Implicit Affective Tagging

Affective Computing and Human-Robot Interaction

Iulian Vlad Serban

1 Implicit Affective Tagging

1.1 Introduction

Implicit affective tagging is the process of automatically extracting affective tags for media content based on analysing a user experiencing the media. The process is implicit, because there is no need for the user to record any labels. For example, a music video could be tagged as *exciting* if the user was filmed shaking her head while watching it.

The goal of this project is to develop a system to extract implicit affective tags for music videos based on physiological signals.

1.2 Why implicit affective tagging?

Hiring people to tag media content manually is both slow and costly. An alternative to manual tagging is machine-based tagging. For example, video analysis algorithms can produce tags based on features of the content. However these algorithms have proven to be ineffective, as they do not understand the context of the media [1].

Why do we expect implicit affective tagging to work? Reeves et al. discuss the *media equation*, which proposes that humans react to media content in the same way they would react if the same events were happening in real life [2]. Based on this, Pantic et al. argues that since affective response is known to contain information about real events, affective response to media must also contain information about the media [3]. Indeed, many empirical studies indicate that media content (such as videos, pictures and music) can be effectively discriminated by affective responses [4], [1] [5]. [6], [7]. [8].

In addition, we should expect implicit tagging to be more robust than explicit manual tagging for two reasons. Firstly, the users are not interrupted during the experience process. The tags are therefore spontaneous and do not suffer from conscious interpretation of the content, which would occur if the users were asked to tag the media explicitly [3] [9]. Secondly, the implicit tags are based on a set of basic emotions, which are universal to human beings to a large extend [3]. They are less influenced by culture and should therefore generalize across users.

1.3 Applications & Ethical Implications

Tags obtained from implicit affective tagging have a number of important applications in media-content-systems. The tags may be used to evaluate the accuracy of the manual tags, or they may be treated on par with the explicit tags to improve search performance [1] [3]. The implicit tags may also be used for user profiling through collaborative filtering [3].

Implicit affective tagging also enables dubious and malicious applications. Since the process works primarily at a low conscious level, the same system may be used to design advertisement content which by-passes the observers conscious thinking and affects them subconsciously. The system might also covertly be used to extract unwanted information from the users interacting with it. For example, it could be used to extract political opinions whenever

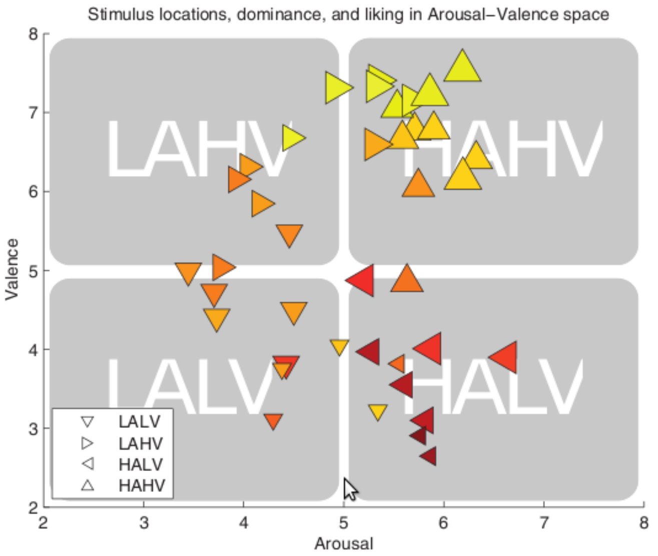


Fig. 1: The mean locations of the stimuli on the arousal-valence space for the 4 conditions (LALV, HALV, LAHV, HAHV). Liking is encoded by color: dark red is low liking and bright yellow is high liking. Dominance is encoded by symbol size: small symbols stand for low dominance and big for high dominance. Taken from Figure 6. in [11].

users watched video clips of political events. This would clearly break the users privacy, as discussed by Fairclough [10].

2 DATABASE

2.1 Description

We use the database DEAP as described by Koelstra et al. in [11]. The database was based on 32 participants, who each watched a 1 min. music video clip from a total of 40 music videos. Physiological signals for each participant were recorded during each clip. The biosensors used were Galvanic Skin Response (GSR), blood volume pressure (BVP), respiration, skin temperature, Electromyography (EMG), Electrooculography (EOG) and Electroencephalography (EEG). All biosensors recorded at 512Hz sampling rate. The placement of the peripheral biosensors are shown in Figure 2.

After each clip participants self-reported on the continuous affective dimensions valence, arousal, dominance and liking on a scale 1-9. Participants used Self Assessment Manikins (SAMs) as shown in Figure 3. Please see appendix for a more details and a discussion of possible issues relating to our project.

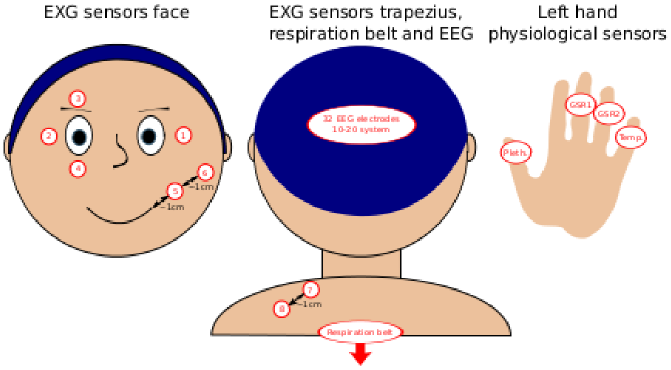

EXG sensors face
EXG sensors trapezius, respiration belt and EEG
Left hand physiological sensors
1
2
3
4
5
6
~1cm
~1cm
32 EEG electrodes 10-20 system
7
8
~1cm
Respiration belt
GSR1
GSR2
Temp.
Pleth.

Fig. 2: Placement of peripheral physiological sensors. Four Electrodes were used to record EOG and four for EMG (zygomaticus major and trapezius muscles). In addition, GSR, BVP, temperature and respiration were measured. Taken from Figure 3. in [11].

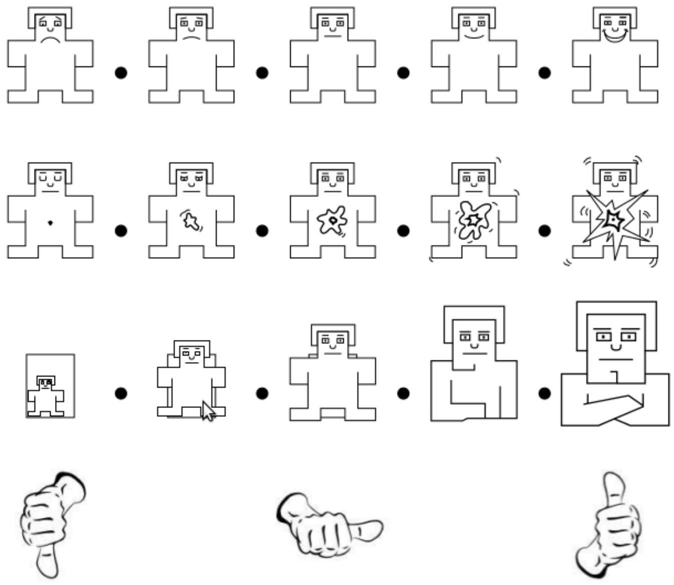


Fig. 3: Images used for self-assessment. Manikins from top to bottom: valence, arousal, dominance and liking. Taken from Figure 5. in [11].

In our work we will make use of the peripheral biosensors: GSR, BVP, respiration, skin temperature, EMG and EOG. These modalities are known to contain a significant amount of affective information for characterising images, video and music [7] [9] [4] [8]. We will refer to the participants as subjects, and to the music videos as either media or stimuli.

3 Methodology

3.1 Tags in Continuous Space

We are going to use the self-reported ratings as the ground truth w.r.t. each subject, and their mean as the ground truth w.r.t. the population. That is, we assume there exists an individual label for each subject and a population label for all subjects. Predicting the population label is the end goal of our system. Taking the self-reports as ground truth could lead to various modelling problems. Ratings could be influenced by culture, personality, mood etc. However, we expect the usage of SAMs to alleviate some of these problems [12].

We will use the affective dimensions valence, arousal, dominance and liking. Valence and arousal have been able to distinguish between a variety of stimuli types, ranging from music to images and advertisements [13] [8] [7] [14] [15]. Dominance has been used to characterize music and advertisements [8] [7]. We also make use of the liking rating, because it captures the notion that on average some media are preferred over other media.

Pantic et al. finds that it is unclear whether a continuous representation or a discrete representation is best for implicit affective tagging [3]. Therefore we evaluated the discriminative power of partitioning valence, arousal, dominance and liking into (low, high) values, where the threshold was set to five. We calculated the mean agreement levels and Fleiss' Kappa across all subjects. The calculated levels range from *low* to *fair*, which indicates that the discrete representation would be a poor choice.

TABLE 1: Agreement Levels

	Valence	**Arousal**	**Dominance**	**Liking**
Mean Agree.	0.696	0.574	0.592	0.667
Fleiss Kappa	0.381	0.119	0.134	0.244

We therefore chose to use continuous affective dimensions. Several studies have shown that media can be separated well by continuous affective labels from self-response [14] [15] [7]. From the continuous representation, any point can afterwards be mapped to a corresponding discrete label that users may then search for. The continuous representation also has the potential to capture more structure between media. For example, media tagged *unhappy* is more related to media tagged *annoyed* than it is to media tagged *enthusiastic*. This could be exploited by the search engine to improve the search.

We normalize all the labels to have zero mean and standard deviation one across all subjects. This normalization enables us to isolate the predictive performance of our models from any bias which may exist in the data.

3.2 Process

Our implicit affective tagging system is illustrated in Figure 4. We first construct regression models, which will predict labels for each (subject, stimuli) pair. The best model is then used to predict affective labels for each subject. We then take the mean of these predictions as the population label.

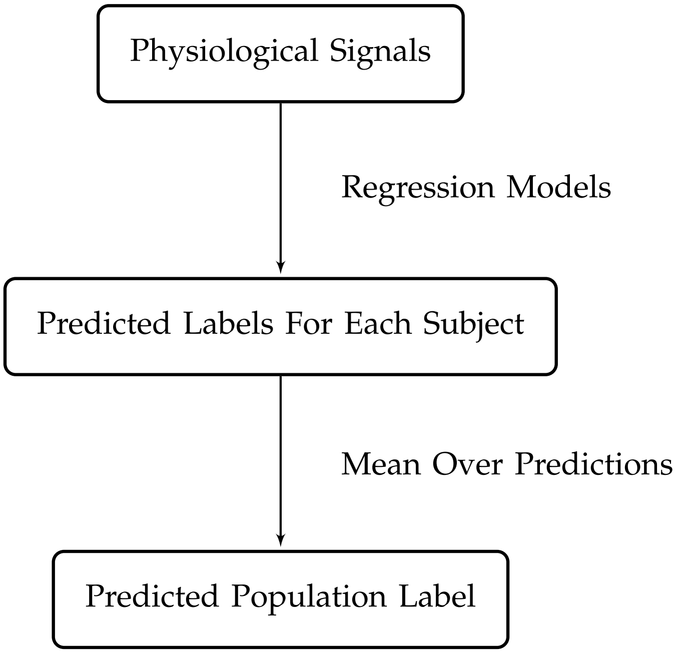


Fig. 4: Data flow of the implicit affective tagging system.

