Towards Recognizing Emotion in the Latent Space

Marios Fanourakis
SIMS group
University of Geneva
Geneva, Switzerland
marios.fanourakis@unige.ch

Guillaume Chanel
SIMS group
University of Geneva
Geneva, Switzerland
guillaume.chanel@unige.ch

Rayan Elalamy
SIMS group
University of Geneva
Geneva, Switzerland
rayan.elalamy@unige.ch

Phil Lopes
Immersive Interaction Group
EPFL
Lausanne, Switzerland
phil.lopes@epfl.ch

***Abstract*—Emotion recognition is usually achieved by collecting features (physiological signals, events, facial expressions, etc.) to predict an emotional ground truth. This ground truth is arguably unreliable due to its subjective nature. In this paper, we introduce a new approach to measure the magnitude of an emotion in the latent space of a Neural Network without the need for a subjective ground truth. Our data consists of physiological measurements during video gameplay, game events, and subjective rankings of game events for the validation of our model. Our model encodes physiological features into a latent variable which is then decoded into video game events. We show that the events are ranked in the latent space similarly to the participants' subjective ranks. For instance, our model's ranking is correlated (Kendall τ of 0.91) with the predictability rankings.**

***Index Terms*—affective computing, video games, emotion recognition, neural networks, appraisal theory, arousal, valence, emotional dimensions**

I. Introduction and Related Work

One of the main motives for automatic emotion recognition has been the improvement of human computer interaction (HCI) by affording machines human-like abilities to better anticipate and adapt to their operators behaviours and needs. Both Cowie et al. [3] and Fragopanagos et al. [5] describe the challenges and opportunities in this endeavour covering not only the need for machines to recognize human emotions but also how machines can influence human emotions.

Since then, there has been an abundance of research literature on the topic of automatic emotion recognition using physiological signals, a topic which is still very active [6], [16]. Affective gaming is an exciting sub-field of HCI where the emotions of video game players are detected and analyzed in the context of gaming. Video games offer a high level of immersion and can elicit a wide range of emotions, making it a popular tool in emotion research [7], [10]. In the literature, emotion recognition in affective gaming (and in general) is achieved, almost invariably, by utilizing various *supervised* learning techniques which require inputs of features like physiological signals, events, facial expressions, etc. The targeted ground truth can be discrete (happy, sad, angry, etc.) or continuous (arousal, valence, etc.) [7], [8], [10], [16]. More recently, deep learning techniques have also made their impact in affective gaming [1].

The quality and reproducibility of the resulting models is closely tied to the quality of the ground truth labels [2]. In general, there are a few common methods to acquire ground truth data: expert annotations, crowd-sourced annotations, self-reports, induced emotions. These different methods to acquire the ground truth make the models difficult to meaningfully compare. Most are unavoidably subjective in nature since the verbalized/communicated emotion does not necessarily reflect the true underlying emotion of the subject [13]. They also depend on the capacity of an individual to assess their own or an other's emotional state [9], [12]. Consequently, these methods only provide a crude approximation of the ground truth.

To create models that better capture the emotion of an individual we must take a step back to look at where emotions come from. Moors [11], does an invaluable comparison of the different theories and concludes that there is much agreement that emotions stem from a combination of component processes. We will focus on a specific component-process representation called appraisal theory that is now well established [14]. The general appraisal process starts with an event which is subsequently appraised and weighed against various criteria which together regulate the specific emotion. In a recent analysis, Scherer and Moors [15], bring to light problems in emotion research like the use of discrete emotions despite that emotions are often a combination of different components with varying amplitudes in a continuous space. Although correlated, internal emotional states, experienced feelings, and expressed feelings are not one and the same. The emotional states are mainly dependent on the various appraisal criteria, the experienced feelings are the conscious representation of an amalgamation of the internal emotional states, the expressed feelings are a further modulation of the experienced feelings based on sociocultural norms and interpersonal relationships.

Recent work of Yannakakis et al. [17] make strong arguments for an ordinal approach to measuring and analyzing emotions. Emotions and the subsequent subjective feelings towards an event are not absolute, they are experienced in relation to the emotions of previous events.Yannakakis et al. show the validity, reliability, and robustness across domains of an ordinal approach to measuring emotions. Hence, we also adopt this approach in our work.

