Group Report - Affective Computing and Human Robot Interaction (COMP0053)

Team Eleos

Sadir Abdul Hadi, Jacob Turton, Mo Afsharmoqaddam
Sachchit Prasad, Shivam Shah, Pranav Nashikkar

Word Count : 2546 (excluding references, contents page, captions)

Contents

1 Introduction

The relationship between music and sentiments has been widely researched in the last few decades. More recently, with the rise of big data methods and music streaming services, multiple platforms which play music depending on a user's mood have emerged. Examples include allmusic.com, moodfuse.com and SoundR. Most of these platforms directly ask participants about their mood, and invite them to choose one out of a limited sets of moods such as happy, nostalgic or energetic.

This project looks to complement Spotify services with an emotion recognition system to improve users' listening experience.

1.1 The choice of emotional states to measure and suitable sensors

As most smartphones and laptops currently have cameras installed, and because GSR (Galvanic Skin Response) sensors are making their way to wearable devices, we use these two types of sensors to measure happiness, sadness and arousal respectively.

The choice of sensors surely has its limitations, including the intrusive nature of continuously monitoring the webcam of a laptop for example, or the fact that mobile phone holders don't always face their device while listening to music. Nevertheless, these two sensors have been extensively used in the literature, where GSR has been shown to be a reliable physiological measure for arousal [1], while webcams have been used to generate reliable results regarding human emotions including happiness and sadness [2]. The perfect scenario to use our system is listening to music while working on a laptop: this minimizes the movement, and guarantees a good angle for the webcam.

When it comes to the choice of the emotion space we want to cover, we have noticed that many academic works have used valence and arousal [3]. Nevertheless, these papers also stated the ambiguity that would sometimes come into play when analysing the results. Scores in the same quadrant, for example can illustrate excitement, happiness, or pleasure, which are different nuanced feelings by definition [3]. For this reason, we have decided to replace valence by a happiness\sadness score.

2 Collecting, Labelling and Refining Own Data

In order to better understand the challenges towards our own problem statement, the team had decided to devise experiments to collect our own data. The purpose of the experiments were to illicit some response, whether facial or physiological from the participant when watching videos and listening to music. There were two stages to this, a preliminary experiment and a final experiment, where the final experiment was informed by our findings in the preliminary experiments. By running independent experiments, labeling the data was also conducted systematically. We initially focused on a sample of the data collected from the preliminary experiment and the labelling practise was reviewed and refined for future core experiments. The labeling practises consisted of a mixture of qualitative and quantitative approaches.

2.1 Preliminary Experiments

For the preliminary experiments, a few initial videos were chosen and then split into happy or sad categories. These videos were then presented in their entirety to each participant in separate sittings, where timings ranged between 2-3 minutes and facial and physiological data were captured using the webcam and the GSR sensor respectively. During the process, the team would mark at which points of viewing the videos the participants elicited the greatest responses, if at all, based on their live GSR results. Once the viewings were over, the participant was then interviewed for their thoughts using a semi-structured interview.

2.2 Preliminary Experiment Findings

The main findings were to streamline data-collection process with many participants, choosing 1 minute long segments of the videos which showed the greatest changes in facial and GSR values as well as removing videos which were not very useful and replacing them with new ones. Interviews at the end made it harder for people to remember their prior emotions, hence a new method would be needed to collect data accurately and partially adding in quantitative data to supplement the heavily qualitative nature.

2.3 Collecting Data From Final Experiments

For the final experiments, participants were shown 6 videos (3 happy, 3 sad), where after watching each video they would mark their arousal and emotional state on our data-collection sheet. After 3 videos were completed, the participant would be interviewed in an open format to get their thoughts on the previous 3 videos, this would be repeated again after the final video. In total 3 participants were used for the preliminary stage and 5 different participants for the final stage.

2.4 Quantitative Data Collection

A quantitative approach enabled us to run descriptive statistics on our data set to derive interesting features. Each participant had fill-in sheets which consisted of two simple questions for each experiment as shown in figure 1.
After a participant finished watching a video, they rated their feelings from a scale of 1-10, where 1 indicated very sad and 10 being very happy. Similarly, they would also rate their arousal from a scale of 1-10 where 1 is not at all aroused and 10 being highly aroused.

2.5 Qualitative Data Collection

Qualitative data often provides subjective, specific and rich information normally shown through the participant's perspective where we assume a dynamic and negotiated environment. After a participant had finished 3 experiments, the data were collected through short semi-structured interviews about any specific feelings and reactions they may have had when completing the experiments. This further enforced and guided our understanding about the themes of the experiments and how the participant truly felt in regards to their quantitative scores.

