Analysis of EEG signals and facial expressions for continuous emotion detection

Mohammad Soleymani, *Member, IEEE,* Sadjad Asghari-Esfeden, *Student member, IEEE* Yun Fu, *Senior Member, IEEE,* Maja Pantic, *Fellow, IEEE*

Abstract—Emotions are time varying affective phenomena that are elicited as a result of stimuli. Videos and movies in particular are made to elicit emotions in their audiences. Detecting the viewers' emotions instantaneously can be used to find the emotional traces of videos. In this paper, we present our approach in instantaneously detecting the emotions of video viewers' emotions from electroencephalogram (EEG) signals and facial expressions. A set of emotion inducing videos were shown to participants while their facial expressions and physiological responses were recorded. The expressed valence (negative to positive emotions) in the videos of participants' faces were annotated by five annotators. The stimuli videos were also continuously annotated on valence and arousal dimensions. Long-short-term-memory recurrent neural networks (LSTM-RNN) and Continuous Conditional Random Fields (CCRF) were utilized in detecting emotions automatically and continuously. We found the results from facial expressions to be superior to the results from EEG signals. We analyzed the effect of the contamination of facial muscle activities on EEG signals and found that most of the emotionally valuable content in EEG features are as a result of this contamination. However, our statistical analysis showed that EEG signals still carry complementary information in presence of facial expressions.

Index Terms—Affect, EEG, facial expressions, video highlight detection, implicit tagging

1 Introduction

Video is a form of visual art that conveys emotions in its content. Depending on users' mood and the context they are in, they prefer different type of emotional content, for example, users in negative mood are more likely to prefer a sad stimulus [1]. Affective characterization in multimedia can be used to improve multimedia retrieval and recommendation [2]. Emotion detection is an effective way to unobtrusively identify emotional traces of videos or other content without interrupting viewers. The collected emotional trace is more reliable since we do not have to rely on emotional self-reports that can be influenced by different social and personality factors, e.g., male participants are less likely to report fear. The focus of this paper is on continuous emotion recognition in response to videos and the cross-modality interference between EEG signals and facial expressions. Emotional highlights of video can be detected based on the emotions expressed by viewers. The moments in which emotions are expressed will build an emotional profile or trace with applications in video summarization, video indexing [3] and movie rating estimation [4]. For example, a movie trailer can be built based on the moments which evoke the strongest emotional response in the audiences.

There are three different perspectives dealing with affect in multimedia, namely, expressed emotions, felt emotions and expected emotions [5], [2]. Expressed or intended emotions are the ones that the artist or content creator is intending to elicit in the audience; independent of the fact that the users feel those emotions or not. Expected emotions are the emotions that arise as a result of the content in most of its audience. Felt emotions are the ones that audience individually feel and can be personal.

In this work, our main aim was to detect felt emotions; as we were labeling and detecting expressions and physiological responses. However, we also annotated the videos by the expected emotions and tried to detect the expected emotions from the responses as well.

Psychologists proposed and identified different models for representing emotions [6]. Discrete emotions such as happiness and disgust are easier to understand since they are based on the language. However, they can fall short in expressing certain emotions in different languages, for example, there was no exact translation for disgust in Polish [7]. On the other hand, emotions can be presented in multi-dimensional spaces which are derived based on studies that identified the axes that carry the largest variance of all the possible emotions. Valence, arousal and dominance (VAD) space is one of the most well known dimensional representations of emotions. Valence ranges from unpleasant to pleasant;

- *Mohammad Soleymani is with the centre interfacultaire en sciences affectives (CISA), University of Geneva, Switzerland. Email: mohammad.soleymani@unige.ch.*
- *Sadjad Asghari-Esfeden is with Department of Electrical and Computer Engineering, College of Engineering, Northeastern University, Boston, MA 02115. E-mail: sadjad@ece.neu.edu.*
- *Yun Fu is with the Department of Electrical and Computer Engineering, College of Engineering, and College of Computer and Information Science, Northeastern University, 403 Dana Research Center, 360 Huntington Avenue Boston, MA 02115. E-mail: yunfu@ece.neu.edu.*
- *Maja Pantic is with the department of computing, Imperial College London, SW7 2AZ, UK and with the Faculty of Electrical Engineering, Mathematics and computer science, University of Twente, the Netherlands e-mail: (see http://www.doc.ic.ac.uk/~maja).*

arousal ranges from calm to activated and can describe emotions intensity; and dominance ranges from not in control to dominant [8]. Given that most of the variance in emotions comes from two dimensions, i.e., arousal and valence, continuous annotation and recognition of emotions are often performed using these dimensions.

The goal of this work was to detect emotions continuously from EEG signals and facial expressions. The utilized data set consists in viewers responses to a set of emotional videos. First, five annotators continuously annotated the emotional responses visible in facial expressions of participants watching videos. We also annotated the stimuli continuously for both valence and arousal to indicate the level of expected arousal and valence with the help of five annotators. The averaged continuous annotations served as the ground truth for our continuous emotion detection system. We analyzed EEG signals and facial expressions of multiple participants and extracted emotional features for continuous emotion detection. Power spectral density (PSD) from EEG signals and facial landmarks were extracted as features. We applied different regression models, including linear regression, support vector regression, continuous conditional random fields and recurrent neural networks, which has been successfully used in the previous studies [9]. We evaluated the valence detection results in a 10-fold cross validation strategy using averaged correlation coefficients and Root-Mean-Square Error (RMSE) metrics.

We performed statistical analyses to identify the relationship between EEG signals and facial expressions. We were particularly interested to identify how much of the emotion detection using EEG signals can be attributed to the electromyogenic artifacts caused by facial expressions. We used a linear mixed-effect model and causality analysis using Granger causality [10] to answer these questions.