In our work we attempt to recover the internal emotional state by creating a machine learning model that does not rely on subjective feelings. We do not yet make any recommenda-

This work was funded by Innosuisse.

tions for the specific layers and components best suited for this task but rather propose a general architecture that resembles parts of the appraisal process.

We show that it is feasible to indirectly measure the amplitude of an emotion in the latent space of a Neural Network without the need for a ground truth of emotion.

Our experimental data consists of physiological signals collected from video game players while they are engaged in gameplay. We create a model with physiological features as input and specific events in the game as targets using a very basic neural network with an encoder/decoder architecture. We then look at how the model's latent/learned variables relate to emotion, therefore avoiding the use of subjective data during model training. To our knowledge this is a new approach to emotion recognition and we hope that our results encourage further development in this direction.

II. Model Architecture

We designed our overall architecture in such a way that it resembles a generic appraisal model where an event goes through an appraisal process (represented in the encoder) which produces an internal representation of emotional states (latent space) that are translated into a measurable physiological reaction (decoder). This is illustrated in the upper part of Figure 1. The encoder and decoder components can include any of the modern neural network layers and their combinations. Our main assumption is that if the inputs and outputs of the model are strongly linked via internal emotional states, then the latent space will capture these states.

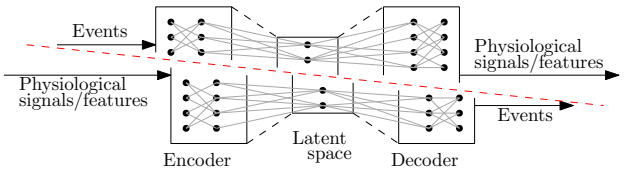

Fig. 1: The proposed general neural network architecture.

Since our goal is to detect the emotional state from physiological signals we invert the model and use physiological signals on the input side and the events on the output like the lower part of Figure 1.

III. Data and Implementation

In our experiment, pairs of players were asked to play a round of 1 vs 1 deathmatch using the Xonotic computer video game. Xonotic is an open source fast paced first person shooter similar to *Quake 4*. The goal of the deathmatch gamemode is to be the first to get 10 frags (kills), the player respawns within a few seconds after each death. There are several items scattered in the environment that the player can pick up and which are replenished after a short time. Several types of game events were automatically recorded during gameplay: weapon pickup, armor pack pickup, damage boost pickup, health pack pickup, health boost pickup, deal damage, die (killed by enemy), suicide (death caused by self damage), kill, receive damage. We also recorded the electrocardiogram (ECG), electrodermal activity (EDA), and respiration of the players using a Bitalino device. Immediately after the participants finished their gameplay session, they were asked to rank a list of game events in terms of four emotional dimensions [4]: arousal, valence, control, and predictability (four different **ranking tasks** of the game events per participant).

In total, we collected physiological data from 19 dyads (38 participants). We visually inspected the signal quality for each participant and discarded a participant's physiological data if the quality of either the EDA or ECG was not good. This left us with physiological data for 19 participants. All 38 rankings for each ranking task remained valid.

A. Participant ranking analysis

To show the disparity between the participants' subjective rankings of the game events, we compared them by calculating the Kendall τ rank correlation between all pairs of ranking in each ranking task. The results of this comparison are shown in Figure 2, where we calculate a histogram of Kendall τ values (in the range of -1 to $+1$). As expected, the participants did not perfectly agree. The median/mean Kendall τ for arousal, valence, control, and predictability were $0.42/0.44$, $0.56/0.52$, $0.42/0.44$, and $0.38/0.34$ respectively.

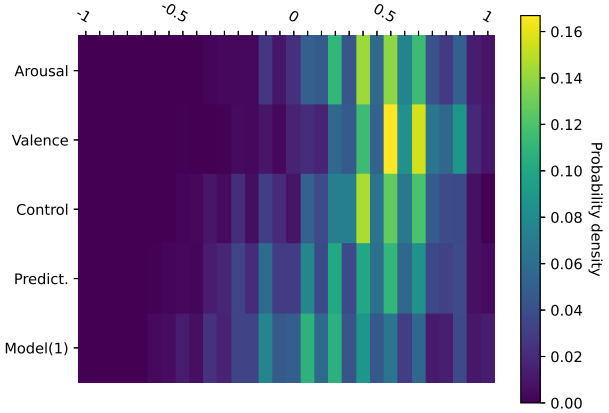

Fig. 2: Kendall τ rank correlation histogram.

