for a hierarchical feature representation in the audio-visual domain to recognise spontaneous emotions. Results showed that improvement of recognition accuracy is achieved when hierarchical features and multimodal information are adopted. In another effort, models are constructed from multiple physiological signals collected from sensors placed on the human body by adopting a multimodal deep learning approach so as to improve their performance and reduce the cost of acquiring physiological signals for real world applications [7]. To classify spontaneous multimodal emotional expressions as positive or negative, researchers proposed a cross channel convolutional neural network (CCCNN) having the capability of learning and extracting general and specific features of emotions relying on body motion and face expression [8].These features were further passed to cross-convolution channels to build the cross-modal feature representation.

Deep leaning-based algorithms can be used for feature extraction and classification. With the use of CNNs the work spent on the pre-processing of the images is greatly reduced since the algorithm is already capable of detecting the best features needed to classify the images. Because CNN-based methods cannot reflect temporal variations, researchers used RNN Long Short Term Memory Network (RNN-LSTM) approaches, in which RNN uses gateway units in addition to the common activation function, which extend its memory [10]. Such an architecture allows the network to learn and remember dependencies over more time steps, linking causes and effects remotely [11]. In recent research, an RNN-LSTM was used

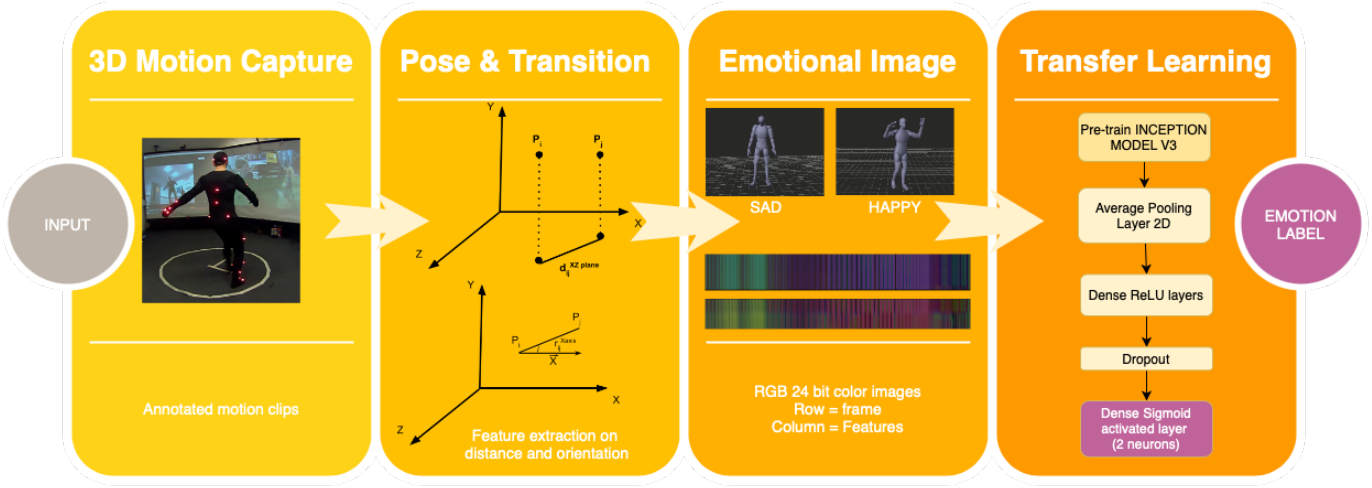

Fig. 1: The overall Architecture

to identify gestures emotion recognition based on low level features inferred from the spacial location and orientation of joints within a track skeleton. [12].

For all the above deep learning approaches, a vast amount of data is needed to perform training and learning. Moreover, encoding raw skeleton data to images and then recognising emotions faces the limitation of a frame by frame representation of emotions. Our method creates features related to time from raw skeleton data and converts them to images.

III. Methodology

The proposed technique is inspired by recent research on action recognition methods that depict skeleton information into image-based representations and create features from 3D skeleton sequences [13]. The feature matrix that is created contains pose and transition dynamics using distance and orientation features.