The main contributions of this work are as follows. First, we detected continuous valence, in both time and space, using EEG signals and facial expressions. Second, we studied the correlation between the EEG power spectral features and features extracted from facial expressions and continuous valence annotations to look for the possible cross modality effect of muscular electromyogram (EMG) activities. The analysis included Granger causality tests to identify the causality relationship between facial expressions and EEG features. We also performed statistical analyses to verify whether EEG features provide any additional information in presence of features extracted from facial expressions. Finally, we applied the models trained with the continuously annotated data on EEG responses that could not be interpreted due to the lack of facial expressions from the users. We found facial expression analysis in combination with the LSTM-RNN gives the best results for this given task. We also observed that an important part of the affective information in EEG signals is caused by electromyogenic interferences from facial expressions. However, we could not completely rule out the existence of emotion related information in EEG signals independent of facial expression analysis. This paper is an extended version of our previously published work [11]. In comparison to that work, the following analysis and methods are extended and added: (i) we redid and simplified the facial expression analysis which led into improved performance; an extended description of the facial features are also added (ii) we added statistical analysis and discussion on the effect of facial expressions on EEG signals and the band-pass filter (between 4Hz-45H) was removed in EEG pre-processing; (iii) an analysis of annotation delay and its effect on performance was added; (iv) different architectures for the Long-Short-Term-Memory Recurrent Neural Network (LSTM-RNN) and its parameters were tested; (v) we added the statistical analysis identifying the relationship between affective information in EEG signals and facial expressions and (vi) we annotated the stimuli continuously for both arousal and valence dimensions and reported the results for expected emotional trace detection.

The remaining of this paper is organized as follows. In Section 2, a background on continuous emotion detection and emotion detection from physiological signals is given. The data collection protocols are presented in Section 3. The method for emotion recognition from facial expressions and EEG signals are given in Section 4. An analysis on the correlations between EEG signals and facial expressions and their relationship is given in Section 5. The experimental results are presented in Section 6. Finally, the paper is concluded in Section 7.

2 Background

2.1 Continuous emotion recognition

Wöllmer et al. [12] suggested abandoning the emotional categories in favor of dimensions and applied it on emotion recognition from speech. Nicolaou et al. [13] used audio-visual modalities to detect valence and arousal on SEMAINE database [14]. They used Support Vector Regression (SVR) and Bidirectional Long-Short-Term-Memory Recurrent Neural Networks (BLSTM-RNN) to detect emotion continuously in time and dimensions. Nicolaou et al. also proposed using an output-associative relevance vector machines (RVM) which smooths the RVM output for continuous emotion detection [15]. Although they showed how it improved the performance of RVM for continuous emotion detection they did not compare its performance directly to the BLSTM recurrent neural network.

One of the major attempts in advancing the state of the art in continuous emotion detection was the Audio/Visual Emotion Challenge (AVEC) 2012 [16] which was proposed using SEMAINE database. SEMAINE database includes the audio-visual responses of participants recorded while interacting with the Sensitive Affective Listeners (SAL) agents. The responses were continuously annotated on four dimensions of valence, activation, power, and expectation. The goal of the AVEC 2012 challenge was to detect the continuous dimensional emotions using audio-visual signals. In other notable work, Baltrusaitis et al. [17] used continuous conditional random fields (CCRF) to jointly detect the emotional dimensions of AVEC 2012 continuous sub-challenge. They achieved superior performance compared to the Support Vector regression (SVR). For a comprehensive review of continuous emotion detection, we refer the reader to [9].

2.2 Implicit tagging for video tagging and summarization

Implicit tagging refers to identifying metadata describing the content, including tags and traces, from users' spontaneous reactions [18].

Movie ratings are indicators of viewers' desired for watching movies and are often used in recommendation applications. Spontaneous reactions were used to estimate these ratings. In a study on estimating movie ratings [4], Galvanic Skin Responses (GSR) were recorded and analyzed from movie audience. Their method could achieve better results incorporating GSR responses in addition to the demography information for two out of three studied movies. Bao et al. [19] used facial expressions, acoustic, motion and user interaction features captured on mobile phones and tablets to estimate movie ratings. They discovered that the individual responses are too diverse and unreliable to be taken into account and focused on processing the collective responses of the audience. McDuff et al. [20], [21] measured the level of smile from the video advertisement's audience to assess their preference for the content. They collected a large number of samples from users' webcams who were recruited using crowdsourcing. Ultimately, they were able to detect fairly accurately the desire to watch the video again and whether the viewers liked the videos.

Spontaneous behavioral reactions were also used for video highlight detection. A video highlight detection method was proposed using facial expression analysis [3], [22]. In their study, Joho et al. used a probabilistic emotion recognition from facial expression analysis to detect emotions. Results were reported for 10 participants watching eight video clips. Different features were developed for video highlight detection including expression change rate and pronounce level. Highly expressive emotions including surprise and happiness received high pronounce levels whereas no expression or neutral had low pronounce levels. Physiological linkage between multiple viewers were proposed for video highlight detection by Chênes et al [23]. Galvanic Skin Response (GSR) and skin temperature were the most informative signals in detecting video highlights using physiological linkage. In a more recent study, Fleureau et al. [24] used GSR responses of a group of audience simultaneously to create an emotional trace of movies. The traces generated from the physiological reposes were shown to match the user reported highlight.

Spontaneous reactions can be also translated into affective tags, for example, a sad video. Physiological signals have been used to detect emotions with the goal of implicit emotional tagging. A video affective representation technique through physiological responses was proposed in [25]. In this affective representation study, physiological responses were recorded from eight participants while watched 64 movie scenes. Participants' self-reported emotion were detected from their physiological responses using a linear regression trained by a relevance vector machine (RVM). Kierkels et al. [26] extended these results and analyzed the effectiveness of tags detected by physiological signals for personalized affective tagging of videos. Emotional labels were reconstructed by mapping the quantized arousal and valence values. The videos were ranked for retrieval mapping queries to valence-arousal space. A similar method was developed for emotional characterization of music videos using ridge regression where arousal, valence, dominance, and like/dislike ratings were estimated from the physiological responses and the stimuli content [27]. Koelstra et al. [28] employed central and peripheral physiological signals for emotional tagging of music videos. In a similar study [29], a multimodal emotional tagging was conducted using EEG signals and pupillary reflex. Khomami Abadi et al. [30] recorded and analyzed Magnetoencephalogram (MEG) signals as an alternative to the EEG signals with the ability to monitor brain activities.

3 Data set and Annotations

The data set, which was used in this study, is a part of MAHNOB-HCI database, which is a publicly available database for multimedia implicit tagging[1]. MAHNOB-HCI database includes recordings from two experiments. The recordings that were used in this paper are from the experiment where we recorded participants' emotional responses to short videos with the goal of emotional tagging.