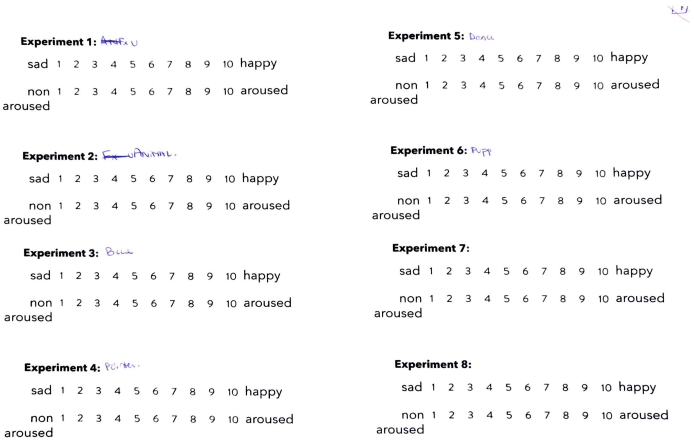

Figure 1: An example of a blank sheet that the participant would fill in

2.6 Labelling and Refining Data

Based off the scores given by the participants, these scores were mapped to create our facial and GSR data-sets. For emotion, a score less than 5 meant the participant was sad/low-aroused and a score higher than 5 was considered happy/aroused. Where a score of 5 was indicated for either metric, based on the interviews we would class this as either happy or sad. Then for each respective participant, the frames of their captured facial data and time-series GSR data would be tagged with 1 for happy and 0 for sad.

3 Other Datasets

In our investigation, we recognised that due to a lack of time and resources, our collected data would likely be of limited size or quality. As a result, we had to look at alternative existing datasets to help train our subsequent Machine Learning models, especially the more data-intensive Deep Learning models.

3.1 DEAP Dataset (Physiological)

This particular dataset contained information about participant ratings and physiological recordings of an experiment where 32 volunteers watched 40 one-minute long excerpts of music videos [4]. The study that accompanied the creation of this dataset was closely related to our context in music, and so encouraged its usage in our investigation [5]. However, the number of physiological data samples ranged around 1200, inhibiting the implementation of Deep Learning models.

3.2 Facial Recognition Dataset

For the facial recognition models, we used a Kaggle competition dataset "Challenges in Representation Learning: Facial Expression Recognition Challenge" [6]. The data consisted of 48x48 pixel

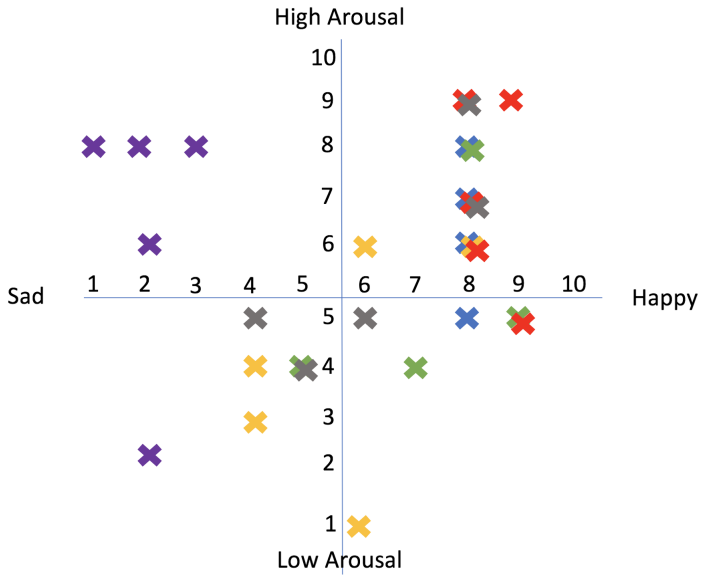

Figure 2: The final graph on which the experimental results were plotted to give a visual representation. Experiment number and colour key : 1 (blue) , 2 (green), 3 (yellow) , 4 (red), 5 (grey), 6 (purple)

