Modeling Emotions as Latent Representations of Appraisals

Fanourakis, Marios Aristogenis; Elalamy, Rayan; Chanel, Guillaume

How to cite

FANOURAKIS, Marios Aristogenis, ELALAMY, Rayan, CHANEL, Guillaume. Modeling Emotions as Latent Representations of Appraisals. In: 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW). Nara, Japan. [s.l.] : IEEE, 2021. p. 1–7. doi: 10.1109/ACIIW52867.2021.9666198

This publication URL: https://archive-ouverte.unige.ch//unige:160177
Publication DOI: 10.1109/ACIIW52867.2021.9666198

Modeling Emotions as Latent Representations of Appraisals

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

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

Guillaume Chanel
SIMS group
University of Geneva
Geneva, Switzerland
guillaume.chanel@unige.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, however, is subjective and not always an accurate representation of the emotional state of the subject. In this paper, we show that emotion can be learned in the latent space of machine learning methods without relying on an emotional ground truth. Our data consists of physiological measurements during video gameplay, game events, and subjective rankings of game events for the validation of our hypothesis. By calculating the Kendall τ rank correlation between the subjective game event rankings and both the rankings derived from Canonical Correlation Analysis (CCA) and a simple neural network, we show that the latent space of these models is correlated with the subjective rankings even though they were not part of the training data.**

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

I. Introduction

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. [1] and Fragopanagos et al. [2] 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 [3], [4]. 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 [5], [6].

In the literature, emotion recognition is almost invariably achieved by utilizing various *supervised* learning techniques which require inputs of features derived from various modalities like physiological signals, events, and facial expressions. The targeted ground truth can take discrete values (happy, sad, angry, etc.), continuous (arousal, valence, etc.), or ordinal [4]–[7]. More recently, deep learning techniques have also made their impact in affective computing [8]–[11].

The quality and reproducibility of the resulting models is closely tied to the quality of the ground truth labels [12]. In general, there are a few common methods to acquire ground truth data: expert annotations of emotions, crowd-sourced annotations, self-reported emotions, and inducing desired emotions. These different methods to acquire the data make the models difficult to 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 and others' emotional state [14], [15]. Consequently, emotion annotations only provide an approximation of the emotion ground truth.

In our work we attempt to train a machine learning model to learn an emotional latent space without relying on subjective feelings or other subjective ground truth. We promote the use of a general architecture that resembles parts of the emotional appraisal process and we use subjective annotations only to evaluate and help interpret the model. Our approach is inspired by Sander et al. [13] whose work makes links between artificial neural network architectures and appraisal processes.

II. Related work

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 [16], 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 [17]. Emotion emerges from a complex system of appraisal components which are triggered by events (stimuli). The general appraisal process starts with an event which is subsequently appraised and weighed against various criteria which together regulate the emotional state. Emotions are therefore tightly linked with these events and their evaluation. In a recent analysis, Scherer and Moors [18], 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. They present evidence that the autonomous nervous system (ANS) responses better correlate with appraisal criteria like

978-1-6654-0021-3/21/$31.00

novelty, goal relevance, and valence. 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, and the expressed feelings are a further modulation of the experienced feelings based on sociocultural norms and interpersonal relationships. Consequently, the objectively measurable components of the appraisal process are the autonomic physiological response, the motor expressions, and the event which caused them. Notably, the internal emotional state is not objectively measurable, we can only get a subjective measure through the verbalized subjective feelings of a subject [13], [18]–[20]. It is important to point out that the vast literature on emotion recognition has been mainly focused on recognizing these subjective feelings which are often confounded with emotions [3], [4].

The success of deep learning techniques on many complex problems has made it an attractive option for multi-modal emotion recognition. Deep learning architectures have also opened the possibility to learn latent representations of many types of data and we see their utility in extracting meaningful features or bounding the latent space of machine learning processes [21]–[24].

Using appraisal theory as an inspiration, in this work we show that the latent variables of a process which maps stimulus events to physiological reactions or behaviours with an adequate bottleneck, are correlated with the subjective feelings of subjects despite that data about their subjective feelings was not used in the training process. This observation is in-line with what we expect when modeling emotions under appraisal theory, and our results may also provide some experimental evidence for certain aspects of appraisal theory. Furthermore, this property may be useful to train models to infer emotions without a ground truth. To our knowledge this is a new approach to emotion recognition and we hope that our results encourage further development in this direction.

