Affective Computing and Human Robot Interaction Mini Project Group Report

Julia Gomes UCABJMG@UCL.AC.UK
Wachira Ndaiga (17095209) UCABSWN@UCL.AC.UK
Jade Savage (40345046) UCABJM5@UCL.AC.UK
Jimmy Xu (12034279) UCABJX1@UCL.AC.UK
Group name: Robo Dancers

1. Introduction

1.1. Background

With the rapid growth of media and entertainment consumption in recent years, online streaming services and content makers are now seeking novel methods to index and categorise their multimedia content to improve user engagement. Because the affective states of a user can influence greatly their choices of multimedia content and reflect their preferences for the content (Soleymani et al., 2008), users' emotional reactions could be used to label multimedia content and help create customised content recommendation systems through a better understanding of users' preferences.

Emotions can be categorised using a number of different methods. Both discrete categorisation, the six basic emotions by (P. Ekman & Ricci-Bitti, 1987), and continuous categorisation, the valence-arousal scale by (Russell, 1980), are used in this project. Results from both categorisations could convey useful information, as multimedia commonly elicits one of the six emotions, and valence-arousal levels indicate value and importance (Clore & E Palmer, 2009).

1.2. Purpose of Project

The purpose of this project is to create an automatic emotion recognition system capable of detecting different affective states of users while they are watching movies or TV shows. This particular system implemented in the project aims to detect seven different emotions (Sad, Disgust, Anger, Neutral, Surprise, Fear, and Happy) from facial expressions, and classify high/low valence and arousal levels from physiological signs, using machine learning models.

To achieve this, we carried out experiments where participants were shown a series of videos chosen to elicit specific emotions, and recorded their facial expressions and physiological signs. The collected data, combined with two larger public data sets, are then used to train a deep learning facial emotion recognition model and two binary physiological data classifiers for high/low valence and arousal.

2. Experiment Setup

2.1. Stimuli Selection

We manually selected two videos for each of the 6 Ekman basic emotions (P. Ekman & Ricci-Bitti, 1987) with an additional neutral emotion included for completeness. We only considered videos which were short and complied with UCL regulations, then voted for the ones that elicited the strongest emotion. We compiled the videos in such a way that the entire experiment would take less than 20 minutes. Stimuli were also limited to one scene to minimize the emergence of subtle emotions that would interfere with the elicitation of the target affect. In total, 31 stimuli videos were collected for further analysis.
The consensus step was formulated using a standard vote-based stimuli selection criteria with both the first and second in class videos forwarded for media preprocessing. Table 1 shows the average stimuli durations for each emotion with Surprise and Neutral taking the shortest and longest durations respectively.

2.2. Materials and Setup

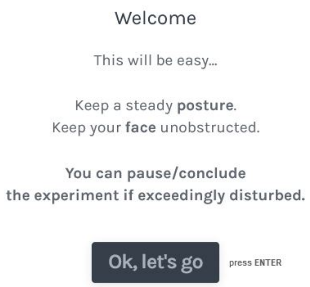
The experiment was performed in a single laboratory with controlled, fluorescent illumination and minimal acoustic/environmental disturbances. The latter was of noted importance as emotion elicitation in a large portion of the stimuli set was dialogue/sound predicated.
Peripheral GSR (Galvanic Skin Response) and BVP (Blood Volume Pulse) physiological signals were

Emotion (Modified Ekman)	**Average Duration** (seconds)
Anger	44
Disgust	15
Fear	53
Happy	42
Neutral	60
Sadness	56
Surprise	10

Table 1. Average duration of emotion-selected stimuli

recorded using an Empatica E4 wristband. Frontal face video was captured using a 720p HD webcam with the ffmpeg tool used to capture the latters' input stream as well as each participants' web-based experiment session. To present the stimuli and record participant self-assessment, the online form response tool, Typeform was used. The latter interface included participant prompts and notices to ensure they were appropriately guided through the experiment. This proved to be inadequate with a possible improvement being the inclusion of an unrecorded practice trial prior to at the commencement of the experiment.