For the *pose distance feature* within any given frame, the joint-to-joint Euclidean distance for all the joint pairs combinations was calculated by projecting the 3D joint coordinates to the three planes perpendicular to the axes x, y, z in a global coordinate system. The pose distance feature between two joints i and j for a given frame t is given by the below equation:

$$\mathbf{D}_{ij}^{t} = [d_{ij}^{XYplane,t}, d_{ij}^{YZplane,t}, d_{ij}^{XZplane,t}] \tag{1}$$

where:

$$d_{ij}^{XYplane,t} = ||P(i_x, i_y)^t - P(j_x, j_y)^t|| \tag{2}$$

In the above equation, P is the 2D point created from the projection of joint i or j on the XY plane for a given frame t.

In a similar way, the *transition feature* calculates the joint-to-joint Euclidean distance for all possible joint pairs combinations but within two consecutive frames:

$$\mathbf{C}_{ij}^{t} = [c_{ij}^{XYplane,t}, c_{ij}^{YZplane,t}, c_{ij}^{XZplane,t}] \tag{3}$$

where:

$$c_{ij}^{XYplane,t} = ||P(i_x, i_y)^t - P(j_x, j_y)^{t-1}|| \tag{4}$$

observe the difference from calculating distance for frames t and t-1. Two additional features are calculated based on joint-to-joint orientations with respect to the horizontal axes X, Y, Z. Calculating the dot product of each joint-to-joint orientation with each of the 3 axis vectors allows the extraction of the orientation angle from the inverse cosine function. An example is given below for a joint-to-joint vector $\overrightarrow{ij}$ and the X axis vector $\overrightarrow{X}$, where the X notation in the denominator represents multiplication of numbers.

$$r_{ij}^{Xaxis,t} = cos^{-1}\left(\frac{\overrightarrow{ij^t} \cdot \overrightarrow{X}}{||\overrightarrow{ij^t}|| \times ||\overrightarrow{X}||}\right) \tag{5}$$

And below is the vector from all 3 axes for a single pair of joints i and j.

$$\mathbf{R}_{ij}^{t} = [r_{ij}^{Xaxis,t}, r_{ij}^{Yaxis,t}, r_{ij}^{Zaxis,t}] \tag{6}$$

In a similar way, a transition of orientation is calculated across two consecutive frames, with the same formula as above but now vector $\overrightarrow{ij}$ is calculated with joint i from frame t and joint j from frame t-1:

$$\mathbf{G}_{ij}^{t} = [g_{ij}^{Xaxis,t}, g_{ij}^{Yaxis,t}, g_{ij}^{Zaxis,t}] \tag{7}$$

The four features are calculated for all applicable joint pairs and are normalized using min and maximum values to (0,1). They are then concatenated in a row to form a feature set for a given frame. The same process is repeated for each frame starting from frame number 2 and moving further taking into consideration the dynamics with the previous frame 3D joint data. Given this configuration, at the end we have a 2D matrix with every row being the data for each frame and every column representing a feature for a particular pair of joints. This data is then converted to a 2D RGB image.

A. *Emotion image generation*

There are various ways of representing emotions, either by using **distinct emotions** like happiness, sadness, fear, anger, surprise, disgust or by measuring and contextualizing emotions according to a **dimensional space** [16] as illustrated in Figure 2, where emotions are represented in two dimensions of *valence in x axis and arousal in y axis* and each emotion can be viewed in the space defined by these dimensions.