3.1 Stimuli video clips

We chose 20 videos as emotion evoking stimuli to cover the whole spectrum of emotions. 14 out of 20 videos were excerpts from movies and were chosen based on the preliminary study. In the preliminary study, the participants self-reported their felt emotion on nine-point scales, including arousal (ranging from calm to excited/activated) and valence (ranging from unpleasant to pleasant). The initial set of videos in the preliminary study consisted in 155 movie scenes from famous commercial movies, including, ”The Pianist”, ”Love Actually”, ”The Shining”; for the full list of movies from which the scenes were selected we refer the reader to [31]. To facilitate emotional self-assessment SAM

[1] http://mahnob-db.eu/hci-tagging/

manikins were shown above the ratings [32]. Three additional videos, well-known for eliciting certain emotions, e.g., funny, were selected based on authors' judgment from online resources and were added to this set (two for joy and one for disgust). To have a number of neutral videos, three old weather forecast reports were selected from YouTube. Finally, we selected 20 videos to be shown as stimuli. Video clips were between 34.9s to 117s long ($M = 81.4s$, $SD = 22.5s$). Psychologists suggested videos from one to ten minutes long for eliciting single emotions [33], [34]. In MAHNOB-HCI data collection, the video clips were kept as short as possible to reduce the chance of co-occurring emotions or habituation to the stimuli [35]. However, the clips were kept long enough to convey the content and elicit the desired emotions.

3.2 Data collection

28 healthy volunteers were recruited on campus, comprising 12 male and 16 female between 19 to 40 years old. Different bodily responses were recorded from the participants watching video clips. EEG signals were acquired from 32 active electrodes on 10-20 international system using a Biosemi Active II system. To capture facial expression, a frontal view video was recorded at 60 frames per second. For a detailed explanation of the synchronization method, hardware setup and the database, we refer the reader to the original database article [31].

A subset of 239 trials (every recorded response of a participant to a video is a trial) containing visible facial expressions was selected to be annotated in this study. A large number of these frontal videos from participant faces do not contain visible facial expressions according to the authors' judgment. The lead author initially annotated all the 540 videos available in this data set and chose the 239 trials that could be further annotated. We then trained five annotators, including the authors and researches from Northeastern University to conduct the continuous annotation from participants' facial expressions in the laboratory. Valence from the frontal videos were annotated continuously using a software implemented based on FEELTRACE [36] and a joystick. Unlike SEMAINE database [14] where the participants were engaged in a conversation with an agent, in this study, they were quiet and passively watching videos. Hence, the annotators were unable to annotate arousal, power or expectation. An example of the continuous annotations and their averaged curve which served as the ground truth are shown in Fig. 1. We also annotated the 20 stimuli videos by five annotators. We asked the annotators to indicate the level of "expected" valence and arousal for a given video continuously using a joystick.

We calculated the Cronbach's alpha from the pairwise correlation coefficients to measure the inter-rater agreement for the expression annotations ($M = 0.53$,$SD = 0.41$); 48.7% of sequences agreement were above 0.7. This is lower that what is reported for the SEMAINE database [14] which is due to the difference in the nature of the databases and the lack of speech in this database. In the case of stimuli annotations, we again calculated the Cronbach's alpha from the pairwise correlations of the annotations; 50% of the sequences' agreement were above 0.7 for arousal ($M = 0.58$,$SD = 0.38$); and 70% of sequences' agreement were above 0.7 for valence ($M = 0.64$,$SD = 0.40$).

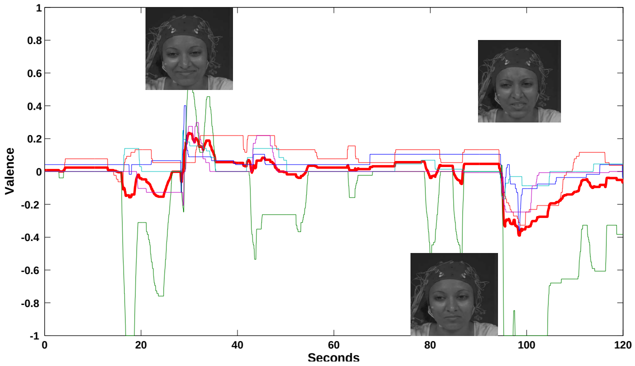


Fig. 1. The continuous annotations for one sequence and their averaged curve (the strong red line) are shown.

4 Methods

4.1 EEG signals

EEG signals were available at 256Hz sampling rate. EEG signals were re-referenced to the average reference to enhance the signal-to-noise ratio. Average re-referencing is performed when there is no reference electrode by subtracting the average amplitude of EEG signals from all the electrodes from every EEG signal recorded from any electrode. This averaged signal includes noise and artifacts that can be detected on the scalp but are not originated from the brain, e.g., electric potentials from cardiac activities.

The power spectral densities of EEG signals in different bands are correlated with emotions [37]. The power spectral densities were extracted from 1 second time windows with 50% overlapping. We used all the 32 electrodes for EEG feature extraction. The logarithms of the PSD from theta ($4Hz < f < 8Hz$), alpha ($8Hz < f < 12Hz$), beta ($12Hz < f < 30Hz$) and gamma ($30Hz < f$) bands were extracted to serve as features. In total number of EEG features of a trial for 32 electrodes and 4 bands is $32 \times 4 = 128$ features. These features were available at 2Hz temporal resolution due to the STFT window size (256).

4.2 Analysis of facial expressions

A face tracker was employed to track 49 facial fiducial points or landmarks [38]. In this tracker, a regression model is used to detect the landmarks from features. It then calculates translational and scaling difference

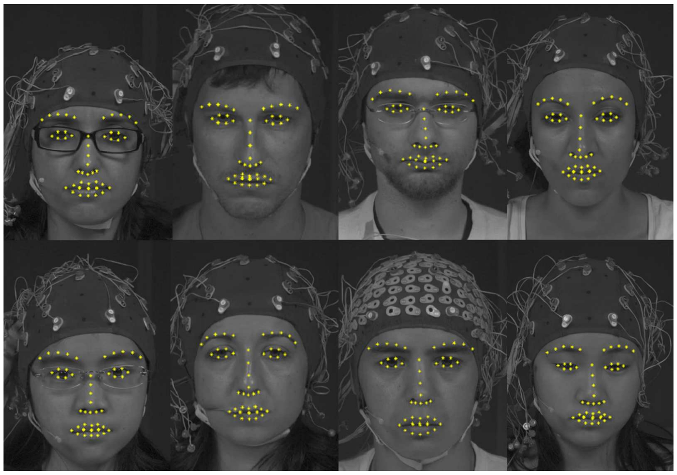


Fig. 2. Examples of the recorded camera view including tracked facial points.

between the initial and true landmark locations. The main idea for feature tracking is to use supervised descent method (SDM) for the detection in each frame by using the landmark estimate of the previous frame. The model is trained with 66 landmarks, and provides the coordinates of 49 points in its output (see Fig. 2). The facial points were extracted after correcting the head pose. An affine transformation was found between a neutral face of each subject and an averaged face of all the subjects and all the tracked points of sequences were registered using that transformation. A reference point was generated by averaging the coordinates of the inner corners of eyes and nose landmarks. We assumed this reference point to be stationary. The distances of 38 point including eyebrows, eyes, and lips to the reference point were calculated and averaged to be used as features.