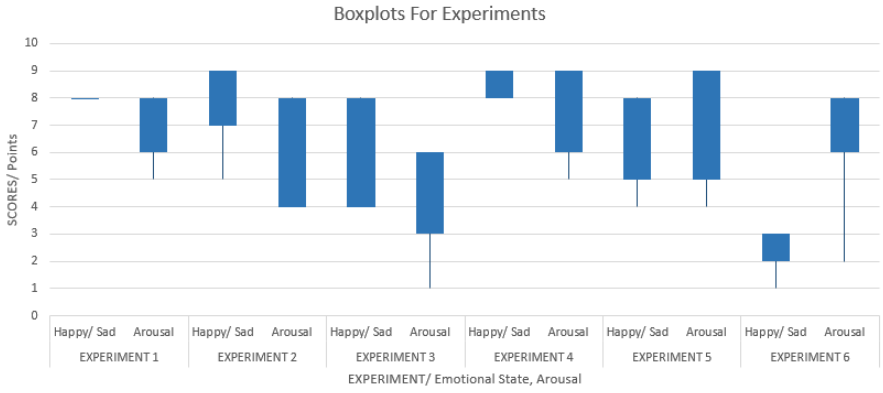

Figure 3: Scores for emotion level and arousal over the 6 experiments, wider bars represent more spread over scores given, where as thinner bars provide us with more certainty

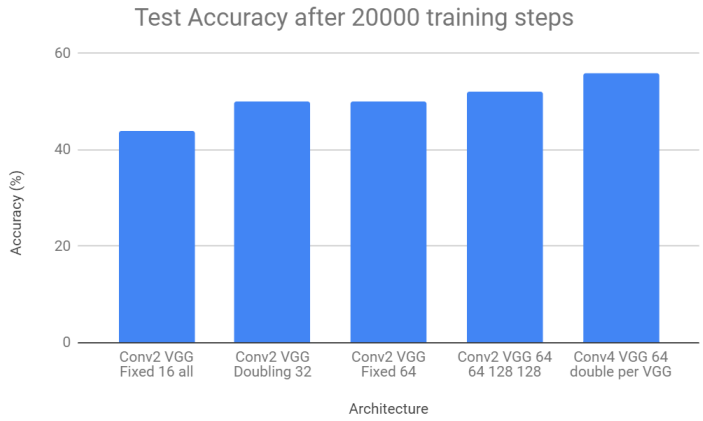

Figure 4: Training Performance of CNN Architectures

grayscale images of faces, with seven categories Anger, Disgust, Fear, Happy, Sad, Surprise, Neutral. More importantly, the training set consisted of 28,709 examples allowing for implementation with Deep Learning techniques.

4 Facial Classification

VGGNets [7] are a particular type of convolutional neural network (CNN) that uses stacked layers of 3x3 filter sizes. This method retains the power of CNNs, but maximises the data-driven feature extraction approach, allowing for minimal human biasing in architectural design. When stacked in layers, these combine into more complex and sophisticated transformations and have proven to be extremely successful in computer vision and specifically facial emotion classification tasks [8] [9] [10].

To match the preprocessing of the dataset, we resized the inputs to 48x48 and grayscaled, preserving most relevant information, and speeding up computation. Furthermore, the reduced dimensionality of the problem makes the task easier and thus improves performance.

4.1 Model Optimisation

Complex architectures enable better representations, however can 'overfit' and learn irrelevant noise, but 'dropout' [11] was applied to address this, boosting performance. Since a VGGNet architecture had already been chosen, the hyperparameters to tune were the number of convolutional layers, and number of filters per layer. Each model was trained for 20,000 training steps on all of the dataset with a 7-class emotion objective (to use all available data).
As seen in 4, a 4 VGG block (8 layer) convolutional neural network doubling numbers of filters at each block performed best. This doubling offsets the downsampling 'pooling' in CNNs to preserve representational power throughout. This model achieved 86% binary accuracy (since the training dataset only has categorical happiness/sadness rather than values to regress on), 0.85 precision,

0.81 recall and 0.87 area under the precision-recall curve (AUC-PR). This is a very strong result, demonstrating that faces can be highly relevant for emotion classification.

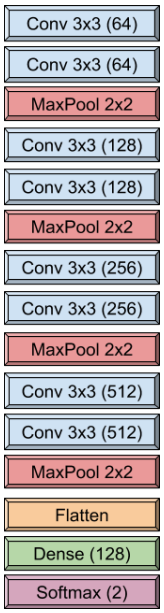

Figure 5: Final CNN Architecture. N.B. ReLU activation was used after each convolution

4.2 Results

Classifying the experimentally gathered data directly had weak results, with only about a 54% accuracy for binary classification; significantly lower than original 86% test. Since our experiment uses the same label for an entire video, this figure could be caused by infrequent facial emotion elicitation. Instead, a classification was made for each input in a video and then the mode class was selected as 'overall' classification, boosting accuracy to 58%. This suggests that the input videos may not be the same emotion throughout or partially unemotional.