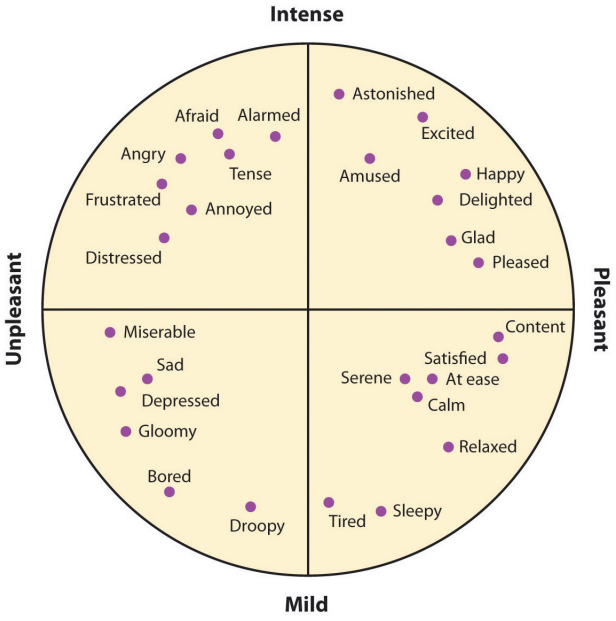

Fig. 2: The Valence-Arousal space

Starting from the hypothesis that motion data can represent emotion information, to prepare motion clips for use in a CNN, we propose the transformation of 3d Spatio-temporal data to pixel data in the form of normalized posture and motion dynamics using an approach that has proven to be successful for action recognition [14]. The posture and motion features are encoded to RGB 24-bit color images. Each row of the image represents a single frame of the clip and each column a different posture or motion dynamics feature as seen in Figure 3.

All clips depict a single skeleton therefore the number of features is the same in all clips, making the width dimension of the image common for all of them. However, since each input motion clip can have a different length in terms of number of frames, the generated images have different sizes with respect to the image's height. To prepare the data for input for the selected pre-trained CNN, images needed to be converted to a standard size. This was achieved by determining the maximum height of all images, witch have variable frames in length and we padded zeros to the remaining images extended pixels.

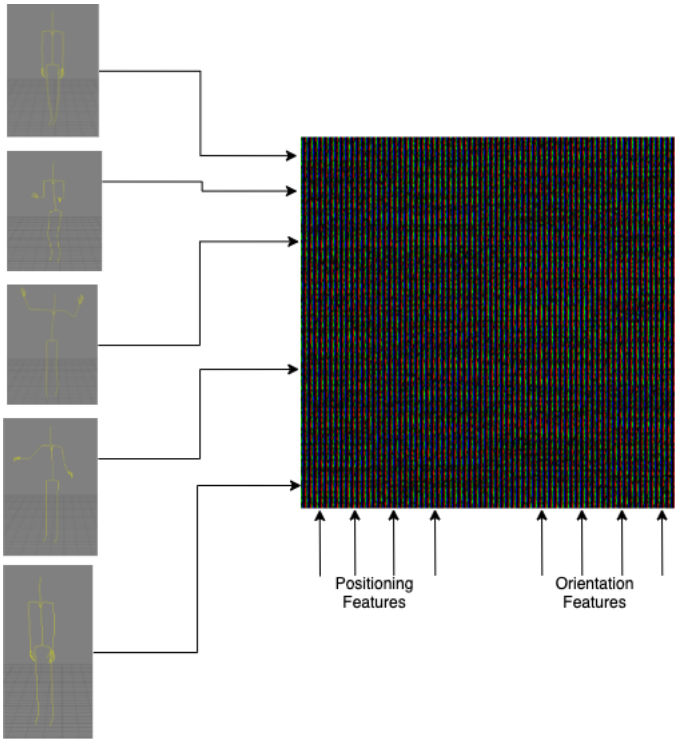

Fig. 3: Image representing a series of postures (rows) with features (columns)

B. *Transfer Learning*

Transfer learning consists of taking features learned on one problem, and leveraging them on a new, similar problem. To address the given classification problem, we used a pre-trained model called Inception V3 [15] which is an image recognition model that has been shown to attain greater accuracy on the ImageNet dataset. The parameters of the Inception module are 24 Million as can be seen in Figure 4. We have removed the last layers of the model adding our own layers, to accommodate our architecture with the total parameters reaching 24.5 million out of which 2.6 million are trainable.