3.3 Subject-Independent Model vs. Subject-Dependent Model

It is an open research question whether or not our models should be subject-dependent. Studies have observed that subjects exhibit individual idiosyncrasies, and that their physiological signals w.r.t. the same stimuli varies considerably on a day-to-day basis [16] [17]. Furthermore, it has been observed that generalization over unseen subjects is a very difficult problem [6] [14]. Nevertheless, some researchers have been able to build subject-independent models successfully [5], [6], which indicates that there exists features which are correlated with affect independently of the subject.

This motivates us to experiment with both subject-independent and subject-dependent models. For the subject-independent model, we will train the model on 24 subjects and test it on the remaining 8 subjects using leave-one-subject-out cross-validation. For the subject-dependent model, we will train the model on 30 stimuli and test it on the remaining 10 stimuli using leave-one-stimuli-out cross-validation. This will maximize the data usage and reserve 25% of the samples for testing.

3.4 Evaluation Metrics

We use the mean squared-error, cosine similarity, top match rate and ranking loss metrics to evaluate our models. These are described by Meng et al. in [18]. We implement top match rate on the absolute numerical values, to investigate absolute changes. We implement ranking loss divided by two, due to a typo in the implementation. To limit the scope of the project, we will not calculate any evaluation metrics for each affective dimension. See [1] and [19] for regression and classification results w.r.t. each affective dimension.

We take the direction of the label to be the most important property, which implies that we identify the best model as the one which maximizes cosine similarity. This is less sensitive to individual subject rating differences than the mean squared-error.

We will evaluate each model w.r.t. its ability to predict labels for (subject, stimuli) pairs, and we will evaluate the best model w.r.t. its ability to predict the population label. We will also carry out two-tailed t-test to identify models which perform significantly better than the Mean Model. The evaluation w.r.t. the population label was also carried out by Koelstra et al. for classification accuracy [1].

We establish two benchmark models. The first is the Mean Model, which predicts a label with the mean over the training set. This is equivalent to linear regression without features from physiological signals. The second model is the Random Model, which takes a standard multivariate Normal random variable as its prediction. See appendix for further details.

4 Experiments

4.1 Feature Extraction

Predicting affective labels from physiological signals has been studied extensively in the literature [11], [6], [7], [14] and [8]. As we are already tackling the less studied problem of implicit affective tagging, we decide to not invent new features from physiological signals. We therefore extract the same features as Koelstra et al. [11]. Their features were picked carefully and capture all immediately relevant information for our problem. Furthermore, by using the same set of features we are able to compare our results to theirs.

For preprocessing, the first three seconds of each physiological signal was taken as baseline and its mean was subtracted from the rest of the signal. We then extracted the following 66 features in Matlab R2013a:

- **GSR**
  - *Basic statistics*: average skin resistance, average of derivative, average of derivative for negative values only, proportion of negative samples in the derivative over all samples
  - *Time domain*: number of local minima in the GSR signal, average rising time of the GSR signal
  - *Frequency domain*: 10 spectral power in the [0-2.4]Hz bands, zero crossing rate of Skin conductance slow response (SCSR) [0-0.2]Hz, zero crossing rate of skin conductance very slow response (SCVSR) [0-0.08]Hz, SCSR and SCVSR mean of peaks magnitude

For the time domain features we applied a moving-average interpolation to reduce the noise of the signal.

The basic statistics and frequency domain features have been noted to work well in many studies [4] [8] [5] [20] [21]. Time domain features, such as peak occurrences, have also been suggested in the literature [8].

- **BVP**
  - *Basic statistics*: Average and standard deviation of HR (Heart-rate),
  - *Time domain*: average and standard deviation of heart rate variability (interbeat interval lengths), heart rate
  - *Frequency domain*: energy ratio between the frequency bands pressure [0.04-0.15]Hz and [0.15-0.5]Hz, spectral power in the bands ([0.1-0.2]Hz, [0.2-0.3]Hz, [0.3-0.4]Hz), low frequency [0.01-0.08]Hz, medium frequency [0.08-0.15]Hz and high frequency [0.15-0.5]Hz components of HRV power spectrum.
- **Respiration**
  - *Basic statistics*: average respiration signal, mean of derivative, standard deviation, range or greatest breath,
  - *Time domain*: breathing rate, average peak to peak time, median peak to peak time
  - *Frequency domain*: band energy ratio (difference between the logarithm of energy between the lower (0.05-0.25Hz) and the higher (0.25-5Hz) bands), breathing rhythm (spectral centroid), 10 spectral power in the bands from 0 to 2.4Hz,
- **Skin temperature**
  - *Basic statistics*: average, average of its derivative,
  - *Frequency domain*: spectral power in the bands ([0-0.1]Hz, [0.1-0.2]Hz)
- **EMG and EOG**
  - *Basic statistics*: energy of the signal, mean and variance
  - *Time domain*: eye blinking rate

We point out that almost all our the features are invariant to perturbations in time. For example, a signal produced by an event at a certain time would yield the same feature values as a signal produced by the same event at a different time. This gives us confidence that the model will generalize to new stimuli.

4.2 Feature Selection

We choose to apply unsupervised learning methods to select a subset of features. This differentiates our approach from Soleymani et al [19].

We first applied Principal Components Analysis (PCA) [22], which we found to be ineffective. We then applied Independent Component Analysis (ICA) [22]. The assumption made by ICA is that independent physiological processes of the human body are responsible for the features. This a strong and bold assumption. Nevertheless, the same assumption has been applied to studying Electroencephalography (EEG) signals and Magnetoencephalography (MEG) images with measurable success [23] [1]. As anticipated, this appeared to work better than PCA. Please see appendix for implementation and a detailed analysis of PCA and ICA.

We normalize all features to have mean zero and standard deviation one across all subjects and stimuli.

At the end of our project, we found that our models based on ICA features did not yield significant results. Therefore we also performed a linear regression analysis on a feature-by-feature basis, to investigate whether a hand-picked set of features would have sufficed. Unfortunately, no such set of features appeared to exist across all subjects. In particular, our analysis yielded evidence in favour of the subject-dependent model. Please see appendix for the detailed analysis.

4.3 Models

We apply Linear Ridge Regression (LRR). If there is a linear relationship between the physiological signals and the labels, LRR should effectively weight the most important parameters. LRR appeared to work for Soleymani et al. on our dataset [19], while Bayesian Ridge Regression was found to work for Koelstra et al. on subjects watching video clips [1].

Linearity is a strong and bold assumption, which certainly does not hold between all our features and labels. We should therefore also consider non-linear models, which work well on small datasets. K-Nearest Neighbours Regression (KNN) is one model, which has been used in several studies [4] [14]. We apply KNN with Euclidean norm, where we only need to learn the parameter for the number of neighbours. If the features clusters well, then KNN should work well on the dataset.

After initial experiments, we found that LRR and KNN did not yield good results. We therefore also applied Kernel Ridge Regression (KRR) with a Gaussian kernel. It is a natural extension to LRR, and it is closely related to the Support Vector Machines for Regression (SVR), which have been applied in the literature previously [14] [4]. We assume the features are normally distributed in the affective dimensions space and choose a Gaussian kernel. Given more time, other kernels should also be applied.

We train a separate model for each affective dimension.

4.4 Fusion

We apply both feature-level fusion (early fusion) and decision-level fusion (late fusion), as both types have shown success in the literature [6] [14].

Each modality responds to an event uniquely. A spike in the signal triggered by a certain event has its own time-lag, duration and amplitude. This could cause problems for feature-level fusion. However, we have chosen all of our features to be almost entirely time-invariant. Furthermore, for GSR, BVP, respiration and skin temperature, we have several features based on spectral power bands, which allows our models to give weight to spikes of certain durations. We therefore do not consider feature-level fusion to be a problem for our models.

For the subject-independent models, we use leave-one-subject-out cross-validation in the training phase. We therefore perform feature-level fusion for all models based on both the extracted features as well as 2, 4 and 6 first ICA components. We also perform decision-level fusion based on the extracted features, as we expect these to contain more information than the ICA components. For each modality we apply KRR to capture the non-linear relationships between features and labels, and then we apply Least-Squares Linear Regression (LR) on the predicted labels to *weight* the predictions for each modality linearly.

For the subject-dependent models, we use leave-one-stimuli-out cross-validation in the training phase. Since we only have 29 samples available, we choose to only perform feature-level fusion for all models based on the first 2, 4 and 6 ICA components.

4.5 Results

The results for the subject-independent models are given in Table 2. We observe that all the models perform significantly better than the Mean Model w.r.t. ranking loss, but not w.r.t. any other evaluation metrics. We also note that several models appear to be better than both the Mean Model and Random Model w.r.t. cosine similarity and mean squared-error. In particular, LRR appear to perform consistently better than the KRR and KNN models. We speculate that KNN does not perform well because the features do not cluster tightly w.r.t. the labels. Since KRR with a Gaussian kernel is considered to be a broader and more powerful model than LRR, this indicates that KRR is overfitting the data [22]. Indeed, this is confirmed from the validation errors in appendix Figure 19, 20 and 21. Interestingly, the decision-level fusion model appears to perform almost as well as the best feature-level fusion model. This indicates that dimensionality reduction and KRR on each modality capture the same amount of information. We conclude that the subject-independent models appear to capture some of the underlying relationship between physiological signals and affective labels.

The results for the subject-dependent models are given in Table 3. It is evident that the Mean Model outperforms all of the other models, and we must therefore conclude that our models have not captured any significant information relating physiological signals to affective labels. This confirms previous results in the literature for subject-dependent models on our the data set [19] for subject-dependent models. We speculate that this is due to the small sample size.

Finally, we investigate the ability of our best subject-independent model, LRR ICA4, to predict predict the population label for a new stimuli. The results are shown in Figure 5. For comparison we have plotted the optimal result in Figure 6, where each (subject, stimuli) prediction is replaced by the true label. This is the performance we would have obtained if our model was perfect. The LRR ICA4 model yields high errors and does not improve with the number of subjects, which shows that the model is inadequate for predicting the population label. We speculate that the poor performance is due to an additional noise factor. The model first predicts the affective label for each subject, which incurs a certain noise. Then the mean is taken over these labels to predict the population label, which incurs an additional amount of noise.