An improvement applied averaging over the class probabilities per video to remove noise by averaging uncertainties, resulting in 61% accuracy. This is especially important in reducing the effects of more neutral images, as the classifier is less certain of them then their effects can be reduced. By thresholding predictions to enforce a minimum model certainty, neutrality can be detected or removed. In practice, this did not increase binary classification accuracy versus averaging however if trained on a large dataset, can act as a third emotional class.

Since the original dataset was emphasised emotion rather than our experimentally elicited data, the features that our model has learned may be sub-optimal for recognising subtler emotions. We were unable to gather enough data to train a good CNN model with limited data. Ideally, we would train the model on only relevant data, so that the CNN can truly learn to extract features and specialise to the task at hand for significantly higher accuracies. Furthermore, we could integrate

the happiness scaling of our experiment as a regression target rather than classification to allow for the prediction of how happy or sad a user may be.

5 Galvanic Skin Response (GSR) and Arousal

The GSR data was used to train models which could predict the arousal felt by participants as indicated by previous research [12]. We treated arousal as a continuous variable and therefore this was a regression problem.

5.1 Features

The GSR data was collected as a time series of values. In order to train the models, features had to be extracted from this data. We followed previous research [13] which suggested that the mean value, standard deviation, mean of the first derivative and mean of the second derivative were features of interest. After extracting these features from the data, they could be fed into the regression models for training.

5.2 Models

We chose five different regression models which could be trained and then compared for performance on an unseen test set. The Scikit-learn package for Python [14] was used as it offers a large range of regression models. From this, five regression models were tested: linear regression, LASSO regression (l1 regularisation), Ridge regression (l2 regularisation), support vector regressor and a random forest regressor.

5.2.1 Linear Regression (including LASSO and Ridge)

A linear regression model is perhaps the most straightforward type of regression model [15]. However, they can be prone to overfitting on the training data. To overcome this problem, regularisation can be used to try and avoid learning an overly complex model. Both Lasso and Ridge regression are forms of regularisation that use slightly different methods.

5.2.2 Support Vector Regression (SVR)

SVR uses a similar methodology to Support Vector Machines (SVM) but for regression problems. The advantage of SVMs is that they can capture non-linear relationships [15].

5.2.3 Random Forest Regression

Random forests use large numbers of decision trees. In theory this helps avoid the overfitting. Like SVMs random forest regressors can capture non-linear relationships [15].

5.3 Data

Two thirds of the data was used for optimisation and model training. The remaining one third was kept aside and not seen by the model until testing. We first trained on the DEAP dataset and

Models	R-squared	MSE
Lasso	-0.04	4.1
Ridge	-0.03	4.1
Linear Regression	-0.04	4.1
Support Vector	-0.02	4.0
Random Forest	-0.15	4.5

Models	R-squared	MSE
Lasso	-0.00	4.7
Ridge	-0.00	4.7
Linear Regression	0.07	4.3
Support Vector	-0.20	5.7
Random Forest	-0.33	6.3

Figure 6: Model performances evaluated on R-Squared and MSE Scores : Lasso, Ridge, Linear Regression, Support Vector, Random Forest

then trained on our own collected data separately. This is because they were measured on different scales and therefore it was not possible to transfer the trained models between the two.

5.4 Model Optimisation

All of the models (apart from basic linear regression) had parameters which can be tuned to optimise the performance. We used k-fold cross validation and grid search to find the best combination of parameters for each of the models. For k-fold cross validation we used 3 folds and the error metrics were averaged across all three. This allowed us to identify the parameter combinations that gave the best overall performance for each model. Some models such as Ridge and Lasso only had one parameter to tune whereas others such as the random forest regressor had many.

5.5 Evaluation

To evaluate the models, two metrics were used: mean squared error (MSE) and R-squared. MSE is the average squared error between the ground truth values and the model's predictions. R-squared measures how well the data fits to the model's regression line. MSE is difficult to interpret alone since it is not clear how small of an error is "good" or "bad". R-squared is easier to interpret straight away. A value of 0 suggests the model performs no better than a model which just predicts the mean value. A value closer to 1 is better performance, whereas a negative value suggests poor performance. MSE and R-squared were both used for the optimisation process and the final evaluation.

Table 1 below shows the final test set results for the DEAP dataset. Table 2 shows the final test set results for our collected data.

The results show that for both datasets features extracted from GSR data were not good at predicting arousal level in participants.