III. Emotion Model Architecture

Our specific approach is inspired by the appraisal process as Sander et al. present it in a 2005 publication [13]. In this process, an external event is input to various appraisal processes which in turn elicit motor behaviours and autonomous nervous system (ANS) responses. We consider these elicited behaviours and responses as the output of the system. We hypothesize that since the events are linked to the responses via the appraisal processes, then a machine learning model which is trained to recognize the responses from the events or vice versa will inevitably capture a representation of the appraisal criteria in its latent space. We present a generic such machine learning model architecture in Figure 1 where we seek to recognize events from physiological features such as Electrodermal Activity, Heart Rate, Facial expressions, etc. Thus allowing us to extract a representation of the appraisal criteria from these physiological features.

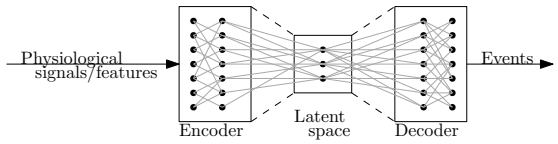

Fig. 1: The proposed architecture.

To validate our approach we must compare with a ground truth, however, the closest we can get to a ground truth is by using the subjective feelings of subjects. We choose to use the emotional dimensions of Arousal, Valence, Control, and Predictability since they are important emotional dimensions [25]. Recent work of Yannakakis et al. [26] makes 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. We do not expect to get perfect correlation with a subjective ground truth using our proposed approach, however we do expect that correlations will be significantly higher than zero.

IV. Data Collection

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 3*. 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. Ten 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. 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 by using a Hamilton segmenter to find the R-peaks.

Immediately after the participants finished their gameplay session, they were asked to complete 4 ranking tasks (see section IV-A), one for each of the emotional dimensions [25] of arousal, valence, control, and predictability. The subjects had to rank the ten game events according to each emotional dimension.

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 one of the following criteria was true: file missing, file corrupted, signal has value 0 (disconnected electrodes or poor electrode contact) for more than 50% of its length, signal has saturated to the maximum value for more than 50% of its length, ECG R peaks are not visually distinguishable from the noise for more than 50% of its length. This left us with physiological data for 19 participants. These criteria were not applicable to the ranking questionnaire, so all 38 rankings for each ranking task remained valid.

A. Participant ranking tasks and ranking analysis

The participants were asked to rank the 10 game events using four ranking questions. Each question addressed one emotional dimension. The questions were asked in the same order as listed below:

1) Rank your ability to keep control during the following events. Please rank (from top to bottom) each event from: "I feel in total control" to "I do not feel I have any control".
2) How pleasurable were the following events during play? Rank each event (top to bottom) from: "This event was positive for me" to "This event was negative for me".
3) How emotionally active were you during the following presented events? Rank each event (from top to bottom) from: "I felt calm" to "I felt excited".
4) How predictable were the following events? Rank each event (top to bottom) from: "I predicted this event" to "I did not predict this event".

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 participant rankings within 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. Moreover, we noticed that the correlations of the Arousal rankings had a bimodal distribution which prompted us to investigate further. The reason for this discrepancy was that the wording of the arousal ranking task asked to rank the events from less to more arousing as opposed to from more to less like the other tasks, creating some confusion. We corrected this problem by reversing the rankings that were significantly (p-value less than 0.05) negatively correlated. Another solution was to flip all rankings where a specific game event (ex. killing an enemy) was on the "wrong" side, but this approach introduces experimenter bias and thus we continued with the previous approach which flipped less of the rankings and was thus more conservative. The mean Kendall τ for the corrected arousal, valence, control, and predictability rankings were 0.41, 0.50, 0.38, and 0.31 respectively. In the remainder of this paper, the corrected arousal rankings were used.

We also compared the participant rankings between ranking tasks. A summary is available in Table I. We notice that the rankings for Arousal, Control, and Predictability are somewhat correlated with each other. Valence and Control are more strongly correlated but Valence is also somewhat correlated with Predictability. Finally, Control and Predictability are strongly correlated within the participants rankings. Therefore, we can expect that comparisons between these ranking tasks and any models we design will show similar relationships. Namely, we expect to see latent variables which will correlate either with Arousal, Control, and Predictability, or with Valence, Control, and Predictability.