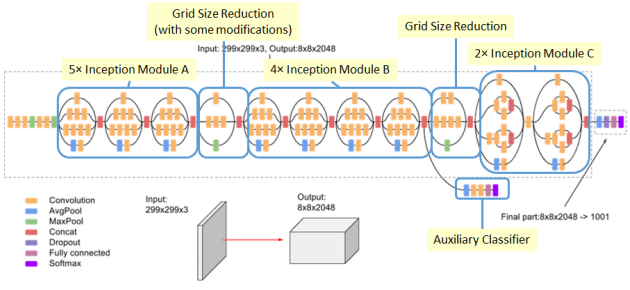

Fig. 4: Inception-V3 model

We use binary cross-entropy as the loss metric as we have 2 target classes (happy and sad). We have added new trainable layers as seen in Figure 5. The new layers contain a global average 2D pooling, then multiple dense RELU activation

layers, and then dropout of 0.3, ending on two neurons for prediction of the targeted two classes.

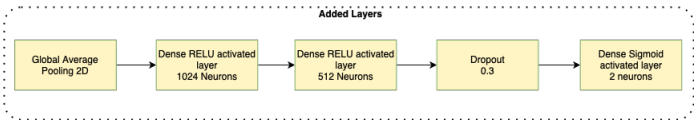

Fig. 5: Added Layers on pre-trained Inception-V3 model

The model learned to convert the existing features into predictions on the new dataset. The summary of the model can be seen in Figure 6. The overall architecture of our emotion recognition method is showed in Figure 1.