6 Conclusion and Future Work

Music has always been a prevalent choice for the depiction and understanding of human emotion, and so we are seeing more research focussing on discovering relationships between music and sentiment. To this end, we looked at complementing existing music streaming applications, such as Spotify, by building an emotion recognition system that could be used to classify a user's mood during their listening experience using physiological and facial recognition data. We found that

for the facial recognition models, classifying the experimentally gathered data directly produced poor results, which was boosted by averaging over the class probabilities per video. With the GSR models we had less success as our results showed the general range of R-squared values being around 0 or less. This indicated that for both the DEAP and our collected datasets, the feature extraction from GSR data was not good at predicting arousal in participants.

We also recognise scope for improvement with our research. To improve accuracies of our facial recognition model, we could use a larger dataset of only relevant data such as our experimentally elicited data. The facial data we used to train our model included over-emphasised emotion that would not be very prevalent in casual listening experiences, and so the features learned were suboptimal for recognising subtler emotions. Additionally, we could experiment further with a different choice of features for the GSR models due to the lack of success we had. Lastly, to build a more robust and useful emotion recognition system especially in the context of industrial application, we could later look at incorporating more classes of emotions such as all seven basic emotions of Anger, Disgust, Fear, Happy, Sad, Surprise, and Neutral.

Affective Computing and Human Robot Interaction - Individual Report

Mo Afsharmoqaddam

COMP0053

Contents

1 Introduction

Music has the power to evoke strong human emotions, where individuals in their everyday lives experience, deal with and enhance their emotions through listening to music. How music stimulates emotions is an active research field and this report will critically discuss the motivations and challenges that emotion recognition specifically through music poses to the Machine Learning community. Moreover, the pros and cons of the methodologies used in the project and how they may be improved will be discussed.

1.1 Aim

The aim of this project evolves around enhancing the experience and recommendation of music by better understanding an individuals emotions through streaming application such as *Spotify*.

2 Importance and Consequences of Emotion Recognition Through Music

2.1 Importance

Music plays an important role in many areas of our lives, through going to a cafe for a coffee to attending weddings, we are constantly involved in a life where music is part of us. However, its use cannot be justified through only such functions, Thompson et. al, (2012) discusses the effect of music is usually personal and shapes different emotional states [16]. Different attributes of music are directly linked to distinct emotional interpretations and they can be changed to convey complex emotional messages. Blood and Zatorre (2001) and Juslin and Sloboda (2001) support the idea that pleasure and emotions are the main motivators when listening to music as well as inducing very strong emotions [17][18].

Hence, recognising emotions indicates an improvement of quality of life for an individual. Our project aims to recognise emotions through listening to music so that better song suggestions can be given to an individual, more appropriate to their mood. This can release stress levels [19], increase the productivity of an individual [20] as well understanding deeper traits of health conditions such as depression and alzheimers [21].

2.2 Consequences

On the contrary, understanding an individual's emotion through music could also have severe consequences. Music combined with emotions could be a very subjective matter, and predicting the wrong emotions and recommending the *wrong* songs could make an individuals feelings worse. At the same time there can be biases when recommending music as well as the recommendation could have influence over the emotions an individual is feeling (e.g you may not naturally want to change your emotional state).

Additionally, ethical issues such as "*knowing*" of an individuals personal feelings and their breach of privacy can also be problematic. Therefore, it is challenging to collectively understand deeply how a user is feeling (as well as protecting their data/privacy rights) and how over time the change of those feelings can be interpreted.

3 Affective States, Dimensions & Labelling

For the purpose of this project we chose the two "*strongest*" and "*easiest*" measurable emotions **happy** and **sad** as well as the **arousal** level. The main reasons for this were such states and dimensions are much simpler to collect and they also provide a good spectrum of opposite emotions

to explore for the limited time given. For similar reasons and to achieve more accurate results we decided arousal would be a better choice than valence and dominance as most individuals do not express very "*extreme*" emotions when listening to music. Valence was also not chosen as it provides many data points, although the project could be extended to take into account the valence as it is a good measure since Spotify's API provides measure of valence amongst other such as energy, temp and danceability of the songs.

