Classification of Emotional State using Multi-Modal Physiological Signals

ALKINOOS SARIOGLOU, ANTONIO ARBUES

Compiled May 29, 2021

In this report, the authors demonstrate the methods used to extract useful features from raw Electrocardiogram (ECG), Electroencephalogram (EEG), Electrodermal Activities (EDA) and facial landmark trajectories (EMO) data in order to train a Random Forest Classifier and predict the emotional state of a subject with higher accuracy than the Zero Rule classifier.

1. FEATURE EXTRACTION AND DATA ANALYSIS

A. ECG

First of all, the Power Spectral Density of the Lead I ECG signal for the first participant watching Clip 1 from 0-40 Hz is plotted in *Figure* 1.

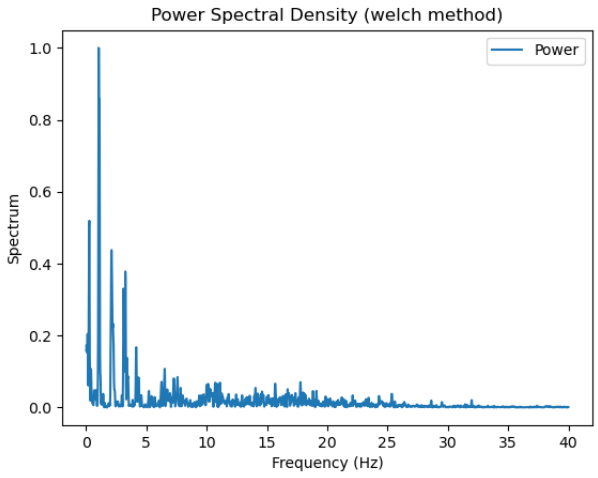

Fig. 1. Power spectral density of the ECG computed with the Welch method for Participant 1 watching Clip 1.

The most significant part of the ECG signal is included in the frequency range between 0-20 Hz as it is visible from *Figure* 1. Therefore, a Butterworth low-pass filter with cut-off frequency 18 Hz is selected. The filter's frequency response is shown in *Figure* 2. After filtering, the filtered signal is plotted in time-domain in *Figure* 3.

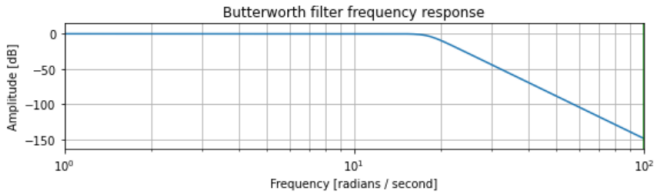

Fig. 2. Low-pass filter for ECG noise removal.

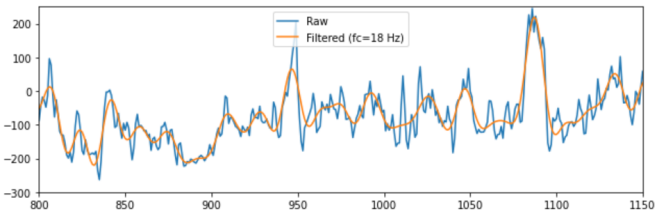

Fig. 3. ECG signal after applying the low-pass filter.

Following this, an algorithm for artifact detection is implemented and it is found that the percentage of ECG data flagged as having artifacts is 7.23%. A signal found to contain artifacts is the ECG trace of participant 41 watching clip 3. This is shown in *Figure* 4.

The additional features selected to improve the classification

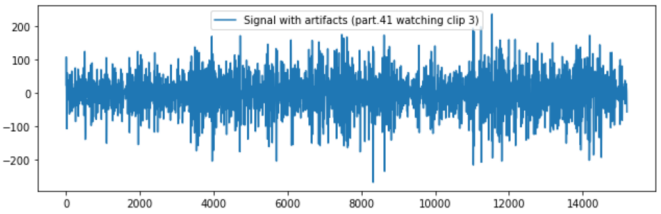

Fig. 4. An ECG signal presenting artifacts.

for emotion recognition are the following:

Ten Mid Frequency [5-10 Hz] PSDs, because there is a lot of information about the QRS complex in this range, which could demonstrate how fast the ventricles depolarize as a result of a change in emotion (faster depolarization would mean higher arousal),

RMSSD (Square Root of Mean of Sum of Successive Differences between adjacent RR intervals), demonstrates differences

in RR intervals over time and can show changes in trend of heart rate and hence emotion,

Ten Mid Frequency [10-15 Hz] PSDs, for the same reasons as above,

SD1SD2 (ratio between short- and long-term fluctuations of the RR intervals), shows how the RR intervals change in long-term when compared to short-term fluctuations,

C1d/C1a (the contributions of heart rate decelerations and accelerations to short-term HRV), can illustrate instantaneous changes in HRV due to heart rate accelerations or decelerations which could be caused by changes in emotions.