The experiment started with a kind welcome and gentle reminder to ensure they do not occlude or change their posture. This is shown in Figure 1.



Welcome

This will be easy...

Keep a steady **posture**.
Keep your **face** unobstructed.

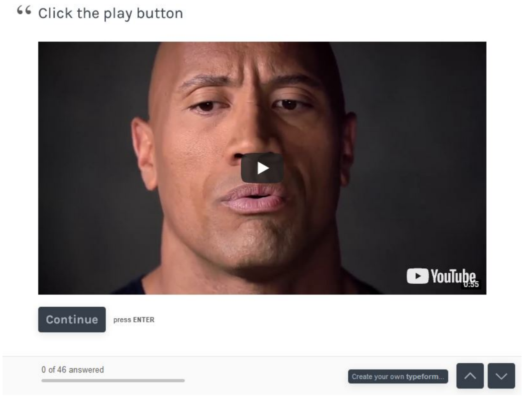
You can pause/conclude
the experiment if exceedingly disturbed.

Figure 1. Screenshot of the start notice of the experiment

This was then followed by a 1 minute baseline recording where a fixation cross was displayed to the user. The following details the constituent elements of the trials run thereafter:

- Tag an empatica event on wristband.
- Stimuli presentation.
- Explicit emotion assessment.
- Self-Assessment Manikins (Valence, Arousal and Dominance).
- Stimuli familiarity.
- Baseline regression activity.

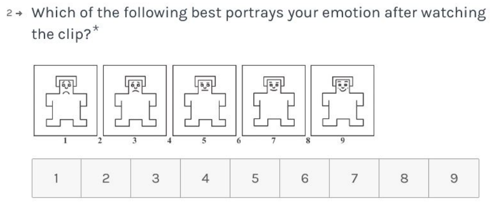
The stimuli presentation was ordered to conform to the ethical constraint that participants exit the experiment in positive mood/state. As such, the Happy emotion stimuli was placed last. A screenshot of the stimuli presentation interface is shown in Figure 2. To minimize eye movements, the presented video stimuli was set to 700 by 400 resolution. Participants were sat at an arms-length from the display screen with the experiment laptops' in-built JBL speakers set to a mid-range volume level.



“ Click the play button

Figure 2. Screenshot of the web interface for stimuli presentation and participant self-assessment

A discrete SAM model was used with the participants able to select numbers between 1 and 9 for all affective dimensions (valence, arousal and dominance) as shown in Figure 3. This could be enhanced by implementing a continuous scale, however the Typeform interface does not provide this capability.



2→ Which of the following best portrays your emotion after watching the clip?*

Figure 3. SAM modelling as implemented for the valence dimension

Participant self-assessment additionally included explicit emotion assessment implemented as a single-choice option list on the web interface.

2.3. Participant Recruitment

Participant recruitment was performed by utilising each experimenters' social network. Given that the participant sample size was small ($n = 4$; with 3 male and 1 female and a participant age range of 19-24), this method worked adequately for our purposes.

3. Subjective Analysis of Collected Data

Because we only collected data from four participants, we did not acquire enough data to build a statistical or deep learning model using our dataset alone. (We later build these models using larger datasets from Kaggle and DEAP). However, it was still a valuable experience to go through the data collection process. In this section, we subjectively analyze whether the experimental design successfully elicited the desired affective states.

Each participant was shown a series of seven videos to induce the following emotions, in order: Sad, Disgust, Anger, Neutral, Surprise, Fear, and Happy. After watching the videos, participants reported which emotion they felt (or "None of the Above"). We found that the majority of videos successfully induced the desired emotional response. Specifically, in 17 of the 28 experiments (61%) in which a participant watched a single video, the participant selected the emotion which the video was intended to elicit. Two emotions in particular were successfully elicited in 100% of experiments: "disgust" and "neutral". This is not surprising, as these are relatively primal and unambiguous emotional states.