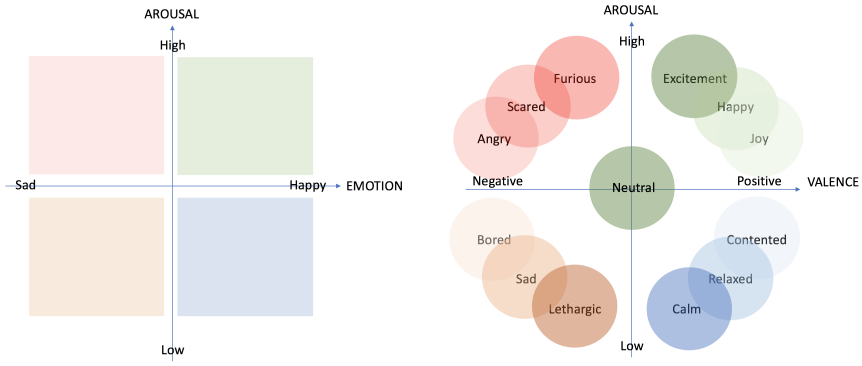

Figure 1: Left: The dimension and states used in the project, Right: An example of an improved model considering more emotions as well as arousal and valence

To understand deeply about data collection and labelling, our project consisted of experiments designed to collect relevant emotional information from participants where, facial and physiological data were collected. During the preliminary testing the collection of data was done after end of 3 videos rather than end of each individual video. This meant that participant would forget how they *truly* felt after a while. To get around this problem we created fill-in sheets which each participant had complete after each video. After each experiment only a small 5 second period was provided for the participants to return to their baseline, however, a more careful approach should be thought for to bring the participants to a "normal" baseline level.

To make the amount of data processing at a reasonable level, approximately, a frame was saved around each second (1fps). Each facial frame recorded from a "*happy*" or "*sad*" video was tagged with a 1 or a 0 respectively as shown in figure 2. Both arousal and emotions were collected quantitatively based on two questions which each participant filled after watching each video. However, this method could be improved if the labelling were done per frame (i.e detecting different emotions through the same video) rather than labelling the same label for all the frames for a video.

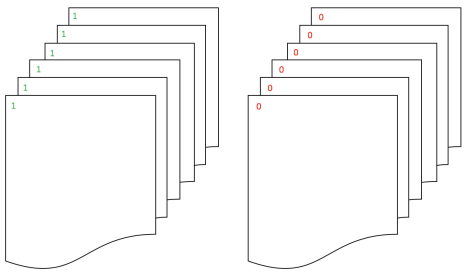

Figure 2: Labelling of a happy and sad video. happy videos tagged as 1 and sad videos tagged as 0

Additionally, as we were in the room with the participants, this may have affected how they reacted to the videos (i.e embarrassment or not being comfortable as they are *watched*) which could cause a bias in the results collected. On top of that, some participants knew some songs shown to them which again could cause some bias and this may increase or decrease the arousal of the participant. Because of the difficulty and the lack of time to collect enough data it was decided to use an external labelled dataset for both physiological and facial from DEAP and Kaggle competition dataset "Challenges in Representation Learning: Facial Expression Recognition Challenge".

Furthermore, we limited this project to only two crude forms of emotions, where in reality emotions are complex and multiple emotions could be derived from a single interaction. Moreover, the order of the videos could introduce a bias, a randomised effort on a larger dataset would remove bias in the expectation.

Eliciting emotions purely on listening to music and music videos proved very difficult as different music provided different subjective views for participants. It also only allowed very subtle emotional changes which were hard to distinguish from each other. To elicit better emotions and to receive a more diverse range, other sources of videos and experiences were used, such as a video of an animal in distress (for eliciting high sadness) and an interactive funny game (for eliciting high happiness).

4 Relevance and Effectiveness of Sensors and Modalities of Expressions

For the purpose of this project two different kinds of sensors were used to measure both the physiological as well as the physical aspects. To do this we decided on choosing a camera and a GSR sensor.

4.1 Facial Recognition - Webcam

Firstly, a camera was opted for as it is fairly cheap to acquire as well as most laptops and phones these days include a camera. The sensor is fairly accurate in capturing participants face and normally when a user is using their phone or laptop, the camera is roughly aimed towards the direction of their face. Certain features that mattered of the face to detect emotions were the eyes, eyebrow movements, cheeks and the mouth. Additionally, different algorithms are also available for facial recognition in regards to emotions, hence very appropriate for industry based applications.

Some limitations of using facial recognition is that when participants listen to music they often have very subtle changes of emotions which does not affect the face. Also elements such as background, wearing accessories, covering of the face, brightness levels of the environment could have an impact of the capture. Additionally, in the real world, if for example a user is using their phone, the face may only partially be visible due to the angle they are holding the phone.

4.2 Physiological - GSR

Secondly, a GSR sensor was used to understand a participants emotional arousals. GSR was chosen as it can reveal "*hidden*" arousal of an individual. With facial recognition a participant can choose to not reveal their true feelings, however, with physiological data, these are often things

that are uncontrollable and are based on instinct. Hence, by understanding physiological data well, perhaps we can understand the true emotions of a person. Therefore, the goal was to use both types of voluntary and involuntary responses to create an accurate emotion recognition system.