Model	Mean squared-error	Cosine similarity	Ranking loss	Top match rate
KNN ICA2	5.48717 ± 1.257	0.04049 ± 0.126	*0.244271 ± 0.025	0.209375 ± 0.105
KNN ICA4	6.30159 ± 1.225	-0.00212 ± 0.082	*0.241667 ± 0.0135	0.228125 ± 0.047
KNN ICA6	5.41899 ± 1.182	0.01662 ± 0.085	*0.233333 ± 0.036	0.209375 ± 0.09
KNN (All modalities)	5.40466 ± 1.066	0.11410 ± 0.211	*0.231771 ± 0.041	0.240625 ± 0.087
KRR ICA2	5.73113 ± 1.369	0.03817 ± 0.086	*0.247135 ± 0.03	0.212500 ± 0.065
KRR ICA4	6.30280 ± 1.283	0.05186 ± 0.12	*0.245313 ± 0.034	0.209375 ± 0.061
KRR ICA6	6.10283 ± 1.443	0.02131 ± 0.057	*0.235417 ± 0.017	0.250000 ± 0.056
KRR (All modalities)	23.9683 ± 46.012	0.05016 ± 0.194	*0.240885 ± 0.040	0.221875 ± 0.146
LRR ICA2	4.97935 ± 1.216	0.16873 ± 0.25	*0.230208 ± 0.057	0.290625 ± 0.238
LRR ICA4	**4.96593 ± 1.19**	**0.17845 ± 0.274**	*0.229167 ± 0.0586	0.290625 ± 0.238
LRR ICA6	4.98389 ± 1.21	0.16583 ± 0.266	*0.227865 ± 0.063	0.290625 ± 0.238
LRR (All modalities)	4.99066 ± 1.233	0.13848 ± 0.222	***0.226302 ± 0.0628**	0.253125 ± 0.201
KRR+LR (Decision-level fusion)	5.37779 ± 1.037	0.15612 ± 0.214	*0.222656 ± 0.051	0.278125 ± 0.183
Mean Model	5.08870 ± 1.272	-0.30864 ± 0.239	0.3244790 ± 0.051	**0.396880 ± 0.208**
Random Model, N(0,1)	8.00000 ± 5.67	0.00000 ± 0.5	0.2500000 ± 0.1225	0.250000 ± 0.435

TABLE 2: Subject-independent models: mean evaluation metric values over each subject and corresponding standard deviations. The best model for each evaluation metric is indicated by bold font. A two-sided t-test was calculated for all models w.r.t. the Mean Model. Asterisk (*) indicates that the test result is significant at 95% confidence level. The reader should observe that our ranking loss metric is divided by two as compared to the definition given by Meng et al. in [18].

Model	Mean squared-error	Cosine similarity	Ranking loss	Top match rate
KNN ICA2	5.41092 ± 1.25	0.009380 ± 0.113	0.237240 ± 0.023	0.259375 ± 0.057
KNN ICA4	5.34627 ± 1.226	0.027682 ± 0.088	0.227865 ± 0.018	0.268750 ± 0.064
KNN ICA6	5.19881 ± 1.096	0.072793 ± 0.097	0.228906 ± 0.024	0.268750 ± 0.099
KRR ICA2	6.01722 ± 1.236	0.122084 ± 0.122	0.235937 ± 0.023	0.309375 ± 0.077
KRR ICA4	5.67165 ± 1.091	0.143894 ± 0.073	0.221354 ± 0.013	0.262500 ± 0.062
KRR ICA6	5.41496 ± 1.129	0.199371 ± 0.086	0.209635 ± 0.018	0.278125 ± 0.066
LRR ICA2	4.98369 ± 1.168	0.211393 ± 0.074	0.196354 ± 0.022	0.284375 ± 0.075
LRR ICA4	5.01024 ± 1.174	**0.220811 ± 0.084**	0.200781 ± 0.021	0.278125 ± 0.049
LRR ICA6	5.01521 ± 1.15	0.219693 ± 0.075	0.201042 ± 0.016	0.281250 ± 0.065
Mean Model	**4.94490 ± 1.227**	0.207494 ± 0.077	**0.194271 ± 0.018**	**0.284375 ± 0.056**
Random Model, N(0,1)	8.00000 ± 5.67	0.000000 ± 0.5	0.250000 ± 0.1225	0.250000 ± 0.435

TABLE 3: Subject-dependent models: mean evaluation metric values over each stimuli and corresponding standard deviations. The best model for each evaluation metric is indicated by bold font. No models performed significantly better than the Mean Model. The reader should observe that our ranking loss metric is divided by two as compared to the definition given by Meng et al. in [18].

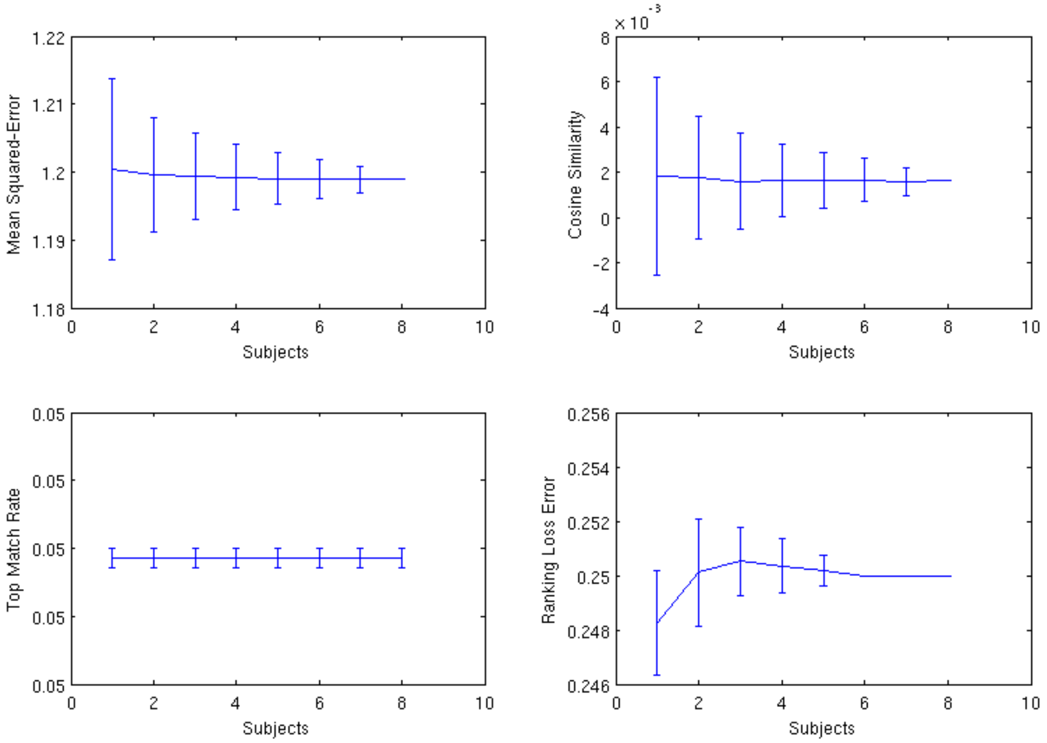


Fig. 5: Evaluation metrics w.r.t. true population label for the subject-independent LRR ICA4 model. The labels are only tested w.r.t. the 8 test subjects, which were previously designated to be the test subjects.

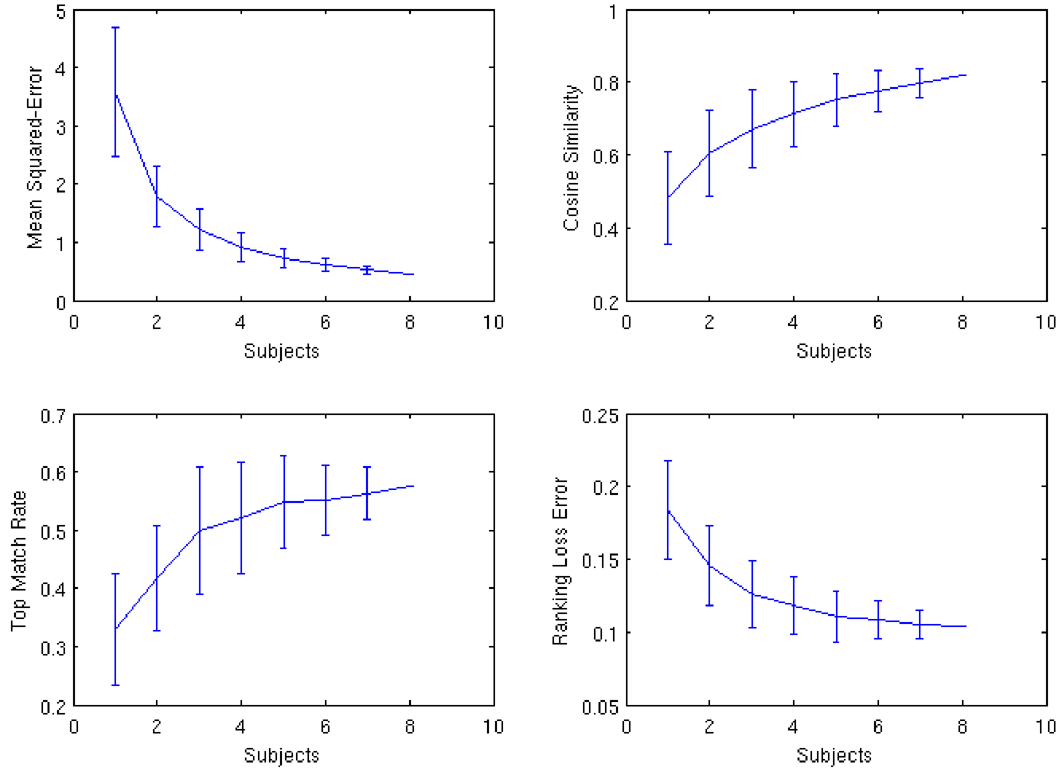


Fig. 6: Evaluation metrics w.r.t. true population label for a model which predicts each (subject, stimuli) pair with its true affective labels. This is a perfect model and the best we can hope to achieve in regard to predicting the population label based on a mean of affective subject labels.

5 Further Work

The features we used were largely taken from previous work on predicting affective response to pictures and videos. However, we may speculate that music videos elicit different physiological signals from these. Therefore further work should be directed at constructing new features specific to music videos.

Adding the modalities video and EEG signals may also improve our system performance significantly [6], [14], [24], [1]. If the user was allowed to move around, we may also take body movement into account, as it can be can be used effectively predict affective state [25]. In particular, it would be interesting to identify laughter from body movement and use it as input to our models [26].

Fisher Projection and Recursive Feature Elimination appear to be promising dimensionality reduction methods [17] [1].

Our limited success with subject-independent models suggest that we should apply Multi-Task Learning to learn the characteristics of each subject [16] [18]. We may also consider gender-dependent models, as males and females have been observed to give markedly different affective responses to video clips [20].

6 Conclusion

Implicit affective tagging is an automatic method for inferring affective tags for media based on analysing a person experiencing the media. In this project we attempted to construct an implicit affective tagging system with the end goal of tagging media content based on peripheral physiological signals from users. We used dataset of 32 subjects watching 1. min music video clips. We took the affective labels self-reported by the subjects to be the ground truth w.r.t. each subject, and their mean as the ground truth w.r.t. the population. We extracted a large set of features, and applied KNN, LRR and KRR to construct subject-dependent and subject-independent models for predicting the affective labels. Our subject-independent models yielded some significant results, but failed to predict the population label. Further work is needed in the direction of constructing new features and fusioning aditional modalities.

Appendix A
Database

We use the database DEAP as described by Koelstra et al. in [11]/, as described briefly in Section 2. The database was based on 32 participants, who each watched a 1 min. music video clip from a total of 40 music videos. Participants were from The Netherlands and from Switzerland, and ranged from 19 to 37 years of age. Participants were fitted with biosensors and their physiological signals occurring during each music video clip were recorded. The biosensors used were Galvanic Skin Response (GSR), blood volume pressure (BVP), respiration, skin temperature, Electromyography (EMG), Electrooculography (EOG) and Electroencephalography (EEG). The BioSemi ActiveTwo system was used, see http://www.biosemi.com/. All biosensors recorded at 512Hz sampling rate, which was afterwards down-sampled to 128 Hz. Frontal face video was also recorded.

After watching each clip, participants self-reported on the continuous affective dimensions of valence, arousal, dominance and liking on a scale 1-9. Participants used Self Assessment Manikins (SAMs), which are known to work better across different cultures [12], see Figure 3. Participants also recorded their on a scale 1-5.