4.3 Dimensional affect detection

Four commonly used regression models for similar studies were utilized for continuous emotion detection, namely, Multi-Linear Regression (MLR), Support Vector Regression (SVR), Continuous Conditional Random Fields (CCRF) [17], and Long Short Term Memory Recurrent Neural Networks (LSTM-RNN) [39].

4.3.1 Long Short Term Memory Neural Networks

Long Short Term Memory Recurrent Neural Networks (LSTM-RNN) have shown to achieve top performances in emotion recognition studies for audio-visual modalities [13], [9]. LSTM-RNN is a network which has one or more hidden layers including LSTM cells. These cells contain a memory block and some multiplicative gates which will determine whether the cell stores, maintains or resets its state. In this way, the network learns when to remember and forget information coming in a sequence over time and therefore it is able to preserve long-range dependencies in sequences. Recurrent Neural Networks are able to remember the short term input events through their feedback connections. LSTM adds the ability to also remember the input events from a longer period using the gated memory cell.

An open source implementation of LSTM[2] which is powered by NVIDIA Inc., Compute Unified Device Architecture (CUDA) technology was used in this paper. We chose to have two hidden layers containing LSTM cells for all the three configurations that we used based on the results we obtained trying different configurations (see Section 6.3). The number of hidden neurons were set to a quarter of the number of the input layer neurons, or features. The learning rate was set to 1E-4 with the momentum of 0.9. The sequences were presented in random order in training and a Gaussian noise with the standard deviation of 0.6 has been added to the input to reduce the problem of over-fitting. The maximum epochs in training were 100. If there was no improvement on the performance, i.e., sum of squared errors, on the validation set after 20 epochs, the training was stopped with the early stopping strategy.

4.3.2 Continuous Conditional Random Fields

Conditional Random Fields (CRF) are frameworks for building probabilistic models to segment and classify sequential data. Unlike Hidden Markov Models (HMM), they do not assume that the observations are conditionally independent and therefore are good alternatives for cases where there is a strong dependency between observations. Continuous Conditional Random Fields (CCRF) [17] were developed to extend the CRFs for regression. CCRF in this configuration acts as a smoothing operator on the per sample estimations from another model; in this work, we fed CCRF with the outputs estimated by a multi-linear regression. CCRF models a conditional probability distribution with the probability density function:

$$P(y|X) = \frac{exp(\Psi)}{\int_{-\infty}^{\infty} exp(\Psi)dy} \tag{1}$$

Where $\int_{-\infty}^{\infty} exp(\Psi)dy$ is the normalization function which makes the probability distribution a valid one (by making it sum to 1).

$$\Psi = \sum_{i} \sum_{k=1}^{K1} \alpha_k f_k(y_i, X) + \sum_{i,j} \sum_{k=1}^{K2} \beta_k g_k(y_i, y_j, X) \tag{2}$$

In this equation, Ψ is the potential function, $X = \{x_1, x_2, ..., x_n\}$ is the set of input feature vectors (matrix with per frame observation as rows, valence estimation from another regression technique such as MLR in our case), $Y = \{y_1, y_2, ..., y_n\}$ is the target, α_k is the reliability of f_k and β_k is the same for edge feature function g_k. f_k, the Vertex feature function, (dependency between y_i and $X_{i,k}$) is defined as:

$$f_k(y_i, X) = -(y_i - X_{i,k})^2 \tag{3}$$

[2] https://sourceforge.net/p/currennt

And g_k, the Edge feature function, which describes the relationship between two estimation at steps i and j, is defined as:

$$g_k(y_i, y_j, X) = -\frac{1}{2} S_{i,j}^{(k)} (y_i - y_j)^2 \tag{4}$$

The similarity measure, $S^{(k)}$, controls how strong the connections are between two vertices in this fully connected graph. There are two types of similarities:

$$S_{i,j}^{(\text{neighbor})} = \begin{cases} 1, & |i - j| = n \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

$$S_{i,j}^{(\text{distance})} = \exp\left(-\frac{|X_i - X_j|}{\sigma}\right) \tag{6}$$

The neighbor similarity, in Equation 5, defines the connection of one output with its neighbors and the distance similarity, in Equation 6, controls the relationship between y terms based on the similarity of x terms (by distance σ).

The CCRF can be trained using stochastic gradient descent. Since the CCRF model can be viewed as a multivariate Gaussian, the inference can be done by calculating the mean of the resulting conditional distribution, i.e., $P(y|X)$.

5 Artifacts and their Effect

There is often a strong interference of facial muscular activities and eye movements in the EEG signals. The facial muscular artifacts and eye movements are usually more present in the peripheral electrodes and higher frequencies (beta and gamma bands) [40]. We, hence, expected that the contamination from the facial expressions in the EEG signals to contribute to the effectiveness of the EEG signals for valence detection. In this Section, the EEG features were extracted from all the 32 electrodes (128 features from 4 bands).