B. EMO

In order to extract the statistical measurements from the EMO signal, the data has been extracted clip-wise from the container. Then, for every clip and every different feature, the mean, standard deviation, skewness, kurtosis, and the percentage of times the values deviate more than one standard deviation from the mean are calculated with the help of the `statistics` and `SciPi` libraries.
The data is in this way appended to an array and is ready to be fed to the classification method.

C. EDA

The first task is to extract the Power Spectral Density (PSD) of the second participant watching clip 1.
After a visual inspection over the data, it is evident that artifacts affect the data heavily.
Hence, an algorithm that filters out consecutive data points with deviation greater than a certain threshold has been crafted and applied.
Now, it is possible to plot a clean PSD, unaffected by evident artifacts, as shown in *Figure* 5.

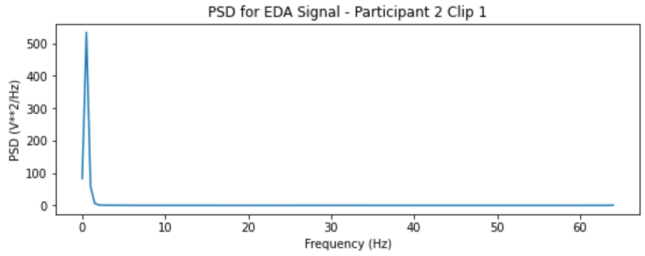

Fig. 5. The PSD plot of the EDA signal of the second participant watching clip 1.

From the plot in *Figure* 5 it is evident that the interesting part of the signal is contained at very low frequencies.
Thus, a low-pass filter with cut-off frequency of 0.8 Hz is crafted, as shown in *Figure* 6.

The application of the low-pass filter results in the clean signal shown in *Figure* 7 in its time domain.

A successful filtering is a key element for the extraction of great features. For this reason, it has been chosen to use the `Neurokit2` library to filter all the EDA signals available in the data. Then, for the last 50 seconds of every recording, the 15 relevant features are computed using self-crafted algorithms and other libraries like `NumPy`, `SciPy`, and `statistics`.
In addition to these standard features, five additional features are proposed.
These are:

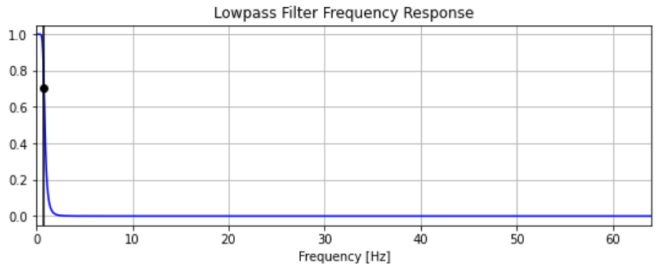

Fig. 6. Frequency response of the low-pass filter used to remove high-frequency noise in the EDA signal.

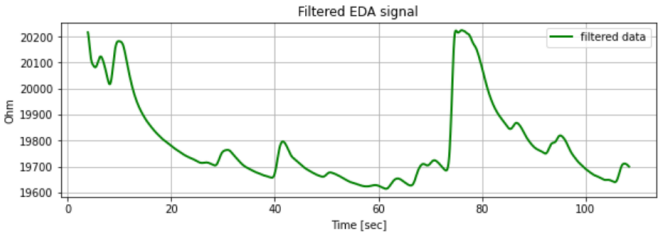

Fig. 7. The time-domain plot of the filtered EDA signal of the Participant 2 watching Clip 1.

The **mean of tonic component of the signal**, since the tonic level is generally considered the background level of activity on top of which rapid GSR responses appear,

The **mean of phasic component of the signal**, since these are generated as a response to a specific event (e.g., visual stimulus or unexpected question) known as event-related SCR (ER-SCR). ER-SCRs are the most common measure used in research to relate changes in emotional arousal to specific stimuli,

The **mean of SCR amplitude of the signal including the Tonic component**, since it might be a good indicator for the average of the arousal,

The **maximum of SCR amplitude of the signal including the Tonic component**, since it might be a good indicator for the maximum of the arousal and

The **minimum of SCR amplitude of the signal including the Tonic component**, since it might be a good indicator for the minimum of the arousal.