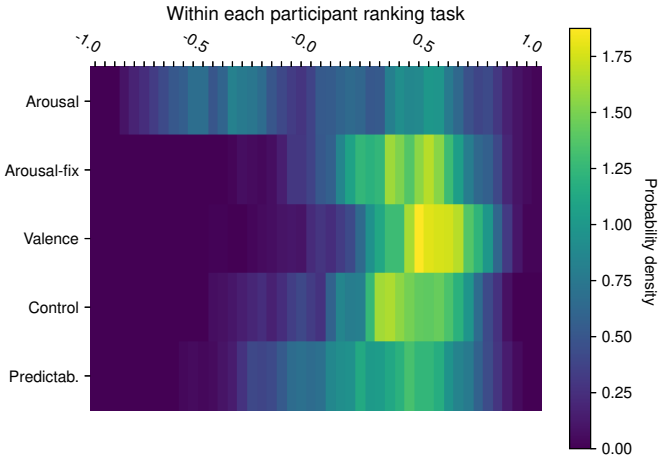

Fig. 2: Kendall τ rank correlation histograms within each of the ranking tasks.

TABLE I: Kendall τ correlation median (M), mean (μ), variance (σ^2) between ranking tasks.

	Valence	**Control**	**Predict.**
Arousal	$M = 0.1429$ $\mu = 0.1664$ $\sigma^2 = 0.0916$	$M = 0.3333$ $\mu = 0.3076$ $\sigma^2 = 0.0984$	$M = 0.3333$ $\mu = 0.265$ $\sigma^2 = 0.132$
	Valence	$M = 0.5238$ $\mu = 0.4585$ $\sigma^2 = 0.0872$	$M = 0.3333$ $\mu = 0.3383$ $\sigma^2 = 0.0878$
		Control	$M = 0.4286$ $\mu = 0.3867$ $\sigma^2 = 0.100$

V. Feature Extraction

For our analysis we tried using engineered features and learned features. The features and event targets were calculated from a 15 second window. Our choice of window size was empirically chosen such that as much of the physiological response to an event is present within the window as possible.

The engineered features consisted of the mean and variance of the standardized (per participant) heart rate (HR) and the mean and variance of the derivative of the electrodermal activity (EDA).

The learned features were extracted using a convolutional autoencoder. The architecture of this network is described in Table II. We used physiological data of good quality from the 19 participants as described in section IV and trained this network with 70% of the data and validated it with the rest

(folds were generated without considering participants). We optimized on the mean squared error loss using an Adam optimizer with a learning rate of 0.001. In each window, we

TABLE II: Convolutional autoencoder architecture

Layer	Type	Params	Activation	Output size
Enc-1	Input	-	-	1x1500
Enc-2	Conv1D	k=201, s=2, 1x16	ReLU	16x750
Enc-3	Conv1D	k=101, s=2, 16x8	ReLU	8x375
Enc-4	Conv1D	k=51, s=2, 8x4	ReLU	4x188
Enc-5	Linear	752x10	-	**1x10**
Dec-1	Linear	10x752	-	1x752
Dec-1b	Reshape	-	-	4x188
Dec-2	ConvTranspose1D	k=51, s=2, 4x8	ReLU	8x375
Dec-3	ConvTranspose1D	k=101, s=2, 8x16	ReLU	16x749
Dec-4	ConvTranspose1D	k=201, s=2, 16x1	-	1x1497
Dec-5	ReplicationPad	-	-	1x1500

extracted 10 features from the HR data, and 10 from the EDA signal. Some examples of the reconstruction performance of the autoencoder is shown in Figures 3 and 4. Based on these results, we found that this particular feature extraction architecture was adequate for our purposes and hence no other changes were considered (ex. adding LSTM layers).

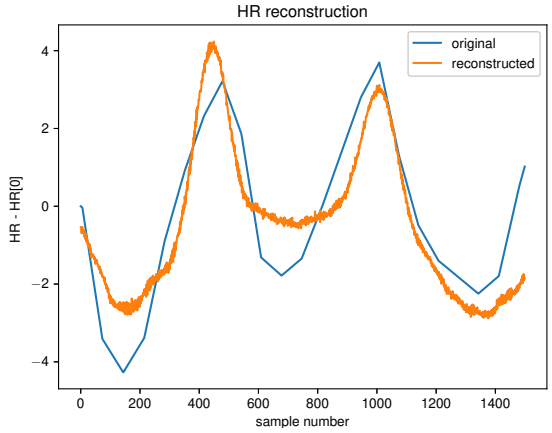

Fig. 3: CNN autoencoder reconstruction of heart rate data.