Having said that, GSR requires a steady hand and for it to be kept still to record meaningful values. Hence, in the real world since many people tend to use their hands often, it would be disruptive to the sensor. It would also require at least a smart watch for the user to wear at all times which is much less common than the camera phone for example. Other possible changes such as the position of the sensor, the tight fitting as well as direct skin connection could cause the results to be inaccurate.

4.3 Other Modalities

We are not limited to only the modalities explained above, other modalities could bring valuable information in conjunction to the modalities described above. Other physiological types of data could be collected such as temperature, breathing patterns and the heart beat of an individual to understand arousal. Electroencephalogram (EEG) could also be used to track and record brain wave patterns which may give more meaningful features, although it is very invasive and not matured enough for everyday use.

Having mentioned such modalities it is good to consider the ethical implications of these methodologies. The data collected from all the sensors are extremely personal and delicate. The face for example highlights the identity, the race, age and many other features of an individual. Additionally, the constant recording of the physiological data could infer information about an individuals health and well-being. Data protection, sharing and anonymisation are key topics that need to be considered heavily if such a *product* is going to be commercialised.

5 Justify and Critique Feature Selection and Optimisation Process, Other Features

5.1 Feature Selection & Optimisation Process

For our facial classification, opting for a deep learning methodology, specifically the use of convolutional neural networks (CNN), it was possible to automatically select features that are most useful and relevant to the problem we are tackling. Deep learning and in particular VGGNets in its layers performs feature selection, as they learn the features from the data instead of handcrafted feature extraction methodologies. However, for this to be done well you would require a very large amount of dataset. We can clearly see that by running the model through our own dataset the best result were approximately around 58% due to the lack of data presented. Conversely, when used on the bigger dataset, 86% accuracy were achieved.

Furthermore, since the original dataset compromised with *emphasised* emotions, rather than experimentally elicited data, the features that the model may have learnt may be sub-optimal towards subtler emotions expressed. We believe deep learning is the way forward for predictions and feature selection, however, only when a large enough dataset is available.

The GSR features extracted to train the models were followed by previous research [13] that suggested that features of interest with most success would be the mean, standard deviation and the mean of first and second derivatives. Although, research practises were followed in order to extract the best features, the results viewed by both our own dataset as well as the DEAP dataset suggested that the features extracted from GSR data were not useful. One reason for this could be that GSR alone may not be sufficient enough to predict the arousal level.

5.2 Other Features

There are other features that could have been explored if the scope of the project was extended. For example there are numerous novel approaches to extract image features such as "*Binary Pattern for Texture Classification*" which is robust to image rotation and noise. Other methods include stacking the extracted features through extra channels within the CNN as well as other feature descriptors such as "*Histogram of Oriented Gradients*" could have been used to see if the model improved in results. One way of improving the GSR feature selection would be to not solely rely on one physiological data and to combine other physiological measures such as temperature or electrocardiogram (ECG).

6 Modelling Algorithms

For facial emotion classification we used VGGNets a particular CNN which holds the power of CNN's but maximises the data-driven feature extraction approach which minimises human bias. VGGNets have proven to be very successful in computer vision and for facial emotion classification tasks. This is due to the fact that if you stack VGGNets layers, these combine in much more powerful, complex and advanced transformations.

Five different regression models were selected to predict the arousal felt by participants. Random forest regression, support vector regression (SVR) and linear regression (including LASSO and Ride). Every model provided their pros and cons, from simplicity of linear regression although prone to over-fitting to more complex models such as SVR and random forest which have advantages of capturing non-linearity relationships.

7 Optimisation And Evaluation Methods

7.1 Optimisation Methods

Optimisation methods were induced in both of the models described above. As complex architectures can learn noise and overfit, the dropout technique was used that boosted performance for the VGGNets. Additionally, the hyper parameters to tune were the number of layers and filters. By exploring different hyper parameters such as using an 8 layer VGG Block and doubling the filters proved the best results.

The regression models were optimised through parameters such as k-fold cross validation and grid search to give the most accurate parameters for every model. The simpler models only had a few parameters to be explored with whereas the more complex models had many. To achieve the best overall performance a 3 folds approach were used and the errors averaged.

7.2 Evaluation & Performance

The evaluation metrics used consisted of mean squared error (MSE) and R-Squared error. They both serve their own purpose as MSE is the average squared error between the ground truth values and the model's prediction, whereas R-Squared measures how well the data fits to the model's regression line.