D. Valence and Arousal

The Pearson Correlation Coefficient between valence and arousal is calculated to be -0.0177.
The value of the Pearson Correlation Coefficient is very close to 0 and very slightly negative, therefore it means that there is no significant correlation between arousal and valence. As a result, the intensity of an emotion (arousal) is not dependent on whether the emotion is positive or negative (valence), because both negative and positive emotions can be intense or less intense.
When considering the distributions of the classes, then the distribution of low arousal and low valence is found to be 311, the distribution of low arousal and high valence is 188, the distribution of high arousal and low valence is 473 and the distribution of high arousal and high valence is 483. That means that most clips created high intensity of positive emotions in the participants and only few participants had low-intensity positive emotions while watching some of the clips. Regarding negative emotions,

most participants had high intensity of negative emotions rather than low-intensity ones while watching the clips.

2. CLASSIFICATION

In order to get started with the classification, it is necessary to organise the features into different containers that are going to be needed for the analysis. One is a clip-wise feature organisation, and the other one is a participant-wise organisation.
At this point the data is inspected and some non-valid - `NaN` - values are discovered. The classification algorithm needs all the data to be valid, hence an imputation algorithm from the `sklearn` library is going to be deployed in the loop to infer the missing data based on the known part of the data. The missing values are imputed using the mean of the values in the columns where the missing values are located.
Next, the `LeaveOneOut` method of the `sklearn` library is used to loop over the data organised clip-wise and participant-wise. In order to classify the input data into the four categories formed by high-low valence and high-low arousal, a Random Forest Classifier is used. Furthermore, to increase the performance of the classifier over the data, a grid-search hyper-parameter tuning method has been deployed to discover the most tailored parameters for this kind of task.

A. Performance Measures on Two Validation Schemes

The classification is performed in two different ways (participant-wise and clip-wise) and the obtained accuracy, as well as precision, recall, and F1-score for each class are reported in *Table* 1 and *Table* 2.

B. Confusion Matrices

The two validation schemes produce slightly different results in the classification. These differences are mirrored in the two confusion matrices that result as the sum of confusion matrices of each cross-validation split. The two confusion matrices are shown in *Figure* 8 and *Figure* 9.

C. Ten Selected Features

An algorithm performing randomized feature selection, model fit, and performance evaluation has been crafted and used in order to spot the most relevant features.
After several iterations, the ten best features were found to be:

EDA: the PSD at 0.1Hz band
ECG: standard deviation of the hearth rate
EEG: kurtosis of the second channel
EEG: skewness of the third channel
EEG: percentage of values below the mean in the second channel
EMO: percentage of values above the mean of the deformation of the right cheek
ECG: percentage of values below the mean of the heart rate
EMO: standard deviation of the vertical deformation of the lower lip
ECG: the PSD at 0.72Hz band
EMO: kurtosis of the vertical deformation of the right lip corner

The results extracted by using these features are shown in *Table* 3.

D. Zero Rule Algorithm

The implementation of the Zero-R algorithm yields an accuracy of 6.6% for the clip-wise validation scheme and 30.7% for the participant-wise validation scheme.
As evident from the results, the Random Forest Classifier outperforms by 3% the Zero-R classifier for the participant-wise validation scheme, and by 30% in the clip-wise validation scheme.
In the participant-wise classification it is evident that the class with the most occurrences in the training dataset is predicted the most, which is why its accuracy is very high. However, this does not mean that the Zero-R classifier achieves a good performance since all the other classes are not predicted at all.
Furthermore, in the clip-wise classification it is observed that the accuracy is really low because the clips generate inherently different emotions.

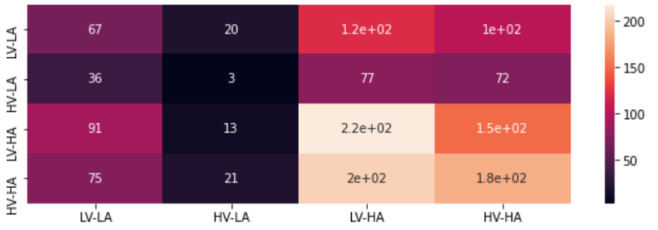

Fig. 8. The combined confusion matrix for the leave-one-participant-out validation scheme.

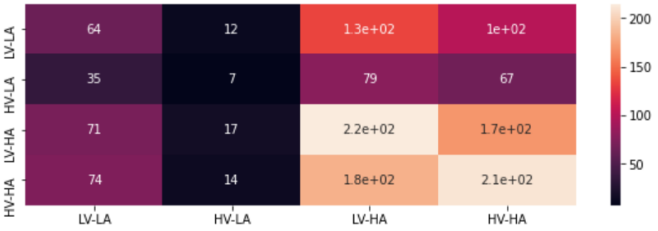

Fig. 9. The combined confusion matrix for the leave-one-clip-out validation scheme.