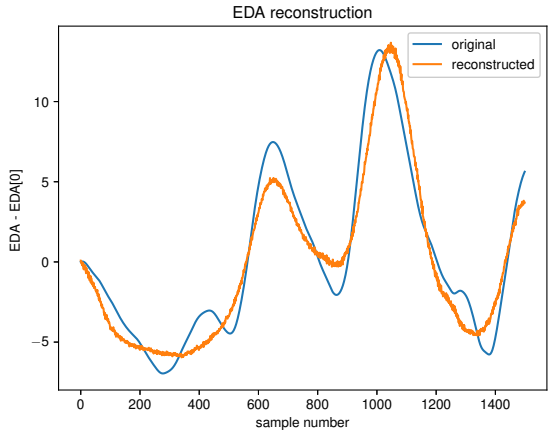

Fig. 4: CNN autoencoder reconstruction of the EDA data.

VI. Models

To show the validity of our hypothesis we explored two approaches. First, we performed Canonical Correlation Analysis (CCA) between the physiological features and events and looked at the extracted canonical loading vectors in relation to the ranking tasks. Then we implemented a simple neural network model and looked at the latent space of the model for correlations with the ranking tasks. Both CCA and the simple neural network were trained using the engineered features and the learned features for comparison.

The targets for CCA and the simple neural network consisted of a multi-hot encoded vector of the events for each 15 second window, where an event had a value of 1 if there was at least one occurrence of that event in the window and 0 otherwise.

A. Canonical Correlation Analysis (CCA)

CCA is a standard statistical technique for finding linear projections of two random vectors that are maximally correlated. Given two column vectors $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_m)$ of random variables, canonical correlation analysis seeks vectors $a \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$ such that the random variables $a^T X$ and $b^T Y$ maximize the correlation $\rho = \text{corr}(a^T X, b^T Y)$. The random variables $U = a^T X$ and $V = b^T Y$ are the first pair of canonical variables. For the second pair of canonical variables we do the same procedure subject to the constraint that they are uncorrelated with the first pair of canonical variables. This procedure may be continued up to $\min\{m, n\}$ times

We used CCA to get an initial sense of the relationship between the features and the targets. The canonical vectors derived from CCA may also be related to the internal emotional state of subjects.

B. Simple neural network model

In our first neural network implementation we chose to use the most simple and basic components of a neural network as a starting point. To achieve this, we used the engineered features and targets described in Section V. The model architecture is an encoder-decoder with a bottleneck in the latent space of 1 neuron. Both our encoder and decoder layers consisted of a fully connected layer with no bias. The lack of a bias facilitates the extraction of rankings from the linear weights. This model is illustrated in Figure 5.

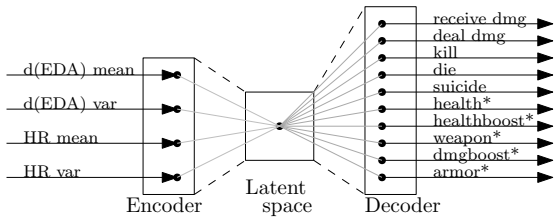

Fig. 5: Implemented model. Events with an asterisk are item pickups.

The data was randomly split into training set and validation set with a $70/30$ ratio.

We used *pytorch* to implement the network in Figure 5 and used the Adam optimizer with a learning rate of 0.001. The loss function was the weighted binary cross entropy loss. The positive weights for the loss function were calculated from the training sets. These weights were necessary to account for class imbalance. 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. This allowed us to maximize the data that would go into the training and validation. Despite the lack of a test set, we paid attention that the loss of the model decreases with each epoch indicating that the model is learning. To evaluate the model's latent space, we extracted rankings based on each target's sensitivity to the latent variable and compared it with the collected subjective rankings.

After some initial results it was apparent that at each training, each with randomly selected training and validation sets, the model tended to converge at slightly different local optima. This is likely due to the significant constraint of a single dimensional latent space. For this reason, we repeated the training 100 times to get a better understanding of were these local optima were. We did not increase the dimensionality of the latent space in order to keep the analysis tenable in this work.