For the remaining 11 experiments, there seemed to be a few different reasons why the participant selected a different emotion than intended. The most likely explanation is that a few of the emotions are somewhat ambiguous in meaning. For example, we used a political video to try to induce "anger". We found instead that a few participants selected "disgust" instead of "anger", likely because "disgust" can be interpreted in more than one way, as "offensive" or "nauseating." Another common inconsistency was to report "fear" toward videos that were meant to incite "surprise". One of the videos we chose for "surprise" involved a scary face jumping out at the camera, so it makes sense that "surprise" and "fear" were conflated.

Overall, we found it was difficult to predict which videos would be most effective. None of the videos we used in this experiment incited a particularly strong reaction. Participants maintained a straight face and relatively steady heart rate for nearly all the videos, which is another reason why we are using alternative datasets. We may have cut the videos too short to incite a strong reaction in the desired affective state. For example, in one of our "sad" videos, we end the clip before the saddest part of the video (in order to not offend any participants), which may have prevented participants from building up a stronger emotional response.

Despite these setbacks, we did obtain some interesting results. Our main realization was that body movement served as a stronger visual indicator of emotion than facial expression. Specifically, when a participant viewed one of the "anger" videos, his facial expression was unchanged but he visibly reacted by moving his head back and forth. This suggests that a comprehensive visual classifier of emotional state would need to incorporate body movement and gaze-tracking into its model in addition to facial expression.

4. Machine Learning Models

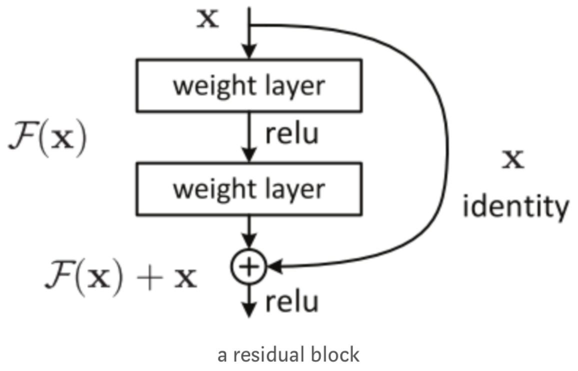
4.1. Deep Learning for Facial Emotion Recognition

Facial emotion recognition (FER) has made significant improvements in recent years due to advances in deep learning techniques for computer vision tasks. Prior to deep learning, conventional FER approaches involved extracting hand-crafted geometric features and appearance features to construct a feature vector. Machine learning researchers would then train a support vector machine (SVM) or multi-class AdaBoost on the data to categorize the facial image as one of the target emotion. (Ko, 2017).

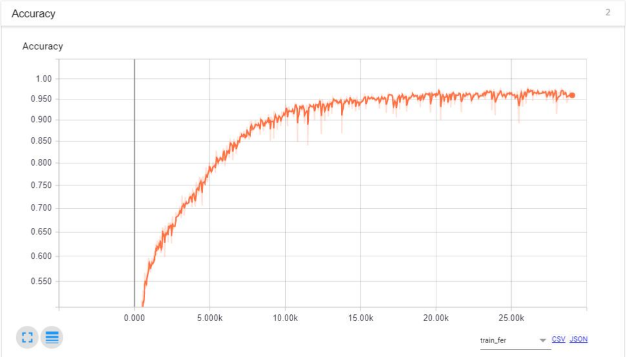
Researchers can now use deep learning techniques to learn the best features from the images, instead of relying on human expertise regarding which features to extract. We chose to take this deep learning approach when classifying the seven basic emotions. However, this sub-field of computer vision is relatively new and there are few datasets that contain enough examples of labelled facial expressions to train a deep learning model. The largest dataset we found available online was the FER-2013 dataset on Kaggle. It consists of over 30,000 48x48 gray-scale images centered on faces expressing the 7 basic emotions: fear, anger, happiness, sadness, disgust, surprise, and neutral. The test set include 3,589 images.