To study this assumption, we used a linear mixed-effect model to test the effect of EEG features on estimating valence (annotated from the expressions) given the information from eye gaze and facial expressions (see Equation 7). Linear mixed-effect model enables us to model the between participant variations in a random effect term while studying the effect of the independent variables (EEG, face and eye gaze) on the dependent variable, valence. The eye movements were taken from the data recorded by the Tobii eye gaze tracker at 60Hz which we resampled to 4Hz to match the other modalities. The facial point movements were defined by how much they moved from one sample to the next. We calculated this in our feature set (see Section 4.2)

$$v_i = \beta_0 + \sum_{j=1}^{N} \beta_k x_{ij} + \sum_{k=1}^{q} b_k z_i + \epsilon \tag{7}$$

In this equation, v_i is the valence at i-th sample, x_{ij} is the feature, including facial, EEG and eye gaze; z_i is the random effect variable representing different subjects and is only one when the sample i belongs to the k-th subject; and ϵ is the random error that is normally distributed with zero mean. The result of the analysis showed that the coefficients of the majority of the EEG features were significantly different from zero. The percentage of the features with significant effect on valence estimations in different bands are as follows: theta band: 41% alpha band: 47% beta band: 50% and gamma: 81%; significance was defined by the results of the ANOVA test rejecting the null hypothesis with p-value smaller than 0.05. In total, the percentage of the EEG coefficient which was significantly different from zero was 55%. The average (absolute) β for the face features was 0.057 in comparison to 0.002 for EEG and 0.0005 for gaze. Given that the face features and eye gaze features were present in the model as fixed effect, this shows that the EEG features add information which was not in the facial movements detected by face tracker and useful for valence detection. It is however, unclear whether this information was originally related to the facial movement artifacts which were not detected by the computer vision based face trackers or not.

We also calculated the correlation between EEG spectral power in different bands and all the electrodes with the valence labels. The topographs in Fig. 3 show that the higher frequency components from electrodes positioned on the frontal, parietal and occipital lobes have higher correlation with valence measurements.

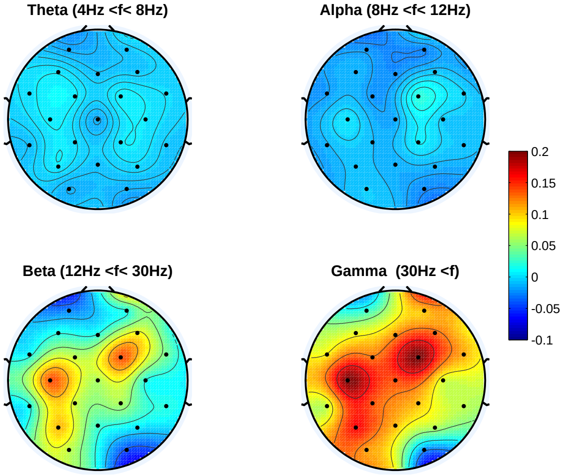


Fig. 3. The correlation maps between PSD and continuous valence for theta, alpha, beta, and gamma bands. The correlation values are averaged over all sequences. In these topographs the frontal lobe, the nose, is positioned on the top.

In order to check how much of the variation of EEG signals are correlated with the facial movement, we used a multi-linear regression model to estimate the EEG features from eye gaze, head pose and face features. In this analysis, we used a larger set of samples including the trials in which the participants did not show any visible facial expression. Eye gaze movement and head pose could not predict the EEG features, however, the features extracted from facial expressions could detect different EEG features with different degrees. Eye movements usually affect low frequency components of EEG signals which were discarded in these set of EEG features. The R-squared results of the estimation of the EEG features from face features are shown in Fig. 4. The association between EEG signals and facial expressions cannot be as a result of motor cortex activation since Mu rhythms are associated with motor cortex activation originated by movement (9-11Hz) [41] were not very strong. Mu Rhythm has the same frequency as alpha waves and originated from the motor cortex; see the alpha activity in Fig. 4. Comparing Figs. 3 and 4, we can see a large overlap between the features detecting valence and the features with a strong correlation and dependency on face features. Even though, using the linear mixed-effect model we found that the EEG features were still significant in detecting valence in the presence of face features, most of the variance which was useful in detecting valence was strongly correlated with facial muscle artifacts. The most effective facial points on EEG signals included the landmarks on lips.

We wanted to further clarify whether the effect is originated from facial expressions or EEG signals. In order to detect the relationship between two time series, we used Granger causality test. Granger [10] stated that $\mathbf{y}$ Granger causes $\mathbf{x}$, if adding time series $\mathbf{y}$ to the auto regressive time series $\mathbf{x}$ reduces the variance of prediction error of $\mathbf{x}$. Given time series $\mathbf{x} = \{x_1, x_2, ..., x_n\}$ that can be described by an auto regressive model

$$x_i = \sum_{j=1}^{k_x} w_{(k_x - j)} x_{(i-j)} + \epsilon_x \tag{8}$$

where k_x is the model order for auto regressive time series, calculated using model selection criteria, like Baysian Information Criterion (BIC) or Akaike Information Criterion (AIC). If time series $\mathbf{y}$ is:

$$y_i = \sum_{j=1}^{k_y} w'_{(k_y - j)} y_{(i-j)} + \epsilon_y \tag{9}$$

we can reconstruct $\mathbf{x}$ by:

$$x_i = \sum_{j=1}^{k_x} w_{(k_x - j)} x_{(i-j)} + \sum_{j=1}^{k_y} w''_{(k_y - j)} y_{(i-j)} + \epsilon_{xy} \tag{10}$$

In equation 10 the first component is $\mathbf{x}$ being described using an autoregressive model, see Equation 8, and the second component models reconstructing $\mathbf{x}$ from the lagged values of $\mathbf{y}$. The presence of Granger causality can be determined by an F-test using the following F-value:

$$F = \frac{(\epsilon_x^2 - \epsilon_{xy}^2)/k_y}{\epsilon_{xy}^2/(n - (k_x + k_y + 1))} \tag{11}$$

We can then verify whether the null hypothesis, or the absence of Granger causality, can be rejected or not; in this paper we used the significance level of 5%.

We generated the squared EEG signals resampled to 60Hz to represent their energy and to match the facial expressions sampling rate. We then performed the Granger causality test from each of 38 facial points to each of 32 EEG signals from all the electrodes and back. We found that in average 28% of the causality test are positive when we test whether the facial expressions are Granger caused by EEG signals whereas this percentage was 54% for EEG signals being under influence of facial expressions. In Fig. 5, we show the averaged percentage of causality tests that came positive for different electrodes when we tested the causality from facial expressions to EEG signals. We can observe that it has a similar pattern to the correlation analysis and further strengthens the idea that a large part of the variance in EEG signals is as a result of facial expressions artifacts.