VII. Results and Discussion

A. Methods

For each trained model we can rank the game events according to how sensitive their corresponding neurons are to the latent variables. This can be achieved by either calculating the inverse of the decoder and activating a target neuron or entering a range of values to the input of the decoder analyzing the output. In our case, since our decoder layer consists of a single linear layer with no bias and one input variable, we can easily determine this sensitivity by simply inspecting the linear layer weights. To derive rankings from the models we applied a simple method of sorting the weights of the single linear output layer which transformed the latent space to the target (event) space. Then, for a list of events corresponding to each of the output dimensions we have that:

$$events\_rank = \text{argsort}(W) \tag{1}$$

Where W can be the weights of the linear output layer (for the neural network model) or the canonical loadings of the targets for each CCA component (for CCA).

To measure the significance of the results with each approach in this section, we compared a set of baseline correlation values against the values measured using each model. The baseline values for each ranking task were the correlation between each of 500 random rankings of the events against each of the rankings in that ranking task. So if we have 40 rankings in a ranking task then the total number of baseline values would be $500 \times 40 = 20000$. The null hypothesis is that for each ranking task, the baseline correlation values and the values from the model come from the same distribution. We performed Monte Carlo permutation tests to reject the null hypothesis with 99.99% confidence and two sided significance levels of less than 0.05, 0.01, and 0.001. The test statistic that we used was the difference of medians between the baseline correlations and the model correlations.

B. CCA

1) With the engineered features: A summary of the CCA correlations using the engineered features is available in Table III. The first component seems to correlate mostly with the Arousal ranking task. The second component correlates mostly with the Control and Predictability ranking tasks although it is also somewhat correlated with the other ranking tasks. The third component correlates mostly with the Valence ranking task although it also somewhat correlates with Control (negatively). Finally, the fourth component somewhat correlates with the Arousal (negatively), Valence, and Predictability (negatively) ranking tasks although not very much.

TABLE III: Kendall τ correlation median (M), mean (μ), variance (σ^2) between CCA (engineered features) rankings and ranking tasks for each component.

	Arousal	**Valence**	**Control**	**Predict.**
Comp1	$\mathbf{M = 0.489}^{***}$ $\mu = 0.464$ $\sigma^2 = 0.033$	$M = 0.067$ $\mu = 0.113$ $\sigma^2 = 0.030$	$\mathbf{M = 0.33}^{***}$ $\mu = 0.297$ $\sigma^2 = 0.066$	$\mathbf{M = 0.33}^{***}$ $\mu = 0.253$ $\sigma^2 = 0.097$
Comp2	$\mathbf{M = 0.133}^{*}$ $\mu = 0.139$ $\sigma^2 = 0.026$	$\mathbf{M = 0.2}^{**}$ $\mu = 0.2$ $\sigma^2 = 0.027$	$\mathbf{M = 0.267}^{***}$ $\mu = 0.263$ $\sigma^2 = 0.024$	$\mathbf{M = 0.267}^{***}$ $\mu = 0.247$ $\sigma^2 = 0.036$
Comp3	$M = -0.022$ $\mu = -0.022$ $\sigma^2 = 0.02$	$\mathbf{M = 0.29}^{***}$ $\mu = -0.26$ $\sigma^2 = 0.018$	$\mathbf{M = -0.16}^{**}$ $\mu = -0.15$ $\sigma^2 = 0.028$	$M = -0.04$ $\mu = -0.03$ $\sigma^2 = 0.03$
Comp4	$\mathbf{M = -0.16}^{**}$ $\mu = -0.09$ $\sigma^2 = 0.039$	$\mathbf{M = 0.156}^{*}$ $\mu = 0.14$ $\sigma^2 = 0.023$	$M = -0.02$ $\mu = -0.06$ $\sigma^2 = 0.039$	$\mathbf{M = -0.2}^{***}$ $\mu = -0.16$ $\sigma^2 = 0.027$

$*$: p-value < 0.05, $**$: p-value < 0.01, $***$: p-value < 0.001

2) With the learned features: A summary of the CCA correlations using the learned features is available in Table IV. The first component correlates negatively with the Arousal, Control, and Predictability ranking tasks, and less so with Valence. The second component correlates somewhat negatively with the Arousal, Control, and Predictability ranking tasks. The third component highly correlates mostly with the Valence ranking task but to a lesser degree also with the Control, Arousal, and Predictability. The fourth component correlates with the Arousal ranking task and somewhat negatively with the Valence ranking task.