Initially, we planned to take a popular, heavily researched network (such as ResNet) pre-trained on the massive ImageNet dataset for object recognition, then use transfer learning to re-train the network on our smaller dataset for facial expression recognition, freezing the first few layers and fine-tuning the rest of the network. However, we realized that we could not use transfer learning on the FER-2013 dataset because the ResNet and most other popular machine vision architectures are designed for RGB images, while the FER-2013 is grayscale. Fortunately, we found a CNN on Github that was specifically designed for the FER-2013 dataset (Freitas, 2017). The Github user safreita1 claimed a relatively high top-1 classification accuracy of 94.8%.

We modified the optimizer to the Adam optimizer (recommended in class) in an attempt to boost the performance, then retrained the model ourselves using Amazon AWS to see if we achieve similar or higher accuracy. The images are zero-centered and normalized in the pre-processing stage. During training, images are randomly flipped in the left/right direction. The network includes nine layers: a convolutional layer with 16 filters of size three, five residual blocks increasing from 16 neurons to 64 neurons, and a final fully-connected layer with a softmax activation to output the seven class probabilities. Residual blocks (illustrated below) are ideal for unfamiliar tasks because they are capable of learning the identity function, allowing them to imitate a shallower network if the model is too deep and thereby preventing the risk of vanishing gradients or saturation (Kaiming He). The network also uses batch normalization, standard relu activation, and a weight decay factor of 0.0001. The network trained over 150 epochs.


x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu
a residual block

The figure below shows the evolution of the training accuracy over time. The training accuracy reaches a little over 95%, leveling off at around 15,000 epochs. Despite the high training accuracy, the final top-1 test accuracy was only 58.9%, which suggests we may have over-fit the model to the training set. It is also surprising that the accuracy of our network was so much lower than that of the Github user safreita1, as we were hoping to verify the performance of their model. However, the accuracy of 58.9% still suggests the model was able to learn the basics of facial emotions recognition, considering the accuracy of random guessing would only be 14.2%.



Next, we fed our own images into the network. As mentioned earlier, none of the participants in our experiment showed strong facial expressions, but we input 6 images where participants appeared to demonstrate slight reactions to the videos and subtle micro-expressions. A few sample images are displayed below.

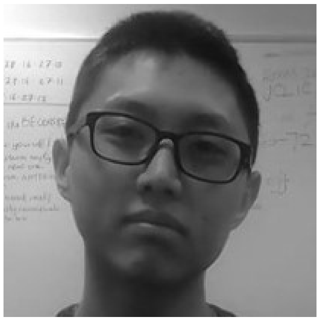


Figure 4. Frontal face capture of Participant 1

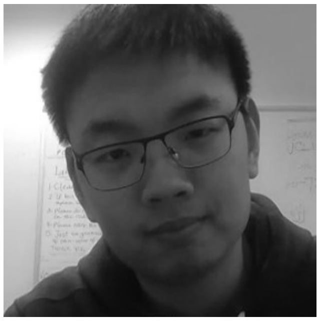


Figure 5. Frontal face capture of Participant 2

Figures 4, 5 and 6 are frontal face captures of three participants. These frames were extracted from web camera video footage taken during the anger elicitation

Figure 6. Frontal face capture of Participant 3

trial. The above images were subjectively chosen after a candidate frame selection process that involved timeframe matching between each participants' screen and frontal face video captures. Certain facial cues are of note, such as the pursed/pressed lips visible in fig 4 and 5 as well as a slightly furrowed brow in fig 6; both expressions known to indicate mounting aggression (E. Sedenberg E., 2017). Interestingly, the model classified all of the images as "angry". Looking at the images of the participants, this is not too surprising, considering the down-turned mouths, narrow eyes and other features. However, we would have expected more "neutral" classifications, which is a more accurate description. Clearly the network is no better at classifying micro-expressions than the human eye.

4.2. Classification using Physiological Data from DEAP

We trained two binary classifiers (low/high valence and low/high arousal) using data provided by DEAP (Koelstra et al., 2012). Initially, we were hoping to produce a single classifier for all participants (population-level as opposed to participant-level). This way we could test the compatibility of our data and assess the impact of having a different experiment protocol for data collection. However, the data did not separate well under the 'global' set-up, and so we decided to opt for a subject-specific classification approach. Refer to figure 7 for a comparison of maximum 'Fisher linear discriminant' values achieved with both methods (larger values indicate higher 'separability' — the formula is outlined in equation (1)).