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
inception_v3 (Functional)    (None, 83, 322, 2048)     21802784
_________________________________________________________________
global_average_pooling2d (Gl (None, 2048)              0
_________________________________________________________________
dense (Dense)                (None, 1024)              2098176
_________________________________________________________________
dense_1 (Dense)              (None, 512)               524800
_________________________________________________________________
dropout (Dropout)            (None, 512)               0
_________________________________________________________________
dense_2 (Dense)              (None, 1)                 513
=================================================================
Total params: 24,426,273
Trainable params: 2,623,489
Non-trainable params: 21,802,784
_________________________________________________________________
```

Fig. 6: Our Transfer learning model architecture

C. Data Capture

We used an acted emotional body movement dataset [14] in order to execute a pilot test with 2 emotions that differ in both dimensions of the Valence-Arousal space. The dataset contained scenarios to perform a typical and natural expression, captured by the Axis Neuron motion capture system. We have selected scenarios of equal male and female actors and in total we used 208 happiness and 194 sadness different inputs. All the data were setup using 17 body joints with both positional and rotational data; we only consider the positional data.

IV. Results

The network is trained for 30 epochs using a learning rate of 10^{-3}. We used 80% of the input clips for training and 20% for validation. All experiments are implemented on an Intel i9-07920x CPU @ 2.9Ghz, with one NVIDIA GeForce RTX 2080 Ti card.

The training model was tested with an un-seen dataset of 16 motion clips (8 happiness, 8 sadness), which resulted in an average of 81% recognition rate as can be seen in table I

Hapiness	*Sadness*	
7(**88**%)	1	*Hapiness*
2	**6**(**75**%)	*Sadness*

TABLE I: Happiness or Sadness classification using Transfer Learning

V. Conclusions and Future work

Previous studies [4] [12] showcased that movement dynamics can be used for emotion recognition. Up to now we have not seen research contributions in the Affective computing domain, that utilise image representations of pose and movement dynamics from 3D skeleton motion data. This technique has been used with success previously for action recognition [13] and the current project attempts to apply it in the context of emotion recognition. The proposed technique utilizes both posture and motion dynamics to construct image representations of the motion clips. The images are then annotated with the emotion class of their source clips. The current technique shows that combining posture and subsequent frame motion dynamics in an image that uses rows as a temporal dimension and columns as dynamic features can capture affective information. While the initial results are promising, the study needs to be extended to a larger set of emotion classes, to determine how descriptive is the encoding of affect into the produced images. Moreover, new representations of images should be tested, such as those derived from other sets of motion dynamics, for example Laban Movement Analysis features [4] [9]. Further to this, the training data can be enriched with standard data augmentation techniques to potentially improve the classification accuracy. The data augmentation can take place either directly to the skeleton data before the creation of images (noise on joint properties, time warping, autoencoder-based among others) or to the resulting images with traditional image-based data augmentation techniques. Finally, while the current pilot study deployed the Inception V3 model, there are other successful pre-trained CNN models that should be tested and compared in terms of performance.

VI. Acknowledgment

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 739578 and the Government of the Republic of Cyprus through the Deputy Ministry of Research, Innovation and Digital Policy.

References

[1] K. Isbister, R. Rao, U. Schwekendiek, E. Hayward, and J. Lidasan. "Is more movement better?: a controlled comparison of movement-based games", 6th International Conference on Foundations of Digital Games ACM, New York, NY, USA, pages 331-333, 2011

[2] G. Castellano, S. Villalba, and A. Camurri. Recognising human emotions from body movement and gesture dynamics. Affective Computing and Intelligence Interaction, (LNCS 4738):71-82, 2007.

[3] G. Cimen, H. Ilhan, and T. Capin. Classification of human motion based on affective state descriptors. computer animation and virtual worlds, 24(3-4):355-363, 2013.

[4] H. Zacharatos, C. Gatzoulis, Y. Chrysanthou, and A. Aristidou. 2013. "Emotion Recognition for Exergames using Laban Movement Analysis." Motion on Games (MIG '13) ACM SIGGRAPH. 61-66. DOI:https://doi.org/10.1145/2522628.2522651

[5] V. Sze, Y. Chen, T. Yang and J. S. Emer, "Efficient Processing of Deep Neural Networks: A Tutorial and Survey," in Proceedings of the IEEE, vol. 105, no. 12, pp. 2295-2329, Dec. 2017, doi: 10.1109/JPROC.2017.2761740.

[6] Y. Kim, H. Lee, and E. Provost. "Deep learning for robust feature generation in audiovisual emotion recognition". IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 3687-3691, May 2013.

[7] W. Liu, W. Zheng, and B. Lu. Multimodal emotion recognition using multimodal deep learning. Cornell University - CoRR, 2016

[8] P. Barros, C.Weber, and S.Wermter. Emotional expression recognition with a cross channel convolutional neural network for human-robot interaction. IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids), pages 582-587, 2016.

[9] Aristidou, A. and Charalambous, P. and Chrysanthou, Y., "Emotion Analysis and Classification: Understanding the Performers' Emotions Using the LMA Entities, Computer Graphics Forum, V34,6, pp.262-276 DOI: https://doi.org/10.1111/cgf.12598

[10] D. Avola, M. Bernardi, L. Cinque, G.L. Foresti, and C. Massaroni. Exploiting recurrent neural networks and leap motion controller for the recognition of sign language and semaphoric hand gestures. IEEE Transactions Multimedia, 21:234-245, 2018.

[11] B. Hermans, M.; Schrauwen. Training and analysing deep recurrent neural networks. Advances in Neural Information Processing Systems, 1:190-198, 2013.

[12] T. Sapinski, D. Kaminska, A. Pelikant, and G. Anbarjafari. "Emotion recognition from skeletal movements". Entropy, 21(7):646, 2019.

[13] H. Thien, H. Cam-Hao,N. Trung-Thanh, K., Dong-Seong. (2019). Image Representation of Pose-Transition Feature for 3D Skeleton-Based Action Recognition. Information Sciences. 513. 10.1016/j.ins.2019.10.047.

[14] Goldberger, A. et.al (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. 215-220.

[15] Szegedy, C. and Vanhoucke, V. and Ioffe, S. and Shlens, J. and Wojna, Z. (2015). Rethinking the Inception Architecture for Computer Vision. 1512.00567, arXiv, cs.CV

[16] Russell, J. A. (1980). A circumplex model of affect. Journal of personality and social psychology, 39(6), 1161.