Comparing the CCA with engineered features against the one with learned features, we observe that the components of the CCA using the learned features have a similar overall pattern but with much more significant correlations indicating that the learned features tend to be better. Overall, the results of CCA point towards the efficacy of learning emotions without using an emotional ground truth. There are clear linear relationships between the CCA components and emotions.

TABLE IV: Kendall τ correlation median (M), mean (μ), variance (σ^2) between CCA (learned features) rankings and ranking tasks for each component.

	Arousal	**Valence**	**Control**	**Predict.**
Comp1	$\mathbf{M = -0.49}^{***}$ $\mu = -0.45$ $\sigma^2 = 0.029$	$\mathbf{M = -0.16}^{**}$ $\mu = -0.2$ $\sigma^2 = 0.034$	$\mathbf{M = -0.42}^{***}$ $\mu = -0.35$ $\sigma^2 = 0.069$	$\mathbf{M = -0.4}^{***}$ $\mu = -0.28$ $\sigma^2 = 0.096$
Comp2	$\mathbf{M = -0.2}^{***}$ $\mu = -0.2$ $\sigma^2 = 0.03$	$M = -0.07$ $\mu = -0.1$ $\sigma^2 = 0.017$	$\mathbf{M = -0.2}^{***}$ $\mu = -0.22$ $\sigma^2 = 0.025$	$\mathbf{M = -0.22}^{***}$ $\mu = -0.22$ $\sigma^2 = 0.039$
Comp3	$\mathbf{M = 0.156}^{*}$ $\mu = 0.142$ $\sigma^2 = 0.022$	$\mathbf{M = 0.378}^{***}$ $\mu = 0.367$ $\sigma^2 = 0.023$	$\mathbf{M = 0.289}^{***}$ $\mu = 0.267$ $\sigma^2 = 0.028$	$\mathbf{M = 0.156}^{*}$ $\mu = 0.107$ $\sigma^2 = 0.047$
Comp4	$\mathbf{M = 0.289}^{***}$ $\mu = 0.276$ $\sigma^2 = 0.033$	$\mathbf{M = -0.24}^{***}$ $\mu = -0.18$ $\sigma^2 = 0.032$	$M = 0.067$ $\mu = 0.02$ $\sigma^2 = 0.041$	$M = 0.111$ $\mu = 0.867$ $\sigma^2 = 0.052$

$*$: p-value < 0.05, $**$: p-value < 0.01, $***$: p-value < 0.001

C. Simple neural network model

1) With the engineered features: Within the 100 trained models using engineered features, we observe two distinct local optima in the results. In Table V we summarize the correlations of each group using a typical result taken from a single model belonging in that group. All results for models within the same group are close to identical with only small variations thus we do not present all 100 for the sake of brevity.

TABLE V: Kendall τ correlation median (M), mean (μ), variance (σ^2) between the simple model (engineered features) rankings and ranking tasks

	Arousal	**Valence**	**Control**	**Predict.**
Grp1	$\mathbf{M = 0.47}^{***}$ $\mu = 0.46$ $\sigma^2 = 0.026$	$M = 0.067$ $\mu = 0.062$ $\sigma^2 = 0.03$	$\mathbf{M = 0.289}^{***}$ $\mu = 0.252$ $\sigma^2 = 0.061$	$\mathbf{M = 0.289}^{***}$ $\mu = 0.2$ $\sigma^2 = 0.1$
Grp2	$\mathbf{M = -0.47}^{***}$ $\mu = -0.46$ $\sigma^2 = 0.025$	$M = -0.02$ $\mu = -0.05$ $\sigma^2 = 0.032$	$\mathbf{M = -0.29}^{***}$ $\mu = -0.25$ $\sigma^2 = 0.059$	$\mathbf{M = -0.29}^{***}$ $\mu = -0.2$ $\sigma^2 = 0.011$

$*$: p-value < 0.05, $**$: p-value < 0.01, $***$: p-value < 0.001

The two result groups appear to be the negative of each other, this is possibly due to the architecture of the neural network and it's symmetric nature which does not restrict the sign of the latent space. Counting the number of models in each group we get that there are 45 models in group1 and 55 models in group2, roughly a $50/50$ split. The results for this simple model, show that the model has captured an emotion related to Arousal in its latent space despite that no emotion ground truth was used in the training. When comparing to the CCA with engineered features, we see a similarity with the first CCA component.