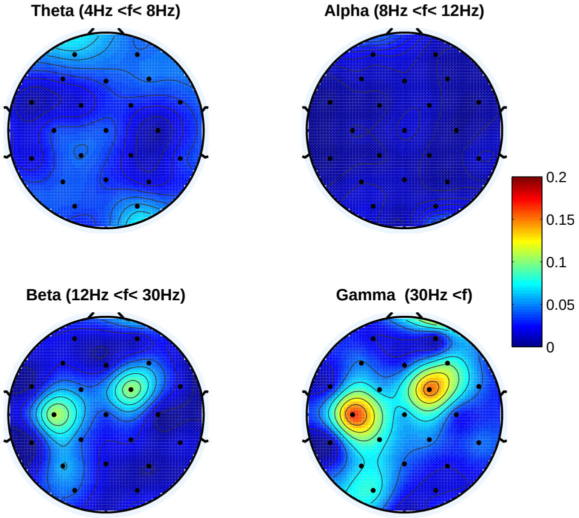


Fig. 4. The R-squared topographs depicting how much the EEG power spectral features could be estimated and as a result of facial movements or expressions.

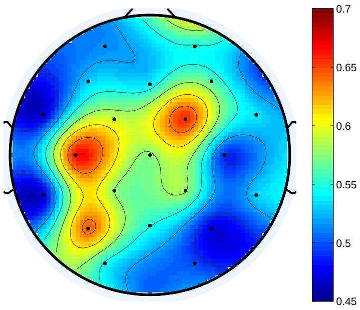


Fig. 5. The average percentage of significant causation from facial points to different EEG signals.

6 Experimental results

6.1 Analyses of features

We calculated the correlation between different facial expression features and the ground truth for each sequence and averaged them over all sequences. The features with the highest correlation coefficients were related to the mouth/lip points, e.g., lower lip angle ($\rho = 0.23$), left and right lip corner distance ($\rho = -0.23$), etc. The results showed that the lip points were the most informative features for valence detection. The openness of mouth and eyes were also strongly correlated with valence.

6.2 Analysis of the lag of annotations

Mariooryad and Busso [42] analyzed the effect of lag on continuous emotion detection on SEMAINE database [14]. They found a delay of 2 seconds will improve their emotion detection results. SEMAINE database is annotated by Feeltrace [36] using a mouse as annotation interface. In contrast, our database is annotated using a joystick which we believe has a shorter response time and is easier to use for such purpose. We shifted our annotations and considered delays from 250ms up to 4 seconds and calculated the averaged correlation coefficients for the valence detection results from different modalities and their combinations. The results, in Fig. 6, showed that in our case the delay of 250ms increased the detection performance whereas longer delays deteriorated it. Therefore, we observed the advantage of using joystick over mouse in this context.

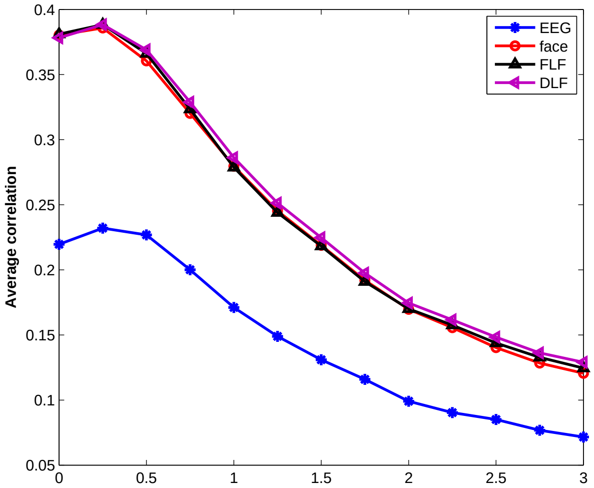


Fig. 6. The resulting $\bar{\rho}$ for different hypothetical annotation delays.

6.3 LSTM structure

We experimented different LSTM structures from 1 hidden layer to two hidden layers and with different number of LSTM cells. The results are summarized in Table 1. Two hidden layers yielded higher detection performance compared to one. We tested with different number of neurons in the hidden layer including 1/8, 1/4, and 1/2 the input layer neurons or the number of features. Different number of neurons did not change the results significantly; we confirmed this with performing a one-sided non-parametric rank-sum test. Therefore, we chose a quarter of the input neurons with two hidden layers for the final settings. Weninger et al. [43] reported that adding the differences between the consecutive targets as outputs improved their regression performance in music emotion estimation from content features. They argue that adding the target differences acts like regularization of the weights. We tried the same technique in which we added the difference between the targets of each point with its neighboring samples ($\Delta(v)/\Delta(t)$) in the training phase and discard those output during testing. Unlike [43], the results of including the target differences ($\Delta(v)/\Delta(t)$) did not have any significant effect on the performance. Hence, we did not use the extra differential targets in our final model.

TABLE 1
The performance of continuous valence detection using different structures of LSTM network.

	One hidden layer		Two hidden layers	
#cells	$\bar{\rho}$	RMSE	$\bar{\rho}$	RMSE
EEG				
16	0.24±0.34	0.053±0.029	0.25±0.34	0.052±0.028
32	0.25±0.34	0.052±0.028	0.26±0.33	0.052±0.029
64	0.22±0.35	0.053±0.029	0.26±0.35	0.053±0.028
Face				
5	0.48±0.37	0.043±0.026	0.48±0.37	0.045±0.027
10	0.47±0.37	0.044±0.026	0.49±0.37	0.043±0.025
19	0.47±0.39	0.044±0.025	0.47±0.38	0.044±0.027
Face and EEG (feature fusion)				
21	0.40±0.33	0.047±0.025	0.42±0.34	0.047±0.024
42	0.39±0.34	0.048±0.025	0.40±0.35	0.047±0.025
83	0.35±0.34	0.050±0.023	0.40±0.35	0.047±0.024

6.4 Continuous emotion detection

In this Section, we present the results of continuous emotion detection in case of felt-emotions. All the features and annotations were re-sampled to 4Hz from their original sampling rates. To recall, EEG features were the power spectral densities in different bands and facial expression features were extracted from the facial landmarks detected in each frame. This re-sampling enabled us to perform multimodal fusion on different levels. All the features were normalized by removing the average value of the features in the training set and dividing by their standard deviation. The results were evaluated in a 10-fold cross validation which is not participant-independent. However, the training, validation and test

TABLE 2
To evaluate the detection performances from different modalities and fusion schemes the averaged Pearson correlation coefficient ($\bar{\rho}$) and Root-Mean-Square Error (RMSE) are reported. The RMSE was calculated after scaling the output and labels between [-0.5, 0.5]. The reported measures are averaged for all the sequences for Multi-Linear Regression (MLR), Support Vector Regression (SVR), Continuous Conditional Random Fields (CCRF) and LSTM Recurrent Neural Network (LSTM-RNN). Modalities and fusion schemes were EEG, facial expressions (face), Feature Level Fusion (FLF) and Decision Level Fusion (DLF). For $\bar{\rho}$ higher is better, and for RMSE the lower is better. The best results are shown in boldface font.