The performance of the facial emotion classification on a large dataset was relatively good. The best model achieved a binary accuracy of 86% with 0.85 precision, 0.81 recall and 0.87 area under the precision-recall curve. This shows that faces can be very impactful when detecting emotions. On the other hand the model based on self collected data managed a performance of 54% accuracy for the binary classification. This was improved to 58% accuracy by making each input in a video and taking the mode. The final improvement was done by averaging over the class probabilities

which achieved an accuracy of 61%. This shows that it is much harder to obtain high quality data through experiments and many obstacles such as subtle emotions as well not having a large enough dataset can hugely impact the performance of the models.

Unfortunately, the performance of all the regression models were very poor without any meaningful results for detecting the arousal level of participants. A method of improving this result is to try different groups of features as well as combining the features of other physiological measures.

7.3 Final Remarks

The performance of the classifier for the arousal of a participants is no where near sufficient for the purpose of our application. An argument could be made that the classification generated by the facial emotion classification is "*good*". However, one should consider that for a real world application, in this case Spotify recommending music based on an individual's emotions, acquires million of users and even with a 86% accuracy the amount of users affected will be in the hundreds of thousands.

Therefore, a very accurate classification is needed (i.e 99%) for any applications in the real world, as we would need to minimise the number of users affected by the technology proposed as it could otherwise have very severe consequences. In addition to that, a special attention should also be payed towards data protection and privacy of an individual as very personal information is being collected. Having said that the findings of this project has been valuable to recognise the value of facial emotion recognition and the problems with physiological data. As well that we have also shown that the technology used in this study is already available and increasing in popularity.

References

[1] G. Kuan, T. Morris, and P. Terry, "Effects of music on arousal during imagery in elite shooters: A pilot study," *PLOS ONE*, vol. 12, pp. 1–13, 04 2017.

[2] D. Duncan and G. Shine, "Facial emotion recognition in real time," 2016.

[3] C. Lin, M. Liu, W. Hsiung, and J. Jhang, "Music emotion recognition based on two-level support vector classification," in *2016 International Conference on Machine Learning and Cybernetics (ICMLC)*, vol. 1, pp. 375–389, July 2016.

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[5] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: A convolutional neural-network approach," *IEEE transactions on neural networks*, vol. 8, no. 1, pp. 98–113, 1997.

[6] M. Matsugu, K. Mori, Y. Mitari, and Y. Keneda, "Facial expression recognition combined with robust face detection in a convolutional neural network," in *Proceedings of the International Joint Conference on Neural Networks, 2003.*, vol. 3, pp. 2243–2246, IEEE, 2003.

[7] M. Matsugu, K. Mori, Y. Mitari, and Y. Kaneda, "Subject independent facial expression recognition with robust face detection using a convolutional neural network," *Neural Networks*, vol. 16, no. 5-6, pp. 555–559, 2003.

[8] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[9] "Challenges in representation learning: Facial expression recognition challenge." https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge.

[10] S. Koelstra, "Deap: A dataset for emotion analysis using physiological and audiovisual signals." https://www.eecs.qmul.ac.uk/mmv/datasets/deap/.

[11] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, p. 1831, 2012.

[12] J. Kim and E. Andre, "Emotion recognition based on physiological changes in music listening," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 12, p. 1831, 2008.

[13] P. Lang, M. Greenwald, M. Bradley, and A. Hamm, "Looking at pictures: Affective, facial, visceral, and behavioral reactions," *Psychophysiology*, vol. 30, no. 3, pp. 261–273, 1993.

[14] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. Springer, 2013.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[16] L. Q. William Forde Thompson, "Music and emotion: Psychological considerations," January 2012.

[17] Z. R. Blood AJ, "Intensely pleasurable responses to music correlate with activity in brain regions implicated in reward and emotion.," sep 2001.

[18] O. S. A. Sloboda, John A, "Emotions in everyday listening to music," *Series in affective science. Music and emotion: Theory and research*, pp. 415–429, mar 2001.

[19] R. B. L. F. U. E. U. M. N. Myriam V. Thoma, Roberto La Marca, "The effect of music on the human stress response," apr 2012.

[20] C. Wilson, "The impact of music on task performance at work," may 2018.

[21] M. M. Shirlene Vianna Moreira, Francis Ricardo dos Reis Justi, "Can musical intervention improve memory in alzheimers patients? evidence from a systematic review," p. 133142, apr 2018.