Based on a literature review, Gunes and Pantic observe that modelling continuous affective dimensions may be quite difficult [6]. They note that valence is not universally understood and that the labelling process may therefore be corrupted. In this regard, Koelstra et al. performed a statistical correlation analysis between the affective dimensions to validate that the participants understood the manikins. Their analysis showed that the participants indeed were able to differentiate between the affective dimensions.

The music videos were chosen according to a previous experiment, such that they would be distributed equally in the four quadrants of the valence-arousal space and induce the maximum affective response. We should therefore assume that any positive results we obtain on this database may be over-confident and that further research should be conducted to verify that the results also hold for music videos with lower affective response. Between music videos participants only had a very short break, so it is very possible that the physiological responses to one video affects the next video. However, since the order of the music videos were chosen randomly for each participant, we will assume that the this overlapping effect is distributed uniformly across samples and should not induce any modelling bias.

Please see Koelstra et al. in [11] for further details on the database.

We limit ourselves to the peripheral physiological signals, as these are known to carry a significant amount of information. Bradley et al. have carried out several studies on pleasant and unpleasant images showing that muscle activity recorded from EOG and EMG is highly correlated with valence [7]. Heart rate acceleration and deceleration is also correlated with valence, while GSR is highly correlated with arousal [7] [4] [8]. There is evidence that physiological signals are also correlated with dominance, see Kim et al. [8].

APPENDIX B
BENCHMARK MODELS

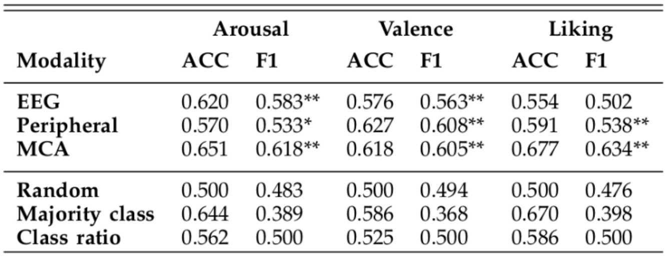
We establish two benchmark models, which we will compare our models against:

Mean Model: For the subject-independent model, the Mean Model predicts a label for a new (subject, stimuli) pair as the mean of all previously observed (subject, stimuli) pairs. For our subject-dependent models, the Mean Model predicts a new stimuli as the mean of all the other labels from the same subject. It is a naive benchmark, but in our setting it is equivalent to training a Linear Regression model without any physiological features. This is the best model we could obtain if we did not observe the physiological features. Any model which performs better than the Mean Model must be capturing some of the underlying relationship between physiological features and labels.

Random Model: The Random Model predicts any (subject, stimuli) pair with a draw from a standard Normal distribution $x \sim N(0, 1)$. This is the best prediction we could give, if all we knew was that the data was distributed according to the standard Normal distribution. The model was evaluated based on $10,000$ simulations in Matlab 2013a.

APPENDIX C
COMPARABLE RESULTS IN THE LITERATURE

Koelstra et al. have already performed some analysis on our dataset [11]. They used Fishers linear discriminant for feature selection and a Gaussian naive Bayes classifier for classification w.r.t. self-reported valence, arousal and liking labels. They extracted the same features as us. They performed feature-level fusion by concatenating all the features together. They performed decision-level fusion by applying the Gaussian naive Bayes classifier to each modality, and then applying the classifier again to the results from each modality. Their results are given in Figure 7. Their models do not yield considerably better results than predicting with the majority class, which indicates that our problem is quite difficult.

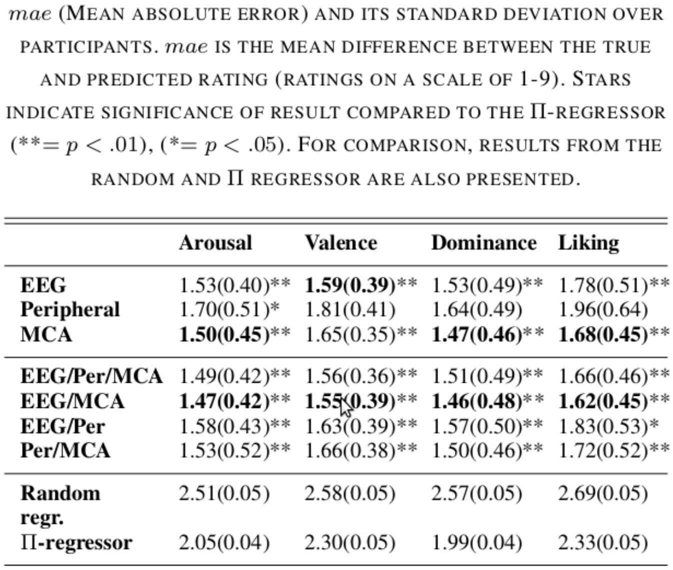


	Arousal		Valence		Liking	
Modality	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**
EEG	0.620	0.583**	0.576	0.563**	0.554	0.502
Peripheral	0.570	0.533*	0.627	0.608**	0.591	0.538**
MCA	0.651	0.618**	0.618	0.605**	0.677	0.634**
Random	0.500	0.483	0.500	0.494	0.500	0.476
Majority class	0.644	0.389	0.586	0.368	0.670	0.398
Class ratio	0.562	0.500	0.525	0.500	0.586	0.500

Fig. 7: Taken from Table 7. in [11].

Our dataset was also used by Soleymani et al. [19]. They extracted the same features as us from the peripheral physiological signals, and applied LRR. They trained subject-dependent regression models, which relates their work closely to ours. Hpwever, contrary to our approach Soleymani et al. did not perform any dimensionality reduction. They also applied Gaussian Process Regression and Relevance Vector Machine regression, but these did not yield any improvement compared to LRR. For the peripheral physiological signals, their results are only significant for the arousal modality, which also indicates that our problem is quite difficult. See Figure 8.

Implicit affective tagging is also investigated by Koelstra et al. in [1] for video clips. They extract a set of features from EEG signals and frontal video of the subject seeing video clips. Then then use Bayesian Ridge Regression, a variant of LRR, to predict affective response w.r.t. valence, arousal and dominance. For feature-level fusion BRR is applied to features from both modalities. For decision-level fusion, BRR is applied to each modality, and then applied again to the outputs from each modality. Their models are significantly better than predicting with the Mean Model w.r.t. arousal and dominance.



mae (Mean absolute error) and its standard deviation over participants. mae is the mean difference between the true and predicted rating (ratings on a scale of 1-9). Stars indicate significance of result compared to the Π-regressor (**$= p < .01$), (*$= p < .05$). For comparison, results from the random and Π regressor are also presented.

	Arousal	**Valence**	**Dominance**	**Liking**
EEG	1.53(0.40)**	**1.59(0.39)****	1.53(0.49)**	1.78(0.51)**
Peripheral	1.70(0.51)*	1.81(0.41)	1.64(0.49)	1.96(0.64)
MCA	**1.50(0.45)****	1.65(0.35)**	**1.47(0.46)****	**1.68(0.45)****
EEG/Per/MCA	1.49(0.42)**	1.56(0.36)**	1.51(0.49)**	1.66(0.46)**
EEG/MCA	**1.47(0.42)****	**1.55(0.39)****	**1.46(0.48)****	**1.62(0.45)****
EEG/Per	1.58(0.43)**	1.63(0.39)**	1.57(0.50)**	1.83(0.53)*
Per/MCA	1.53(0.52)**	1.66(0.38)**	1.50(0.46)**	1.72(0.52)**
Random regr.	2.51(0.05)	2.58(0.05)	2.57(0.05)	2.69(0.05)
Π-regressor	2.05(0.04)	2.30(0.05)	1.99(0.04)	2.33(0.05)

Fig. 8: Taken from Table II in [19].

Taken together, these studies indicate that other modalities, such as frontal video or EEG signals, may be required to obtain effective implicit tags.

APPENDIX D
PRINCIPAL COMPONENTS ANALYSIS

Principle Component Analysis (PCA) is an effective method to extract a smaller set of features as a linear combination of the original features [22]. PCA extracts the K orthogonal vectors which span the directions of maximum variation in the data. By only taking these directions into account, we hope to remove only a minimum amount of information from the physiological signals.

We used the built-in Matlab function *princomp* and applied PCA to our data set. From this, we observed that the first three principle components were not effective at separating samples w.r.t. valence and arousal. The samples seemed to be more separable for a single subject, but clearly we would still need a highly non-linear function to separate them.

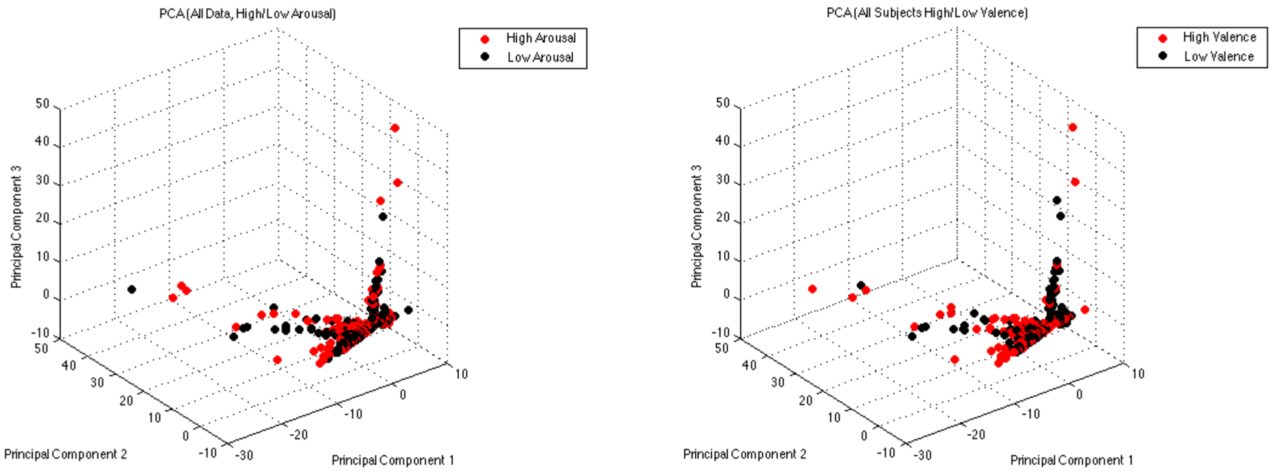
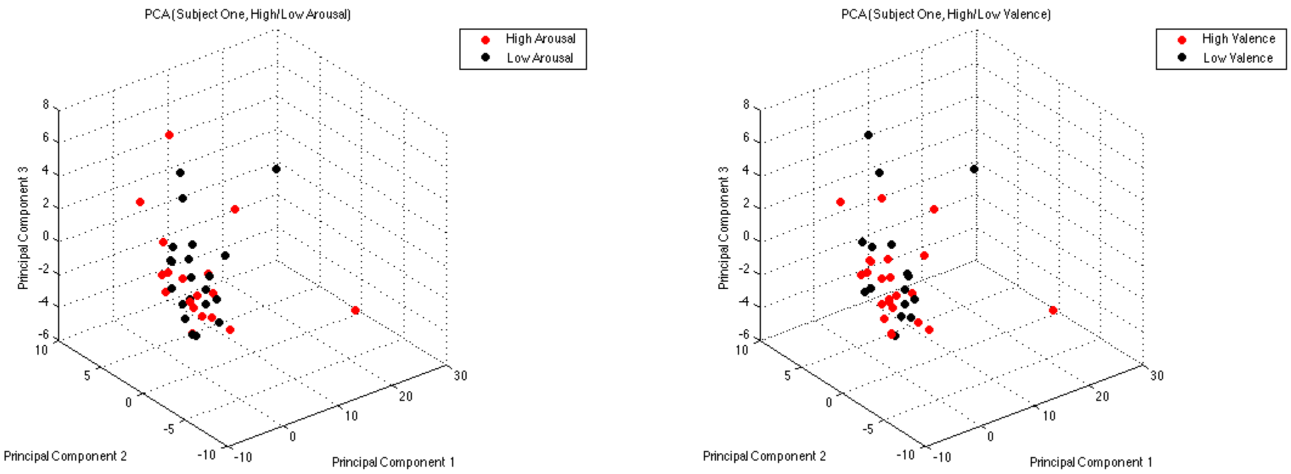


Fig. 9: First three PCA components of all features across all subjects labelled by high / low valence and arousal respectively. Scores greater than 5 are defined as *high* and lower as *low*. The data appears to be highly inseparable.