2) With the learned features: Again, we trained 100 models using the learned features this time and observe a variety of results which indicates the existence of multiple local optima. The selection of the training and validation sets made a significant difference in the model's results. Even within these 100 models there were multiple potential patterns starting to emerge and reveal the local optima but the relatively small number of trained models was not sufficient for conclusive remarks and hence we do not present them. We hypothesize that the learned features are richer and more complex, something that CCA is able to overcome by virtue of using the well defined Singular Value Decomposition (SVD) on the cross-covariance matrix between input and output whereas the latent emotion neural network encoder and decoder are not able to cope with only their single layer.

D. Discussion

When using the engineered features, both the simple latent emotion model and CCA had latent components which correlated mostly with Arousal. In the case of CCA, there were components that correlated with Arousal, and Control or with Valence (refer to Section IV-A for possible cause of these groupings). Since the simple latent emotion model only had one variable it was mostly correlated with Arousal but inevitably (c.f. Section IV-A) it was also somewhat correlated with the other emotional dimensions. Therefore, it appears that the engineered features are more suited for the detection of Arousal. The learned features faired similarly with CCA but the increased complexity rendered the simple latent model too sensitive to the training/validation sets and therefore unreliable. It is important to recall that the subjective rankings that we used for this validation do not correspond precisely to the internal emotional state of the subjects, but we do expect them to be correlated. The models may be capturing emotional dimensions that are different from those of the ranking tasks but still correlated with them. More experiments may be necessary to make definite conclusions.

VIII. Conclusion and Future work

In this paper, we wanted to show that automatically inferring the internal emotional state of a person is possible without the use of a subjective ground truth. Our approach is to design the learning process in such a way that it closely mirrors appraisal theory.

To test our hypothesis, we used data from a video game experiment where physiological signals and game events were recorded. We used this data to perform CCA and train machine learning models and derive game event rankings which we compared to subjective rankings of those game events. Our results show that the CCA and model rankings correlate with the subjective rankings for various emotional dimensions confirming our hypothesis. Specifically, CCA using learned features had components which were strongly correlated with the emotional dimensions of Arousal, Valence, and Control. The single-dimensional latent space of a simple neural network using the engineered features showed correlation with Arousal while when using the learned features the simple neural network was not reliable.

These results indicate that our approach has potential but there are several limitations in the present work which we must address. Our dataset does not include enough data per participant to reliably train individual models. As a result, the inter-participant variability within our dataset makes it more

difficult to learn an emotional latent space. Furthermore, we did not log every possible game event in our dataset which makes it difficult to properly qualify the data in terms of emotion. Further experiments must be designed such that these issues are addressed.

The decoder of the simple neural network (in Figure 5), which consists of a single dense layer, cannot span the target space, especially if the latent space is small (1-2 variables). We must include more hidden layers in the decoder to overcome this limitation which will then require a different strategy to derive game event rankings. Calculating the inverse of the decoder is not always possible in this case, so a better option might be to manually activate variables in the latent space with a range of values and observing the decoder output. A larger latent space with more hidden neurons that has independent variables is also desirable for the neural network. With CCA we were able to extract several independent components with different properties and it would be desirable to have something similar for a neural network by increasing the number of hidden neurons at the latent space. Our current implementation does not guarantee that the latent space variables are independent from each other resulting in a less interpretable latent space, although we have not tested this. A possible solution for this is the use of Variational Auto-Encoder (VAE) architectures. Finally, it was apparent from the results (in Table V) that the current simple architecture has two local optima which are the negative of each other. This needs to be addressed in future architectures by including assymetrical components in the encoder and decoder layers.

Our main contribution in this work is that we have shown that learning latent representations or using dimensionality reduction techniques on objective data such as physiological signals can result in a latent space with emotionally correlated structures even though emotion labels were not part of the training data. It is yet to be seen if these structures are similar across datasets and across domains outside of video-games, a subject of further research. If these structures are significantly different across datasets/domains then the applicability of such unsupervised methods is reduced since at least some ground truth data will be needed to interpret these different structures.