Of the DEAP dataset, we considered the galvanic skin response (GSR) and blood volume by plethysmograph of 32 participants across 40 videos, (1,280 trials in total, data sampling rate of 128Hz). This corresponds to EDA and BVP output, respectively, from the Empatica bracelet. Altogether, 29 features were extracted from those attributes, to be used as potential inputs (refer to table 2 for a full breakdown (Ayata et al., 2017)). The participant ratings were used as the ground truth, and a threshold placed in the middle of the 1-9 scale (at 5), so that videos rated up to five were classified as 'low' and those rated higher than five were classified as 'high'.

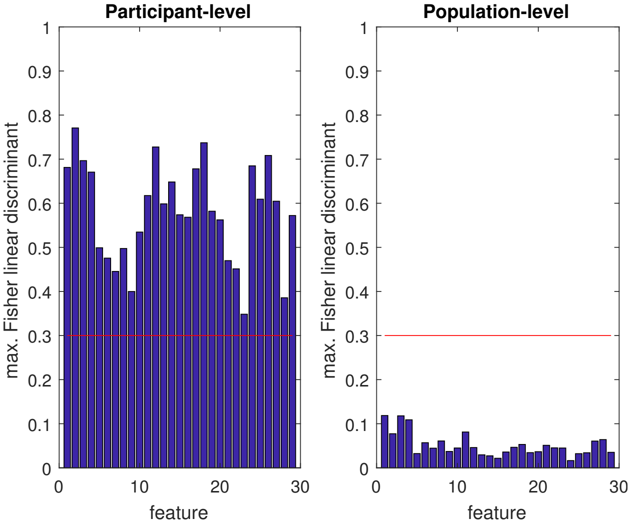


Figure 7. Maximum Fisher's linear discriminant of each feature: participant-level (left) vs. population-level (right) valence classifiers; each bar represents one of the features listed in table 2; both charts have [0,1] y-axis to emphasize low 'separability' at the population level; red line indicates adopted threshold of 0.3

	Extracted features
GSR	minimum, maximum, arithmetic mean, median, standard deviation, variance, 3rd moment, 4th moment, 5th moment, skewness coefficient, kurtosis coefficient, number of zero-crossings, energy, number of peaks
BVP	minimum, maximum, arithmetic mean, median, standard deviation, variance, 3rd moment, 4th moment, 5th moment, skewness coefficient, kurtosis coefficient, number of zero-crossings, energy, mean HR*, standard deviation of HR

Table 2. Low-level features extracted from physiological signals; *HR = heart-rate

We performed leave-one-(video)-out cross validation for each subject and assessed the strength of each classifier by considering the average accuracy and F1-scores (across participants, F1-score averaged over both classes). At each step of the cross-validation, features were first standardised to have a mean of 0 and standard deviation of 1, and then training features were selected using Fishers linear discriminant, with a threshold value of 0.3, in line with the DEAP paper. The formula is given by:

$$J(f_i) = \frac{|\mu_{i1} - \mu_{i2}|}{\sigma_{i1}^2 + \sigma_{i2}^2} \tag{1}$$

where μ_{i1} and σ_{i1} are the mean and standard deviation of the ith features that classified as low, and μ_{i2} and σ_{i2} are the mean and standard deviation of the ith features that classified as high. A Gaussian naïve Bayes classifier was subsequently used to classify the emotional response to the test-video. The classifier (naïvely) assumes independence of all features, and is given by:

$$G(f_1, ..., f_n) = \underset{c \in \{\text{low,high}\}}{\text{argmax}} \; P(C = c) \prod_{i=1}^{n} P(F_i = f_i | C = c)$$

$$= \underset{c \in \{\text{low,high}\}}{\text{argmax}} \; P(C = c) \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_{ic}^2}} \exp\left(-\frac{(f_i - \mu_{ic})^2}{2\sigma_{ic}^2}\right)$$