* *

Fig. 10: First three PCA components of all features for the first subject (subject one) labelled by high / low valence and arousal. Scores greater than 5 are defined as *high* and lower as *low*. The data appears to be inseparable.

APPENDIX E
INDEPENDENT COMPONENTS ANALYSIS

We applied Independent Component Analysis (ICA) [22] to our extracted features. We based our implementation on that by Brian Moore (brimoorumich.edu) at http://www.mathworks.com/matlabcentral/fileexchange/authors/277668, which centers and whitens the data, and then applies ICA assuming the data is multivariate Normally distributed.

To the naked eye, it was impossible to assess whether first three ICA components were able to separate the samples better than the first three PCA components. To evaluate whether or not ICA would be a good dimensionality reduction method, we instead compared its coefficients to those of PCA. From these it appeared that ICA was more selective regarding the features. In its first component ICA put the majority of its weight on $6-8$ features, while PCA put its weights uniformly across a much larger set of features. Due to time constraints, we therefore chose to continue with ICA.

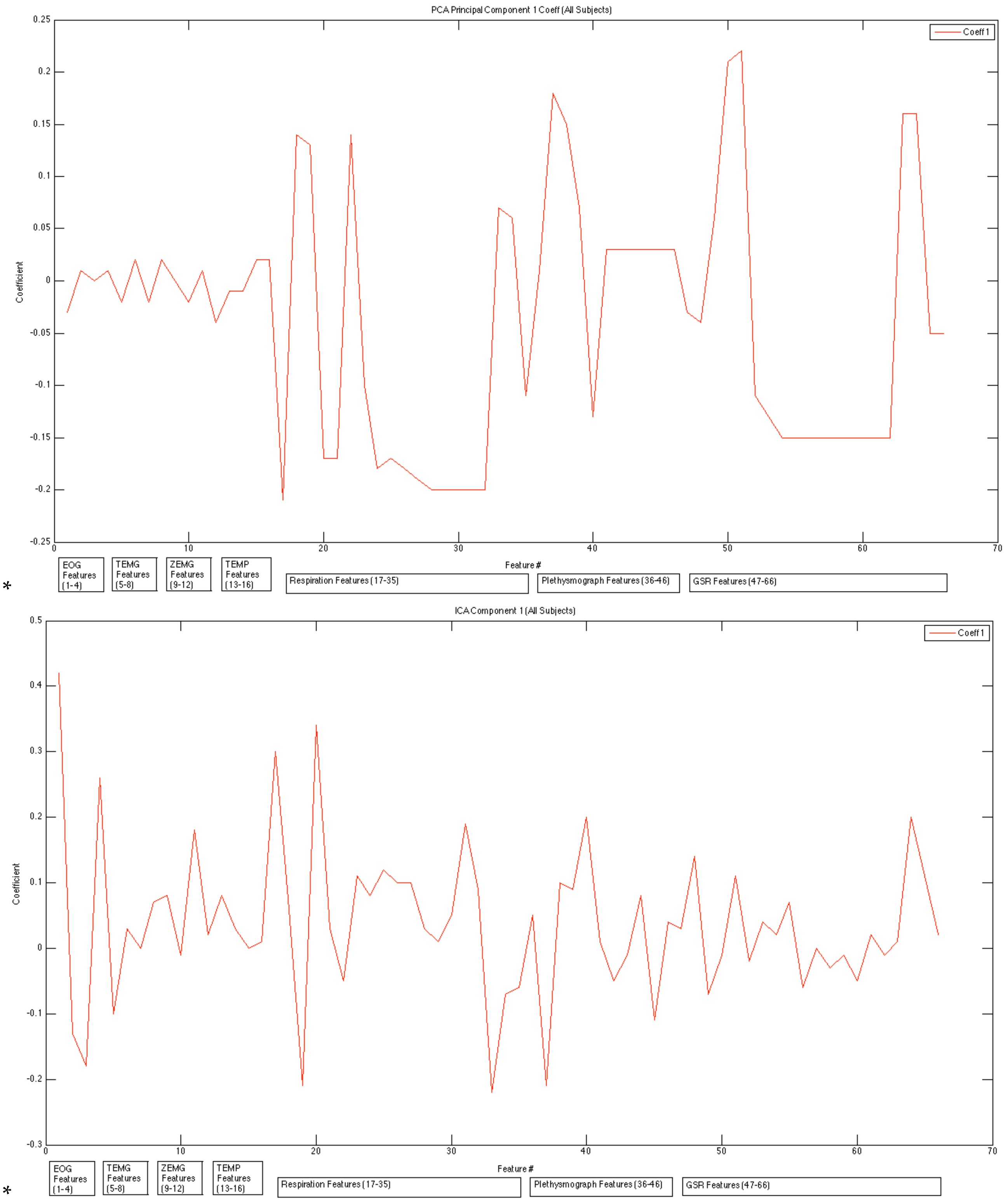


Fig. 11: Top) Weights for the first PCA component. Bottom) Weights for the first ICA component.

Appendix F
Linear Regression Analysis

We performed a linear regression analysis to evaluate whether or not a small subset of features could be used. For each affective dimension and each feature, we fitted a linear regression with intercept and estimated the mean parameter value of the feature and its 95% confidence interval. We used the built-in Matlab function *glmfit*. This was done for all subjects and stimuli, and for subject one and all stimuli. The mean value of the parameter for each feature, together with its confidence interval, should give us a good indication of how effective it is at predicting the affective score.

As expected from the literature, we observed that GSR and PLTH features were highly correlated with arousal, while EOG, ZEMG and BVP features were correlated with valence across all subjects. Interestingly, we also observed that EOG and respiration were correlated with dominance and that GSR was the main modality correlated with liking. Furthermore, we also observed that the correlations were often unique to each subject. For example, Figure 12 show the linear regression analysis across all subjects and for only subject one w.r.t. arousal. From this we observe that BVP features correlate very differently for subject one than for all subjects. These, and other plots, supports our hypothesis of a model for each participant.

The figures below show the linear regression analysis for all affective dimensions across all subjects and for subject one only. The small box to the left shows the five features with highest absolute value.

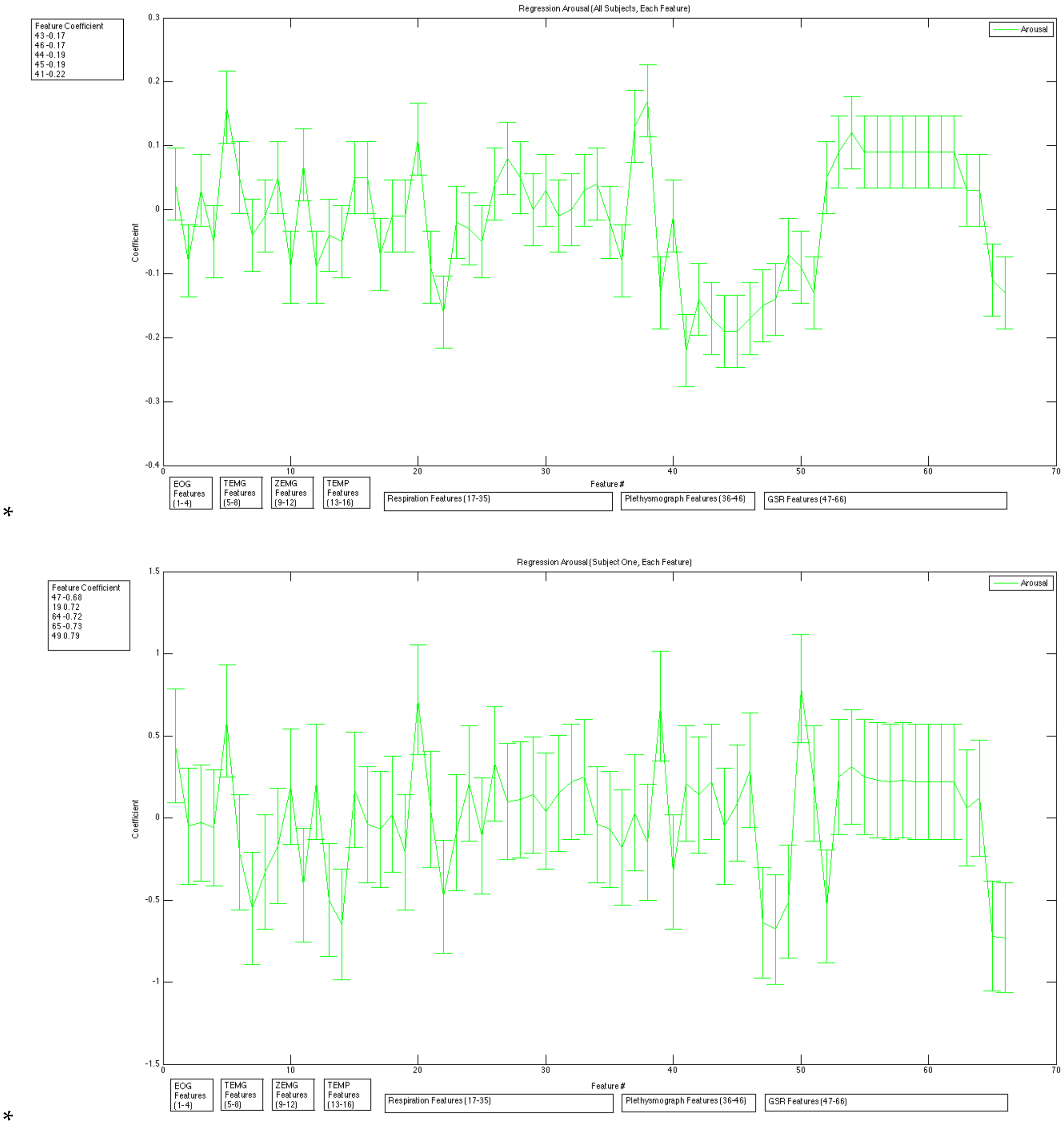
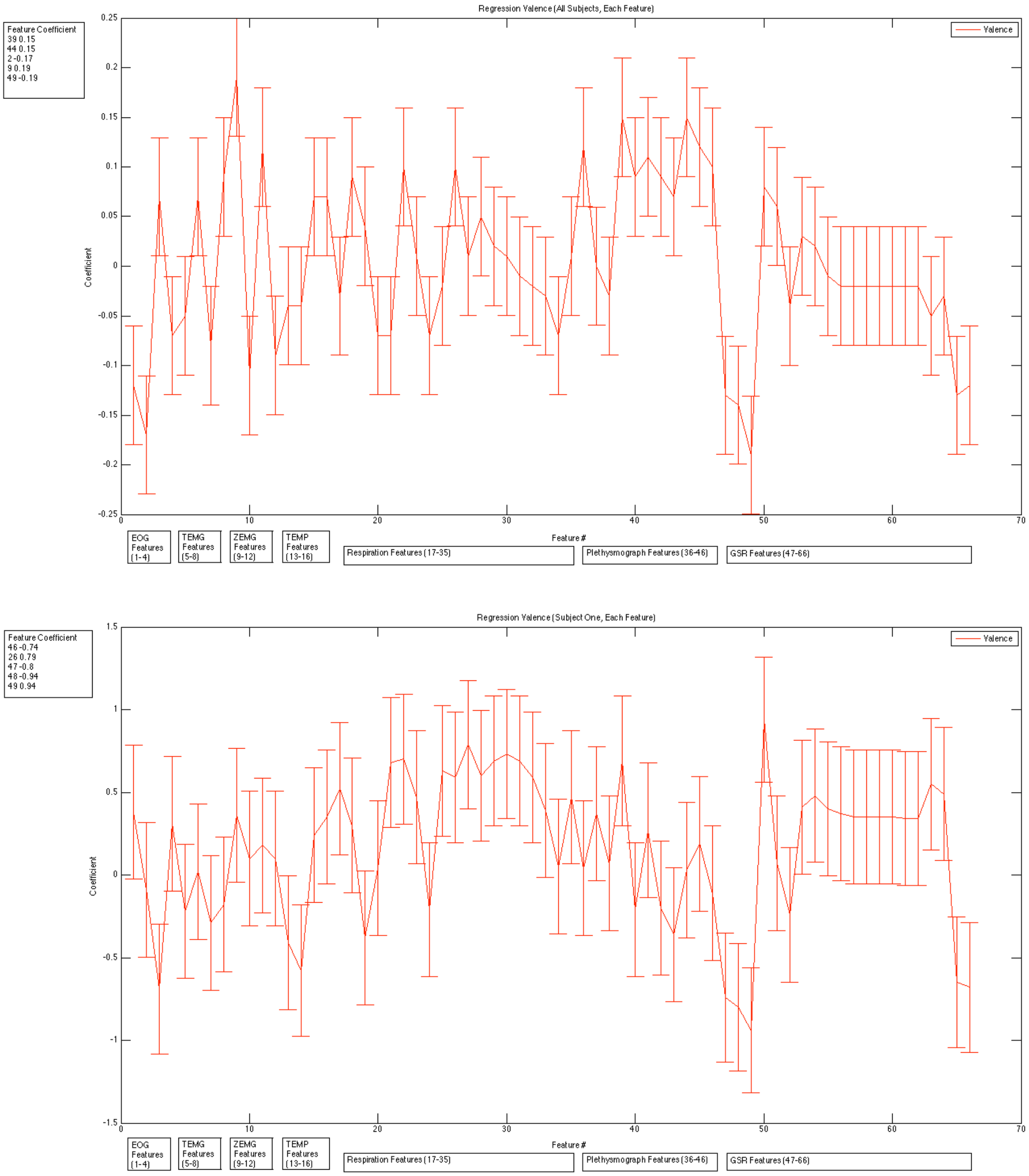