Model	**MLR**		**SVR**		**CCRF**		**LSTM-RNN**	
Metric	$\bar{\rho}$	RMSE	$\bar{\rho}$	RMSE	$\bar{\rho}$	RMSE	$\bar{\rho}$	RMSE
EEG	0.22±0.36	0.055±0.030	0.21±0.35	0.060±0.027	0.26±0.49	0.048±0.035	0.24±0.34	0.053±0.029
Face	0.38±0.35	0.049±0.026	0.38±0.36	0.051±0.025	0.44±0.41	0.053±0.027	**0.48±0.37**	**0.043±0.026**
FLF	0.38±0.34	0.049±0.025	0.33±0.33	0.055±0.024	0.44±0.40	0.53±0.028	0.40±0.33	0.047±0.025
DLF	0.38±0.37	0.047±0.028	0.36±0.37	0.050±0.026	0.42±0.46	0.050±0.029	**0.45±0.35**	**0.044±0.026**

sets do not contain information from the same trials. In every fold, the samples were divided in three sets. 10% were taken as the test set, 60% of the remaining samples (54% of the total number of samples) were taken as the training set and the rest were used as the validation set. For the Multi-Linear Regression (MLR), only the training sets were used to train the regression models and the validation sets were not used. A linear ϵ-SVR with L2 regularization, from Liblinear library [44], was used and its hyper-parameters were found based on the lowest RMSE on the validation set. We used the validation sets in the process of training the LSTM-RNN to avoid overfitting. The output of MLR on the validation set was used to train the CCRF. The trained CCRF was applied on the MLR output on the test set. The CCRF regularization hyper-parameters were chosen based on a grid search using the training set. The rest of the parameters were kept the same as [17].

Two fusion strategies were employed to fuse these two modalities. In the Feature Level Fusion (FLF), the features of these modalities were concatenated to form a larger feature vector before feeding them into models. In the Decision Level Fusion (DLF), the resulting estimation of valence scores from different modalities were averaged.

The emotion recognition results are given in Table 2. We reported the averaged Pearson correlation to show the similarity between the detected curves and the annotations. Root-Mean-Squared Error (RMSE) is also reported to show how far the estimations were in average compared to the ground truth. RMSE penalizes the larger errors more than the smaller ones. Consistently, facial expressions outperformed EEG features. This might be as a result of the bias of the data set towards the trials with expressions. We could show that in this particular case unlike the work of Koelstara and Patras [45], facial expressions can outperform EEG signals for valence detection. It is worth noting that the facial landmark detector used in this study is a more recent landmark tracking technique compared to the one in [45] whereas the EEG features are identical. Moreover, compared to the previous work, we were using a different set of labels which were not based on self-reports. Finally, [45] was a single trial classification study whereas here we are detecting emotions continuously.

A one sided Wilcoxon test showed that the difference between the performance of the decision level fusion compared to the results of a single modality facial expression analysis using LSTM-RNN was not significant. Therefore, we conclude that the fusion of EEG signals is not beneficial and the LSTM-RNN on a single modality, i.e., facial expressions, performed the best in this setting. Using the non-parametric Wilcoxon test, we found that the difference between correlations resulted from LSTM-RNN and CCRF is not significant, however, RMSE is significantly lower for LSTM-RNN ($p < 1E - 4$). Although, direct comparison of the performance is not possible with the other work due to the difference in the nature of the databases, the best achieved correlation is in the same range as the result of [46], the winner of AVEC 2012 challenge, on valence and superior to the correlation value reported on valence in a more recent work, [47]. Unfortunately, the previous papers on this topic did not report the standard deviation of their results; thus its comparison was impossible. We have also tested the Bidirectional Long Short Term Recurrent Neural Networks (BLSTM-RNN), but their performance was inferior compared to the simpler LSTM-RNN for this task.

Two examples of detection results are given in Fig. 7. We observed that positive emotions were easier to detect compared to the negative ones. Smiles are strong indicators of pleasantness of emotions and there are a large number of instances of smile in the current data set. It was also shown that smile detection has a high performance in spontaneous expression recognition [48]. Our analysis on the face features also showed that most of the highly correlated features were the points on lips.

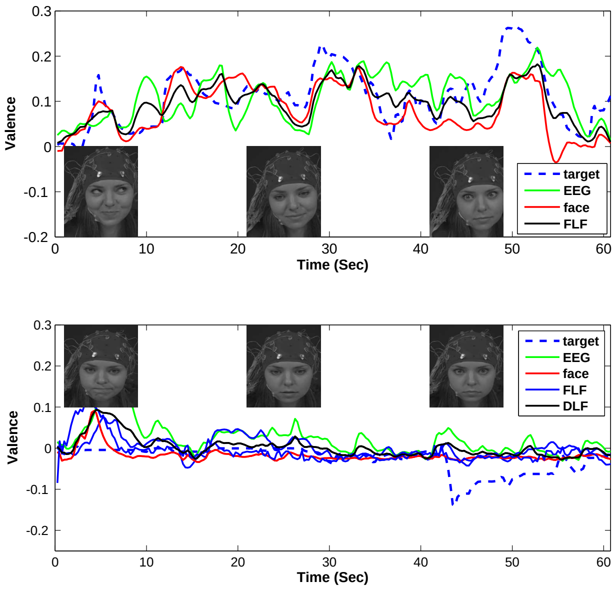


Fig. 7. Two examples of the detected valence traces. The results depicted on top (ρ_{FLF}=0.80, RMSE$_{FLF}$=0.034) is a good example of correct detection of the trend in a positive stimulus. The example shown in the bottom (ρ_{FLF}=-0.11, RMSE$_{FLF}$=0.036), on the other hand is an example where the detection did not pick up the correct trend for this negative stimulus.

TABLE 3
The detection results with the aim of recovering the expected emotion using the labeled stimuli.