Seeing how the participant rankings are correlated, we wanted to capture a global representative ranking for each ranking task. To achieve this, we used the Schulze Condorcet method to combine all rankings into one for each ranking task. The resulting rankings are shown in Table I.

Next, we wanted to see which events had the highest agreement among participants in each ranking task. For two events i, j, we can count how many pairs of rankings have them in the same relative order (number of concordances, $conc_{i,j}$) or disagree on their relative order (number of discordances, $disc_{i,j}$). Then we can calculate a concordance score, cs_i, for each event, i, using the following equation:

$$cs_i = \frac{\sum_j conc_{i,j}}{\sum_j conc_{i,j} + \sum_j disc_{i,j}} = \frac{\sum_j conc_{i,j}}{\binom{R}{2}(J-1)} \quad (1)$$

Where R is the total number of rankings for the ranking task, and J is the total number of events to rank. The results from

TABLE I: Condorcet rankings of each ranking task. Game events with an asterisk are item pickups.

	armor*	dmgboost*	deal dmg	die	healthboost*	health*	kill	receive dmg	suicide	weapon*
Arousal	9th	6th	4th	3rd	7th	10th	1st	5th	2nd	8th
Valence	6th	5th	9th	2nd	8th	7th	10th	3rd	1st	4th
Control	9th	5th	6th	2nd	8th	10th	4th	3rd	1st	7th
Predictability	8th	6th	5th	2nd	7th	10th	3rd	4th	1st	9th
Model	7th	6th	5th	2nd	8th	9th	3rd	4th	1st	10th

our data are shown in Figure 3, where we notice that the *'suicide'* game event has the highest concordance score among all ranking tasks except arousal, followed by the *'die'* game event. This tells us that the rank given to the *'suicide'* game event had the highest agreement between participants.

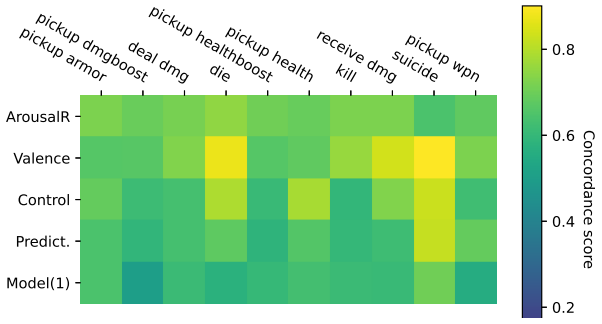

Fig. 3: Concordance score of events for each emotional dimension.

B. Model implementation and training

In our implementation we chose to use the most simple and basic components of a neural network as a starting point. To achieve this, our inputs consisted of pre-processed physiological features calculated from 15 second rolling windows (13 second overlap) of the filtered signals. The EDA signal was filtered using a low-pass Butterworth filter of order 4 with a cutoff frequency of $5Hz$. The ECG signal was filtered using a FIR bandpass filter of order 33 with a low frequency cuttoff of $3Hz$ and high frequency cutoff of $45Hz$. Then the heart rate (HR) was calculated from the filtered ECG signal. The input features to the model were the mean and variance of the standardized (per participant) HR, and the mean and variance of the derivative of the filtered EDA signal. The targets consisted of a multi-hot encoded vector of the events where an event was 'hot' (had a value of 1) if there was at least one occurrence of that event in the first 8 seconds of the main 15 seconds window and 0 otherwise. The reason for including events only in the first 8 seconds of the main 15 second window is to ensure that we do not include events which have not yet influenced the physiological signals in the main window, and to also include as much of the physiological response as feasible for the events in that window.

Both our encoder and decoder layers consisted of a fully connected layer with no bias. The latent space had a single dimension. The full implemented model is shown in Figure 4.