REFERENCES

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[2] N. Fragopanagos and J. G. Taylor, "Emotion recognition in human-computer interaction," *Neural Networks*, vol. 18, no. 4, pp. 389–405, 2005.

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

[4] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang, "A review of emotion recognition using physiological signals," *Sensors (Switzerland)*, vol. 18, no. 7, 2018.

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

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

[7] J. M. Kivikangas, G. Chanel, B. Cowley, I. Ekman, M. Salminen, S. Järvelä, and N. Ravaja, "A review of the use of psychophysiological methods in game research," *Journal of Gaming & Virtual Worlds*, vol. 3, no. 3, pp. 181–199, sep 2011.

[8] H. P. Martinez, Y. Bengio, and G. N. Yannakakis, "Learning deep physiological models of affect," *IEEE Computational Intelligence Magazine*, vol. 8, no. 2, pp. 20–33, 2013.

[9] P. Barros, E. Barakova, and S. Wermter, "A deep neural model of emotion appraisal," *arXiv preprint arXiv:1808.00252*, 2018.

[10] M. Maier, D. Elsner, C. Marouane, M. Zehnle, and C. Fuchs, "Deepflow: Detecting optimal user experience from physiological data using deep neural networks." in *AAMAS*, 2019, pp. 2108–2110.

[11] G. Chanel and P. Lopes, "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.

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

[13] D. Sander, D. Grandjean, and K. R. Scherer, "A systems approach to appraisal mechanisms in emotion," *Neural Networks*, vol. 18, no. 4, pp. 317–352, 2005.

[14] J. D. Parker, D. H. Saklofske, P. A. Shaughnessy, S. H. Huang, L. M. Wood, and J. M. Eastabrook, "Generalizability of the emotional intelligence construct: A cross-cultural study of North American aboriginal youth," *Personality and Individual Differences*, vol. 39, no. 1, pp. 215–227, 2005.

[15] C. MacCann and R. D. Roberts, "New paradigms for assessing emotional intelligence: Theory and data." *Emotion*, vol. 8, no. 4, pp. 540–551, 2008.

[16] A. Moors, "Theories of emotion causation: A review," *Cognition and Emotion*, vol. 23, no. 4, pp. 625–662, 2009.

[17] K. R. Scherer, "Appraisal Theory," in *Handbook of Cognition and Emotion*, 2005.

[18] K. R. Scherer and A. Moors, "The Emotion Process: Event Appraisal and Component Differentiation," *Annual Review of Psychology*, vol. 70, no. 1, pp. 719–745, jan 2019.

[19] K. R. Scherer, "The dynamic architecture of emotion: Evidence for the component process model," *Cognition & Emotion*, vol. 23, no. 7, pp. 1307–1351, nov 2009.

[20] J. W. Fernando, Y. Kashima, and S. M. Laham, "Alternatives to the fixed-set model: A review of appraisal models of emotion," *Cognition and Emotion*, vol. 31, no. 1, pp. 19–32, 2017.

[21] S. Latif, R. Rana, J. Qadir, and J. Epps, "Variational autoencoders for learning latent representations of speech emotion: a preliminary study," *Interspeech 2018: Proceedings*, pp. 3107–3111, 2018.

[22] S. Parthasarathy, V. Rozgic, M. Sun, and C. Wang, "Improving emotion classification through variational inference of latent variables," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 7410–7414.

[23] J. Li, S. Qiu, C. Du, Y. Wang, and H. He, "Domain adaptation for eeg emotion recognition based on latent representation similarity," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 12, no. 2, pp. 344–353, 2020.

[24] H. Fei, Y. Zhang, Y. Ren, and D. Ji, "Latent emotion memory for multi-label emotion classification," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, pp. 7692–7699, Apr. 2020.

[25] J. R. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, "The World of Emotions is not Two-Dimensional," *Psychological Science*, vol. 18, no. 12, pp. 1050–1057, dec 2007.

[26] G. N. Yannakakis, R. Cowie, and C. Busso, "The Ordinal Nature of Emotions: An Emerging Approach," *IEEE Transactions on Affective Computing*, vol. 3045, no. c, pp. 1–1, 2018.