3. DISCUSSION

A. Most Relevant Sensing Modalities

As shown from the results *Table* 1 and *Table* 2, the modalities that achieve highest accuracy are the ECG features (32.9%) and the EMO (32.6%).
However, only ECG would be unobtrusive and would allow continuous tracking in real life if the electrode patches were attached on the chest and the leg and they were also capable of transmitting the signals wirelessly. All the other modalities require hardware mounted on body parts such as hands or head that would make everyday activities difficult for the subjects.

B. Generalization Across Participants and Clips

The proposed classifier is a Random Forest Classifier whose hyper-parameters have been tuned through cross-validation.
In particular the Classifier settings are: Maximum depth of each

Feature Class	Labels	Accuracy	Class Accuracy	Precision	Recall	F1-Score
All the features	Low-Arousal Low-Valence	0.336	0.215	0.291	0.215	0.247
	Low-Arousal High-Valence		0.010	0.044	0.010	0.017
	High-Arousal Low-Valence		0.460	0.346	0.460	0.395
	High-Arousal High-Valence		0.418	0.367	0.418	0.391
ECG features	Low-Arousal Low-Valence	0.329	0.180	0.234	0.180	0.203
	Low-Arousal High-Valence		0.015	0.062	0.015	0.025
	High-Arousal Low-Valence		0.437	0.338	0.437	0.381
	High-Arousal High-Valence		0.440	0.382	0.440	0.409

Table 1. The table shows the performance of the Random Forest Classifier in the leave-one-participant-out validation scheme.

Feature Class	Labels	Accuracy	Class Accuracy	Precision	Recall	F1-Score
All the features	Low-Arousal Low-Valence	0.342	0.186	0.250	0.186	0.213
	Low-Arousal High-Valence		0.026	0.102	0.026	0.042
	High-Arousal Low-Valence		0.473	0.348	0.473	0.401
	High-Arousal High-Valence		0.436	0.436	0.418	0.415
EMO features	Low-Arousal Low-Valence	0.326	0.209	0.288	0.209	0.242
	Low-Arousal High-Valence		0.053	0.175	0.053	0.081
	High-Arousal Low-Valence		0.391	0.313	0.391	0.347
	High-Arousal High-Valence		0.445	0.369	0.445	0.403

Table 2. The table shows the performance of the Random Forest Classifier in the leave-one-clip-out validation scheme.

Feature Class	Labels	Accuracy	Class Accuracy	Precision	Recall	F1-Score
Selected features, clip-wise	Low-Arousal Low-Valence	0.359	0.238	0.292	0.238	0.262
	Low-Arousal High-Valence		0.069	0.188	0.069	0.101
	High-Arousal Low-Valence		0.461	0.375	0.461	0.414
	High-Arousal High-Valence		0.449	0.393	0.449	0.419

Table 3. The table shows the performance of the Random Forest Classifier in the leave-one-clip-out validation schemes for the ten most performing features.

tree = 100, Minimum number of samples for splitting = 2, Number of trees = 16, Balanced sub-sample weights.
The classifier generalizes well across participants, especially in the classes of high arousal, achieving an accuracy of 40%, but worse for the other two classes where the accuracy drops to 10%. On the other hand, the classifier generalizes better across different clips as it achieves a general accuracy of 36%, where the high-arousal accuracy is of 45% and around 15% for the low-arousal classes.

C. Classifier Performance Across the Four Classes

The classifier performs visibly better for high-arousal conditions, as it is evident from the confusion matrices in *Figure* 8 and *Figure* 9.
The reason is two-fold. First of all, due to the distribution of the classes in the training dataset, the classifier is more biased towards the two most occurring labels. Additionally, it is easier for the used sensors to register an emotional reaction with high arousal rather than one with milder intensity.

D. Additional Sensing Modalities

Two additional sensing modalities that would aid in improving the classification accuracy of the emotion recognition classifier would be speech recognition and the measurement of blood pressure.
Regarding speech recognition, there are multiple characteristics on the way of speaking such as speed, number of pauses or tone that give information on the current emotional state of a person. Therefore, it would be a useful modality to record an audio clip of the participant describing what they see on the video while they are watching it.
Furthermore, it is an unobtrusive way of recording signals from the subject and it could be used for continuous tracking in everyday life.
In addition to that, another modality that would help is measuring the blood pressure through a wearable sensor. This modality would certainly give additional information on the emotional state of the subject as blood pressure can demonstrate very effectively when the feelings of a person increase in intensity with the increase of their blood pressure. Blood pressure tracking would also be relatively unobtrusive and could be used for continuous tracking in daily life as it can be measured by wearable watches nowadays.