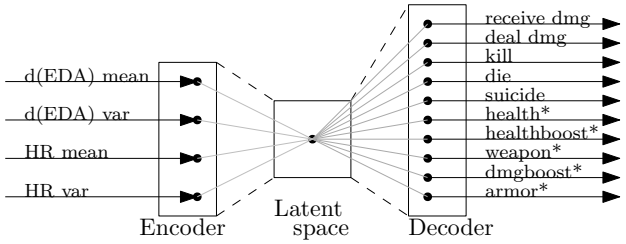

Fig. 4: Implemented model. Events with an asterisk are item pickups.

We trained 50 models with different training and validation sets. We split the data by dyads - for each model 70% were chosen randomly (without replacement) and the remaining 30% were used for validation. We used a weighted binary cross entropy loss function. For each model, the weights for the loss function were calculated from their respective training sets. We did not use a test set since we were not interested in how well the model could predict the game events from the physiological signals and to maximize the data that would go into the training and validation. Despite the lack of a test set, we did pay attention that the loss of the model decreases with each epoch indicating that the model is learning.

IV. Results and Discussion

For each model we can rank the game events according to how sensitive their corresponding neurons are to the latent variables. This can be achieved by entering a range of values to the input of the decoder. In our case, since our decoder layer consists of a single linear layer with no bias we could use the decoder weights directly to order the events. For example, if the events [kill, die, pickup armor] have weights [-2,5,1], then the rank would be [kill, pickup armor, die]. We compared the ranking correlation between all model ranks (50 in total) in the same manner as with the participant rankings. The median and mean Kendall τ were 0.022 and 0.037 respectively indicating that there is no correlation between the models. Upon inspection, it was clear that many of the model rankings were reversed (had negative Kendall τ). This was due to the nature of the architecture we used which allowed for the weights of the encoder and decoder to both be multiplied by -1 with no effect to the training performance. To solve this issue, we anchored the *'suicide'* game event to one side of the ranking, i.e. we reversed ranks that were deemed to be flipped with respect to that event. Our choice of this event was based on the event concordance score from the participants' rankings (see Figure 3) where it had a higher score overall. In Figure 2,

we see that the models correlation (post anchoring) increased significantly but still remains weak with the median and mean Kendall τ being 0.2 and 0.24 respectively.

The weak correlation between the models tells us that this particular neural network architecture cannot easily distinguish emotions given the particular physiological features that we used. Another factor has to do with the individuality of the appraisal process. Our training data included features from multiple participants resulting in conflicting information.

A. Model Condorcet

Like with the participant rankings, we computed the models' Condorcet ranking so that we can easily compare with the participant Condorcet rankings. It is included in Table I. In Figure 5 we show the Kendall τ correlation of the Condorcet rankings of the emotional dimensions (participant data) along with the Condorcet ranking of the models.

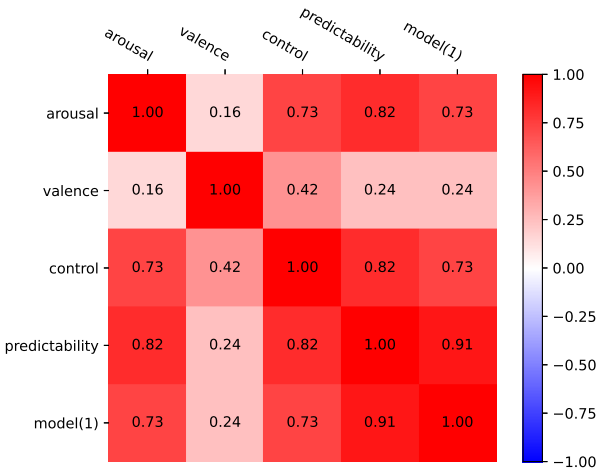

Fig. 5: Rank correlation between the Condorcet ranks.

Arousal, control, and predictability are highly correlated with each other while valence is not significantly correlated with the rest. The model ranking is most correlated with predictability but is also significantly correlated with arousal and control. This leads us to believe that the model successfully captured the most significant emotional factor that relates the physiological signals to the game events.

V. Conclusion and Future Work

In this paper, we wanted to reveal a path towards automatically recognizing the internal emotional state of a person parallel to the current field of automatic emotion recognition which focuses on expressed subjective feelings. Our approach is to create a neural network architecture which closely mirrors the well established appraisal processes.