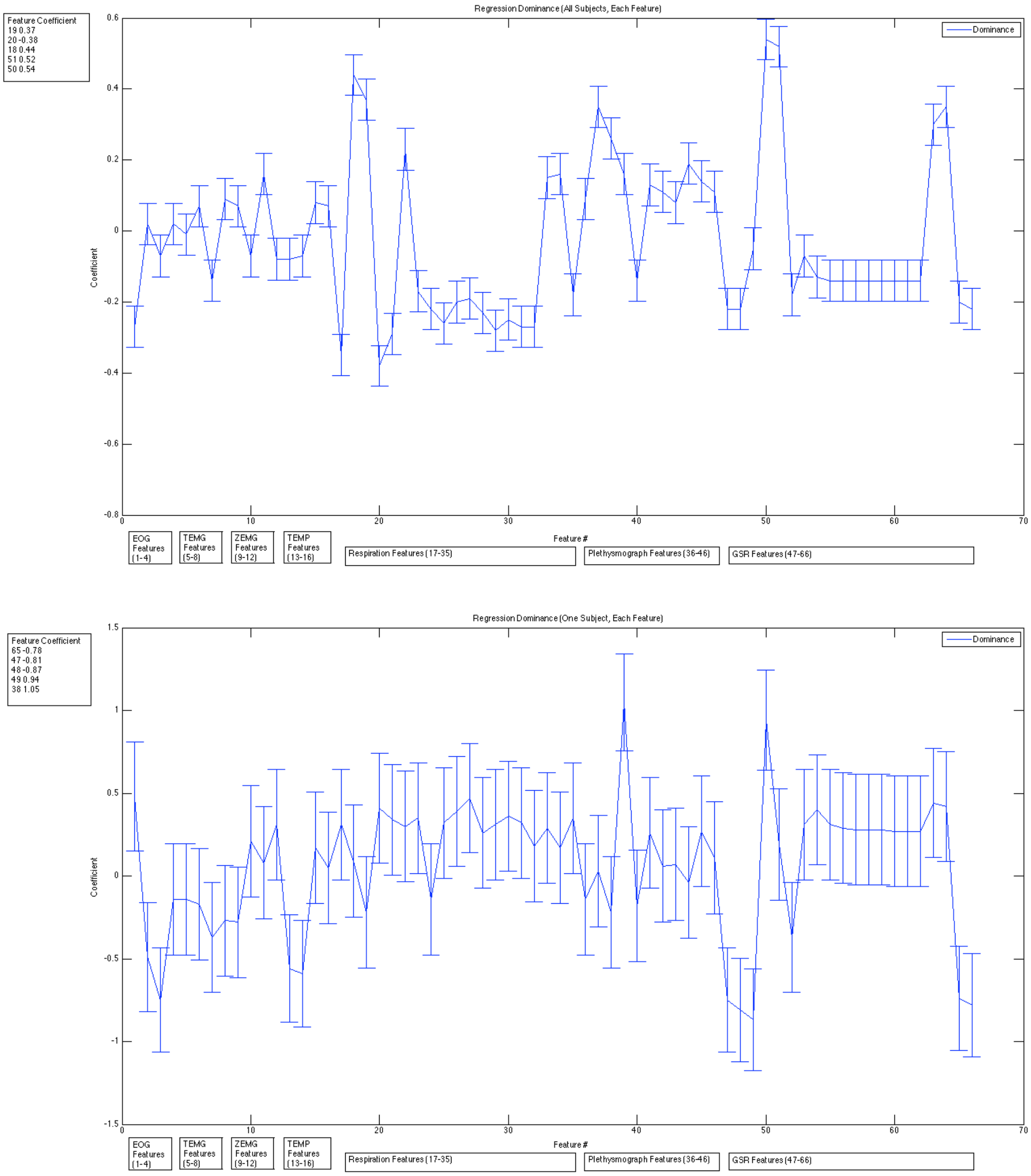
Fig. 12: Top) Linear regression analysis for each feature, with intercept, across all subjects and stimuli w.r.t. arousal. Errorbars indicate 95% confidence intervals for the parameter value. Bottom) Linear regression analysis for each feature, with intercept, for first subject and all stimuli. Errorbars indicate 95% confidence intervals for the parameter value.



*

*

Fig. 13: Top) Linear regression analysis for each feature, with intercept, across all subjects and stimuli w.r.t. valence. Errorbars indicate 95% confidence intervals for the parameter value. Bottom) Linear regression analysis for each feature, with intercept, for first subject and all stimuli. Errorbars indicate 95% confidence intervals for the parameter value.



*

*

Fig. 14: Top) Linear regression analysis for each feature, with intercept, across all subjects and stimuli w.r.t. dominance. Errorbars indicate 95% confidence intervals for the parameter value. Bottom) Linear regression analysis for each feature, with intercept, for first subject and all stimuli. Errorbars indicate 95% confidence intervals for the parameter value.

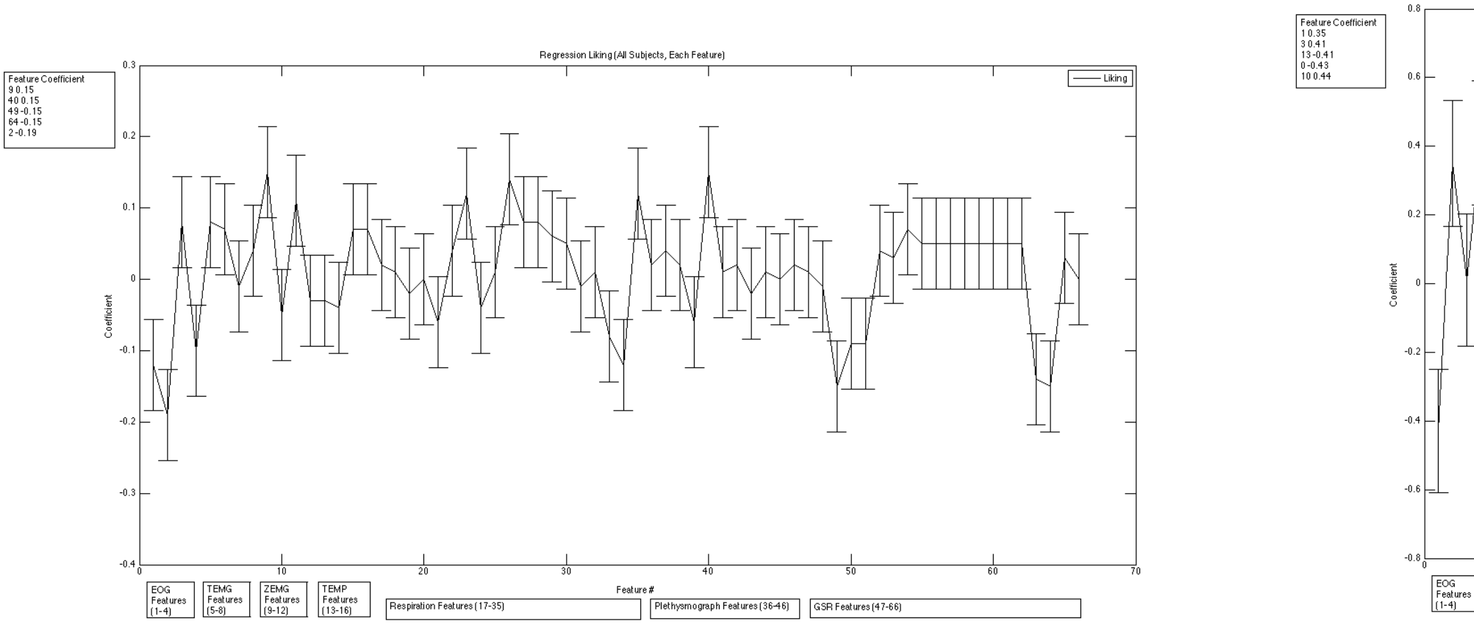


Fig. 15: Top) Linear regression analysis for each feature, with intercept, across all subjects and stimuli w.r.t. liking. Errorbars indicate 95% confidence intervals for the parameter value. Bottom) Linear regression analysis for each feature, with intercept, for first subject and all stimuli. Errorbars indicate 95% confidence intervals for the parameter value.