$$= \underset{c \in \{\text{low,high}\}}{\text{argmax}} \left(\log(P(C = c)) + \sum_{i=1}^{n} -\frac{1}{2}\log(2\pi\sigma_{ic}^2) - \frac{(f_i - \mu_{ic})^2}{2\sigma_{ic}^2} \right)$$

where f_i represents a selected feature. The mean (μ_{ic}) and standard deviation (σ_{ic}) of the constituent Gaussian (for that particular feature) is computed from the training data, and P(C=c) is determined by simply counting the overall occurrence of each class across the whole training set.

	Valence			Arousal		
	ACC	**F1**		**ACC**	**F1**	
Robo Dancers	0.563	0.563	**	0.579	0.564	*
DEAP	0.627	0.608	***	0.570	0.533	**
Random	0.500	0.494		0.500	0.483	
Majority class	0.586	0.368		0.644	0.389	
Class ratio	0.525	0.500		0.562	0.500	

Table 3. Average accuracies (ACC) and F1-scores (F1, average of score for each class) over participants when classifying low/high valence and low/high arousal according to perihperal physiological features. Stars indicate whether the F1-score distribution across subjects has a mean significantly higher than the 'class ratio' of 0.5 according to an independent one-sample t-test ($* * * = p < 0.01$, $** = p < 0.05$, $* = p < 0.10$). Refer to the DEAP paper for detailed explanations of 'random', 'majority class' and 'class ratio', which are given here for comparison.

Average accuracies and F1-scores are displayed in table 3, and are broadly comparable to the DEAP paper (which used more extensive features). With our technique, both classifiers were found to have an F1-score higher than 0.5, but results were slightly less significant than those for DEAP based on an independent one-sample t-test (valence, p = 0.0225; arousal, p = 0.0720).

Confusion matrices are depicted in figures 8 and 9. A good classifier will have the highest values down the main diagonal (highlighted in blue). That is the case for our classifiers, although misclassification is still high.

Predicted class

Actual class	**H**	**L**	**total**
H	414	294	708
L	265	307	572
total	679	601	

Figure 8. Confusion Matrix of results for **valence** classifier; H = high, L = low

Predicted class

Actual class	**H**	**L**	**total**
H	467	270	737
L	269	274	543
total	736	544	

Figure 9. Confusion Matrix of results for **arousal** classifier; H = high, L = low

5. Conclusions and Future Directions

References

Ayata, Deger, Yaslan, Yusuf, and Kamasak, Mustafa. Emotion Recognition Via Galvanic Skin Response: Comparison of Machine Learning Algorithms and Feature Extraction Methods. 17(1):3129–3136, 2017. ISSN 13030914.

Clore, Gerald and E Palmer, Janet. Affective guidance of intelligent agents: How emotion controls cognition. 10:21–30, 04 2009.

E. Sedenberg E., J. Chuang. Smile for the camera: Privacy and policy implications of emotion ai. 2017.

Freitas, Scott. Resnet emotion recognition. https://github.com/safreita1/Resnet-Emotion-Recognition, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren Jian Sun. Identity mappings in deep residual networks.

Ko, Byoung Chul. A brief review of facial emotion recognition based on visual information. 2017.

Koelstra, S., Muhl, C., Soleymani, M., Lee, J. S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., and Patras, I. Deap: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, Jan 2012.

P. Ekman, W. V. Friesen, M. OSullivan A. Chan I. Diacoyanni Tarlatzis K. Heider R. Krause W. A. LeCompte T. Pitcairn and Ricci-Bitti, P. E. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology*, 53(4):712–717, 1987.

Russell, James. A circumplex model of affect. 39: 1161–1178, 12 1980.

Soleymani, M., Chanel, G., Kierkels, J. J. M., and Pun, T. Affective characterization of movie scenes based on multimedia content analysis and user's physiological emotional responses. In *2008 Tenth IEEE International Symposium on Multimedia*, pp. 228–235, Dec 2008.