Our model learned an internal relation in its latent space by only using physiological features and the occurrences of the game events themselves. The latent space was found to be highly correlated with the emotional dimension of predictability based on a comparison of the aggregate participant rankings of events and the aggregate model rankings of events. This leads us to conclude that our model has learned to recognize an internal emotional state using physiological features.

It is important to note that there are challenges in this analysis that must be addressed. The use of engineered features in the input is a limitation on the learning capacity of a neural network since important information about the signal is lost. The current model's latent space is not sign consistent since all the weights can be multiplied by -1 without affecting the loss. The architecture lacks the capacity to learn non-linear relationships and also lacks memory. Both of these are important in the appraisal process and should be incorporated in the model. We must incrementally integrate more complex components into our architecture that retain or improve the interpretability of the latent space.

References

[1] Chanel, G., and Lopes, P. User Evaluation of Affective Dynamic Difficulty Adjustment Based on Physiological Deep Learning. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (2020).

[2] Constantine, L., and Hajj, H. A survey of ground-truth in emotion data annotation. In *2012 IEEE International Conference on Pervasive Computing and Communications Workshops* (mar 2012), no. March, IEEE, pp. 697–702.

[3] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine 18*, 1 (2001), 32–80.

[4] Fontaine, J. R., Scherer, K. R., Roesch, E. B., and Ellsworth, P. C. The World of Emotions is not Two-Dimensional. *Psychological Science 18*, 12 (dec 2007), 1050–1057.

[5] Fragopanagos, N., and Taylor, J. G. Emotion recognition in human-computer interaction. *Neural Networks 18*, 4 (2005), 389–405.

[6] Jerritta, S., Murugappan, M., Nagarajan, R., and Wan, K. Physiological signals based human emotion recognition: A review. *Proceedings - 2011 IEEE 7th International Colloquium on Signal Processing and Its Applications, CSPA 2011* (2011), 410–415.

[7] Karpouzis, K., and Yannakakis, G. N. *Emotion in Games*, vol. 4 of *Socio-Affective Computing*. Springer International Publishing, Cham, 2016.

[8] Kivikangas, J. M., Chanel, G., Cowley, B., Ekman, I., Salminen, M., Järvelä, S., and Ravaja, N. A review of the use of psychophysiological methods in game research. *Journal of Gaming & Virtual Worlds 3*, 3 (sep 2011), 181–199.

[9] MacCann, C., and Roberts, R. D. New paradigms for assessing emotional intelligence: Theory and data. *Emotion 8*, 4 (2008), 540–551.

[10] Madeira, F., Arriaga, P., Adriao, J., Lopes, R., and Esteves, F. Emotional gaming. In *Psychology of gaming*, Y. Baek, Ed. Nova Science Publishers, Inc., New York, New York, USA, 2013, pp. 11–29.

[11] Moors, A. *Theories of emotion causation: A review*, vol. 23. 2009.

[12] Parker, J. D., Saklofske, D. H., Shaughnessy, P. A., Huang, S. H., Wood, L. M., and Eastabrook, J. M. Generalizability of the emotional intelligence construct: A cross-cultural study of North American aboriginal youth. *Personality and Individual Differences 39*, 1 (2005), 215–227.

[13] Sander, D., Grandjean, D., and Scherer, K. R. A systems approach to appraisal mechanisms in emotion. *Neural Networks 18*, 4 (2005), 317–352.

[14] Scherer, K. R. Appraisal Theory. In *Handbook of Cognition and Emotion*. 2005.

[15] Scherer, K. R., and Moors, A. The Emotion Process: Event Appraisal and Component Differentiation. *Annual Review of Psychology 70*, 1 (jan 2019), 719–745.

[16] Shu, L., Xie, J., Yang, M., Li, Z., Li, Z., Liao, D., Xu, X., and Yang, X. A review of emotion recognition using physiological signals. *Sensors (Switzerland) 18*, 7 (2018).

[17] Yannakakis, G. N., Cowie, R., and Busso, C. The Ordinal Nature of Emotions: An Emerging Approach. *IEEE Transactions on Affective Computing 3045*, c (2018), 1–1.