APPENDIX G
MODEL IMPLEMENTATIONS

We implemented the models, cross-validation and testing procedures in Matlab R2013a. The model implementations were based on former lecture notes, but are equivalent to those given by Hastie et al. in [22]. We also used the Matlab built-in KNN search method *knnsearch*.

Appendix H
Subject-Independent Models

The figures below show the validation error for the subject-independent models based on the ICA features. We observe a significant amount of overfitting.

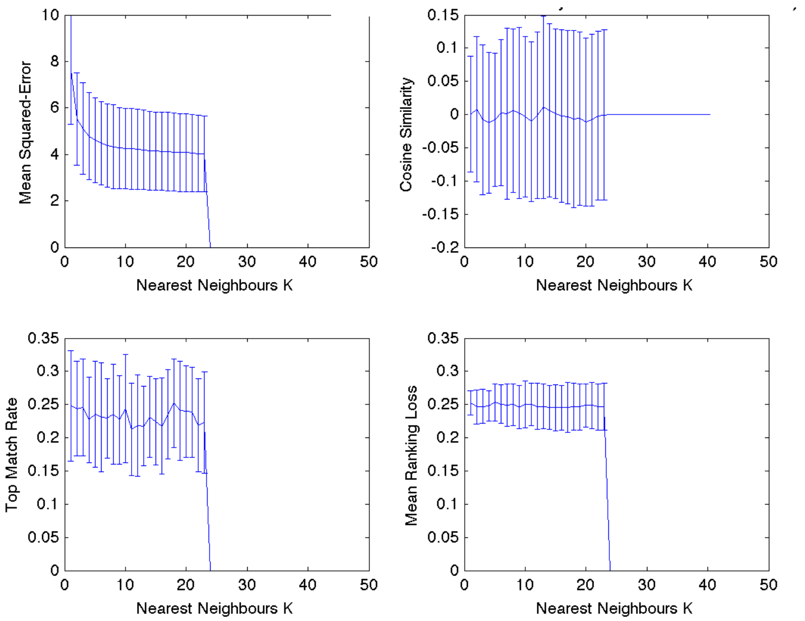


Fig. 16: KNN ICA2 validation error as a function of K, number of neighbours.

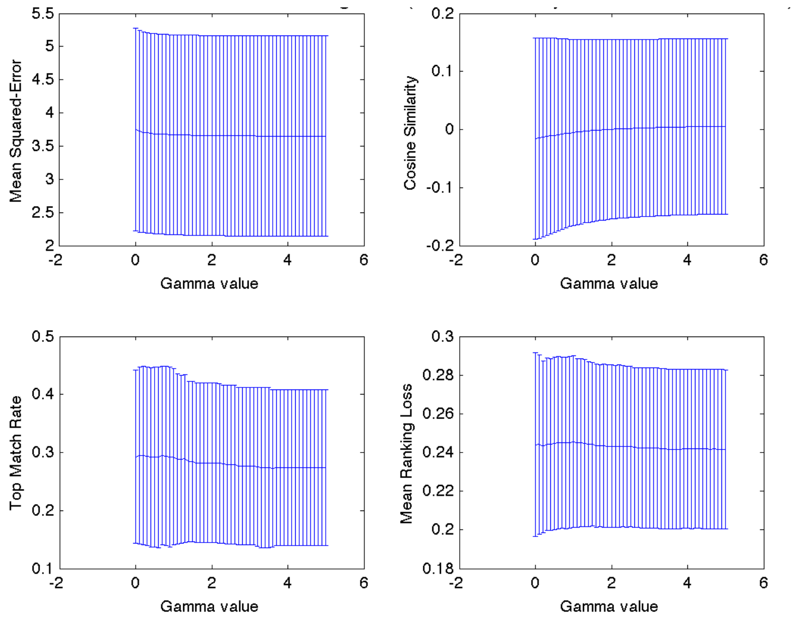


Fig. 17: LRR ICA2 validation error as a function of $\gamma \in \{0, 0.1, \ldots, 5.0\}$, the regularization parameter. Higher γ values imply more regularization and therefore lower parameters weights.

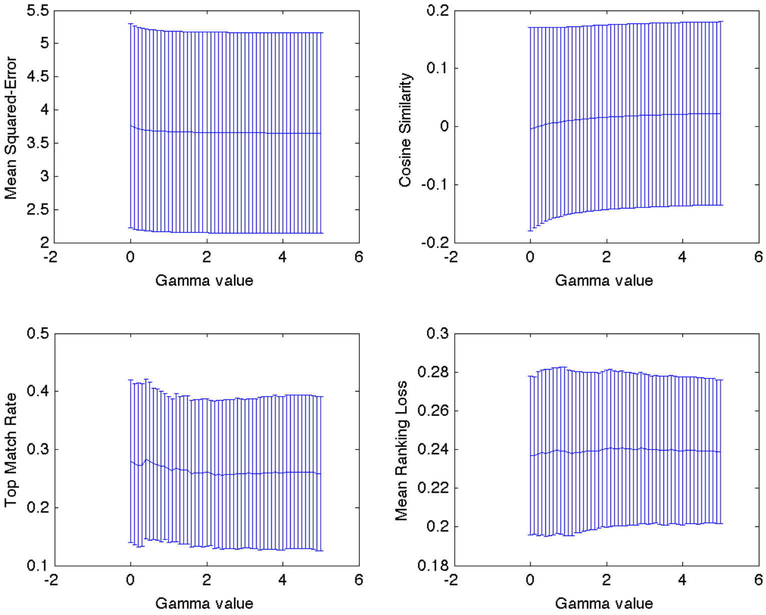


Fig. 18: LRR ICA4 validation error as a function of $\gamma \in \{0, 0.1, \dots, 5.0\}$, the regularization parameter. Higher γ values imply more regularization and therefore lower parameters weights. This was the best subject-independent model.

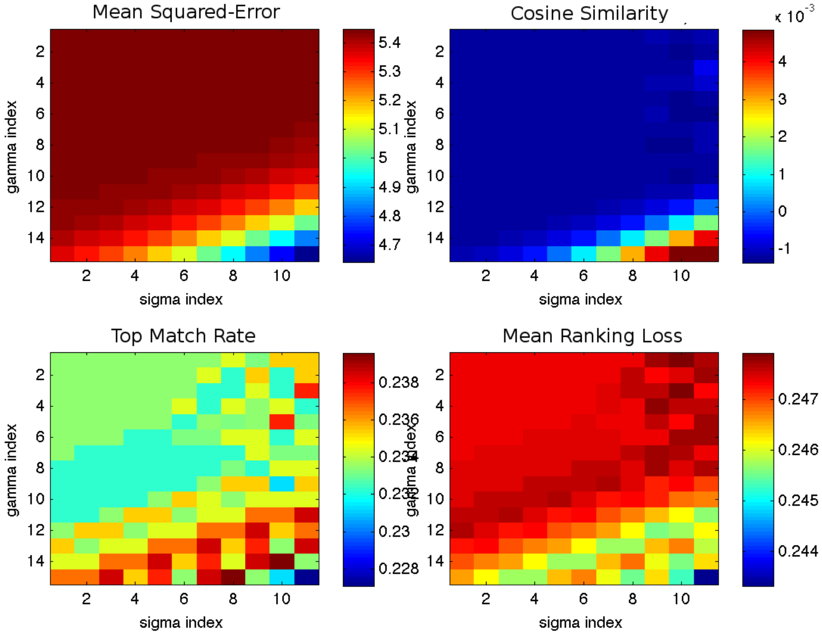


Fig. 19: KRR ICA2 validation error as a function of $\gamma \in \{2^{-40}, 2^{-39}, \dots, 2^{-26}\}$, the regularization parameter, and $\sigma \in \{2^{7}, 2^{7.5}, \dots, 2^{12}\}$. Higher γ values imply more regularization and therefore lower parameter weights, where as higher σ values imply that more training samples are used to predict the label of a new sample.

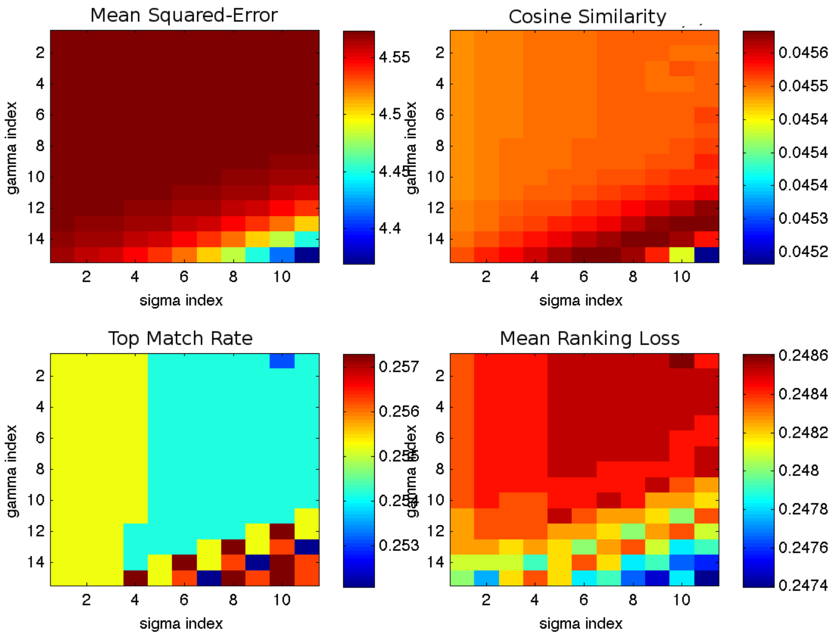


Fig. 20: KRR ICA4 validation error as a function of $\gamma \in \{2^{-40}, 2^{-39}, \dots, 2^{-26}\}$, the regularization parameter, and $\sigma \in \{2^{7}, 2^{7.5}, \dots, 2^{12}\}$. Higher γ values imply more regularization and therefore lower parameter weights, where as higher σ values imply that more training samples are used to predict the label of a new sample.

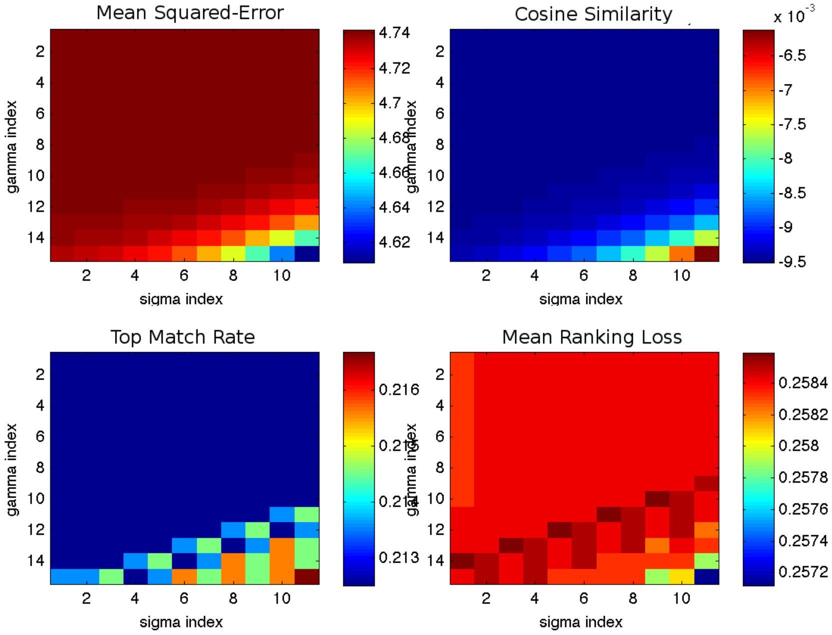


Fig. 21: KRR ICA6 validation error as a function of $\gamma \in \{2^{-40}, 2^{-39}, \dots, 2^{-26}\}$, the regularization parameter, and $\sigma \in \{2^{7}, 2^{7.5}, \dots, 2^{12}\}$. Higher γ values imply more regularization and therefore lower parameter weights, where as higher σ values imply that more training samples are used to predict the label of a new sample.

Appendix I
Subject-dependent Models

The figures below show validation error for the subject-dependent models based on the ICA features. We again observe a significant amount of overfitting.

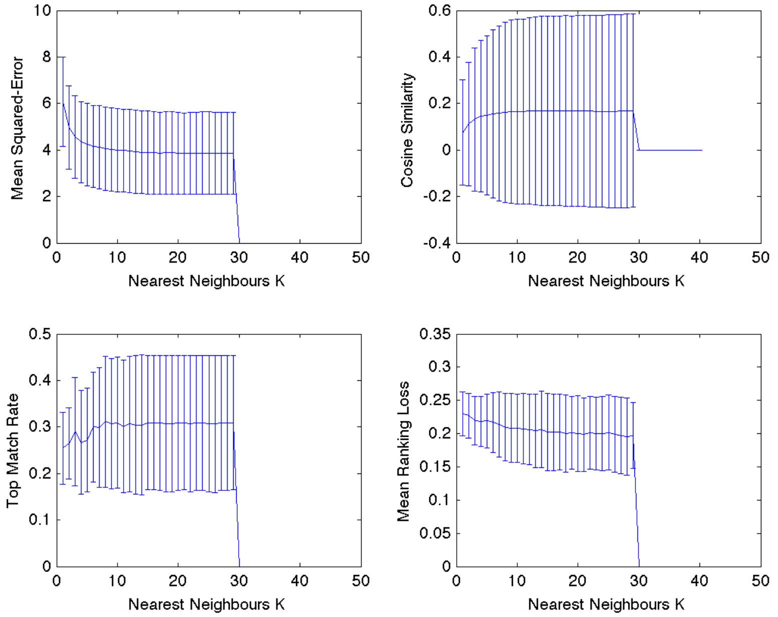


Fig. 22: KNN ICA2 validation error as a function of K, number of neighbours.

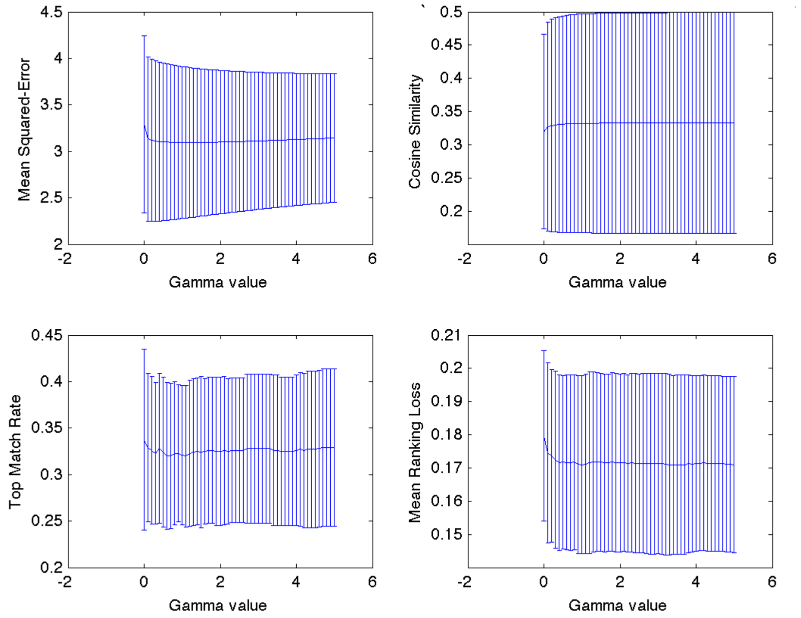


Fig. 23: LRR ICA2 validation error as a function of $\gamma \in \{0, 0.1, \ldots, 5.0\}$, the regularization parameter. Higher γ values imply more regularization and therefore lower parameters weights.

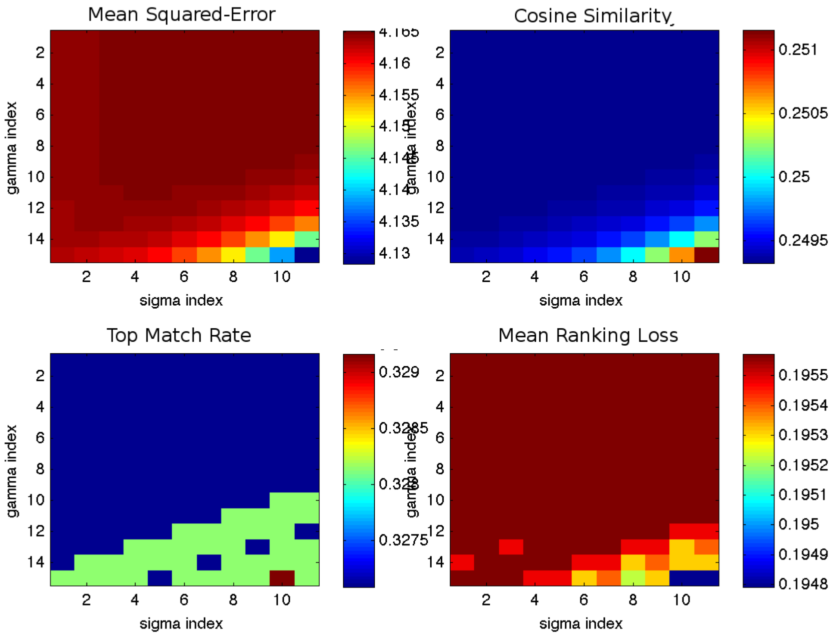


Fig. 24: KRR ICA2 validation error as a function of $\gamma \in \{2^{-40}, 2^{-39}, \ldots, 2^{-26}\}$, the regularization parameter, and $\sigma \in \{2^{7}, 2^{7.5}, \ldots, 2^{12}\}$. Higher γ values imply more regularization and therefore lower parameter weights, where as higher σ values imply that more training samples are used to predict the label of a new sample.

Acknowledgments

The author would like to thank Aaron Perlmutter and Florin Brasoveanu for good team work, as well as Dr. Nadia Bianchi-Berthouze and Dr. Hane Aung for helpful discussions and feedback.

References

[1] S. Koelstra and I. Patras, "Fusion of facial expressions and eeg for implicit affective tagging," *Image and Vision Computing*, vol. 31, no. 2, pp. 164–174, 2013.

[2] B. Reeves and C. Nass, *How people treat computers, television, and new media like real people and places*. CSLI Publications and Cambridge university press, 1996.

[3] M. Pantic and A. Vinciarelli, "Implicit human-centered tagging [social sciences]," *Signal Processing Magazine, IEEE*, vol. 26, no. 6, pp. 173–180, 2009.

[4] S. Wioleta, "Using physiological signals for emotion recognition," in *Human System Interaction (HSI), 2013 The 6th International Conference on*. IEEE, 2013, pp. 556–561.

[5] L. Li and J.-h. Chen, "Emotion recognition using physiological signals," in *Advances in Artificial Reality and Tele-Existence*. Springer, 2006, pp. 437–446.

[6] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *International Journal of Synthetic Emotions (IJSE)*, vol. 1, no. 1, pp. 68–99, 2010.

[7] M. M. Bradley and P. J. Lang, "Measuring emotion: Behavior, feeling, and physiology," *Cognitive neuroscience of emotion*, vol. 25, pp. 49–59, 2000.

[8] J. Kim and E. André, "Emotion recognition based on physiological changes in music listening," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 12, pp. 2067–2083, 2008.

[9] R. W. Picard and S. B. Daily, "Evaluating affective interactions: Alternatives to asking what users feel," in *CHI Workshop on Evaluating Affective Interfaces: Innovative Approaches*, 2005, pp. 2119–2122.

[10] S. Fairclough, "Physiological data must remain confidential." *Nature*, vol. 505, no. 7483, pp. 263–263, 2014.

[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," *Affective Computing, IEEE Transactions on*, vol. 3, no. 1, pp. 18–31, 2012.

[12] J. D. Morris, "Observations: Sam: the self-assessment manikin an efficient cross-cultural measurement of emotional response," *Journal of advertising research*, vol. 35, no. 6, pp. 63–68, 1995.

[13] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[14] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image and Vision Computing*, vol. 31, no. 2, pp. 120–136, 2013.

[15] H. Gunes, "Dimensional and continuous analysis of emotions for multimedia applications: a tutorial overview," in *Proceedings of the 20th ACM international conference on Multimedia*. ACM, 2012, pp. 1531–1532.

[16] B. Romera-Paredes, M. S. Aung, M. Pontil, N. Bianchi-Berthouze, A. de C Williams, and P. Watson, "Transfer learning to account for idiosyncrasy in face and body expressions," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 2013, pp. 1–6.

[17] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 10, pp. 1175–1191, 2001.

[18] H. Meng, A. Kleinsmith, and N. Bianchi-Berthouze, "Multi-score learning for affect recognition: the case of body postures," in *Affective Computing and Intelligent Interaction*. Springer, 2011, pp. 225–234.

[19] M. Soleymani, S. Koelstra, I. Patras, and T. Pun, "Continuous emotion detection in response to music videos," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*. IEEE, 2011, pp. 803–808.

[20] M. Soleymani, G. Chanel, J. J. Kierkels, and T. Pun, "Affective characterization of movie scenes based on content analysis and physiological changes," *International Journal of Semantic Computing*, vol. 3, no. 02, pp. 235–254, 2009.

[21] J. Wang and Y. Gong, "Recognition of multiple drivers emotional state," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*. IEEE, 2008, pp. 1–4.

[22] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin, "The elements of statistical learning: data mining, inference and prediction," *The Mathematical Intelligencer*, vol. 27, no. 2, pp. 83–85, 2005.

[23] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4, pp. 411–430, 2000.

[24] B. N. Cuthbert, H. T. Schupp, M. M. Bradley, N. Birbaumer, and P. J. Lang, "Brain potentials in affective picture processing: covariation with autonomic arousal and affective report," *Biological psychology*, vol. 52, no. 2, pp. 95–111, 2000.

[25] A. Kleinsmith, N. Bianchi-Berthouze, and A. Steed, "Automatic recognition of non-acted affective postures," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 41, no. 4, pp. 1027–1038, 2011.

[26] H. J. Griffin, M. S. Aung, B. Romera-Paredes, C. McLoughlin, G. McKeown, W. Curran, and N. Bianchi-Berthouze, "Laughter type recognition from whole body motion," in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, 2013, pp. 349–355.