Model	Arousal		Valence	
Metric	$\bar{\rho}$	RMSE	$\bar{\rho}$	RMSE
EEG	0.16±0.29	0.145±0.047	0.14±0.34	0.205±0.083
Face	0.27±0.38	0.142±0.053	0.26±0.42	0.188±0.087
FLF	0.21±0.30	0.143±0.050	0.21±0.34	0.193±0.083
DLF	0.25±0.32	0.143±0.049	0.25±0.37	0.185±0.077

6.5 Expected emotion detection

If the goal of emotion detection is to identify the emotional highlight and emotional trace of videos, then the ground truth should reflect the expected emotions. For this reason, we repeated the same procedure for detecting the expected emotion this time using both continuous arousal and valence labels given to the stimuli. When we annotated the stimuli directly we did not have the same limitation as annotating facial expressions that made annotating arousal impossible. This time, we only report the results of the best performing model, i.e., LSTM-RNN. The results are given in Table 3. Again the facial expressions results are superior to the EEG signals and fusion of two modalities do not outperform facial expressions. It is worth noting that even without the labels crated based on the expressions, see Section 6.4, facial expressions outperform the EEG signals. The results are in average inferior to the results obtained using the expression-based labels in Section 6.4. This might be as a result of different factors that we summarize as follows. Not all the viewers at all times feel the emotions that are expected; e.g., a person might have already watched a surprising moment in a movie scene and it is surprising to her for the second time. Often times we only express or feel high level of emotion or arousal only in the first moment of facing a stimulus; the emotional response usually decays afterward due to the habituation to the stimulus [35]. Moreover, the personal and contextual factors, such as mood, fatigue and familiarity with the stimuli have effect on what we feel in response to videos [2]. One solution to these problems can be to combine several users' responses. Multi-user fusion has been shown to achieve superior results in emotional profiling of videos using physiological signals compared to single-user response [24].

In order to verify whether the model trained on annotations based on facial expression analyses can reflect on the case without any facial expressions. We chose one of the videos with distinct emotional highlight moments, the church scene from “Love Actually”, and took the EEG responses of the 13 participants who did not show any significant. Since these responses did not include any visible facial expressions, they were not used for the annotation procedure and were not in any form in our training data. We extracted the power spectral features from their EEG responses and fed it into our models and averaged the output curves. Fig. 8 show that despite the fact that the participants did not express any facial expressions and likely did not have very strong emotions, the valence detected from their EEG responses covaries with the highlight moments and the valence trend in the video. The CCRF provides a smoother trace which matches better with the overall trend compared to the other methods. The snapshots in Fig. 8, show the frames corresponding to three different moments. The first one, at 20 seconds, is during the marriage ceremony. The second and third frames are the surprising and joyful moments when the participants bring their musical instruments in sight and start playing a romantic song unexpectedly.

7 Conclusions

We presented a study of continuous detection of valence using EEG signals and facial expressions. Promising results were obtained from EEG signals and facial expressions. We expected the results from facial expressions to be superior due to the bias of the ground truth towards the expressions, i.e., the ground truth was generated based on the judgment of the facial expressions. The results from the statistical test on the linear mixed-effect model showed that EEG modality carries useful information for detecting valence. The information gain from including EEG signals however did not improve

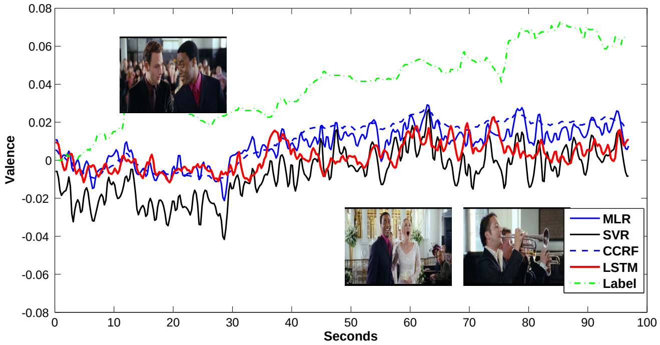


Fig. 8. The average valence curve, emotional trace, resulted from the EEG signals of participants who did not show any facial expressions while watching a scene from Love Actually. The general trend is correlated with the emotional moments. The labels given directly to the stimulus were divided by 10 for visualization purposes and are in green dash dots. The figure is best seen in color.

the performance of the detection in our different fusion approaches. It is also possible that the rest of the valence related information in EEG signals were also related to subtle facial expressions which were not captured by the machine vision based face tracker. The results of detecting the labeled stimuli again demonstrated that facial expressions had superior performance compared to the EEG signals for emotion detection. In summary, facial expressions outperformed EEG signals for emotion detection in this study. The data set that was used in this study consisted of responses with apparent facial expressions with face videos captured in high quality with limited head movements and pose variance. However, this conclusion cannot be generalized to all conditions and all emotion detection scenarios as the opposite was reported previously [45].

Even with inferior performance, EEG signals can still be considered in applications where we cannot record users' faces. The current wearable headsets, e.g. EMOTIV[3], demonstrate the potentials in EEG analysis. Moreover, with correct electrode placement electromyogenic signals and the artifacts from facial expression on EEG signals can be used to detect facial expressions to replace the need for a front-facing camera capturing users' faces.

The analyses of the correlation between the EEG signals and the ground truth showed that the higher frequency components of the signals carry more important information regarding the pleasantness of emotion. The analysis of correlation and causation between EEG signals and facial expressions further showed that the informative features from EEG signals were mostly originated from the contamination from facial muscular activities. The analysis of the valence detection also showed that the results were superior for the sequences with positive emotions compared to the negative ones. The continuous annotation of facial expressions suffers from a lag and we showed the superior time response of annotators while using joystick instead of using a mouse [14], [42].

In the future work, it will be beneficial to record multimodal data for the sole aim of understanding the correlation and causation relationships between facial expressions and EEG signals. Recordings of the future work can include acted and spontaneous expressions in different situations.

[3] http://emotiv.com/

Acknowledgments

This work of Soleymani was partially supported by the European Research Area under the FP7 Marie Curie Intra-European Fellowship: Emotional continuous tagging using spontaneous behavior (EmoTag). The work of Pantic is supported by the European Community Horizon 2020 [H2020/2014-2020] under grant agreement no. 645094 (SEWA). The work of Asghari-Esfeden and Fu was supported in part by the NSF CNS award 1314484, ONR award N00014-12-1-1028, ONR Young Investigator Award N00014-14-1-0484, and U.S. Army Research Office Young Investigator Award W911NF-14-1-0218.

References

[1] P. G. Hunter, E. Glenn Schellenberg, and A. T. Griffith, "Misery loves company: Mood-congruent emotional responding to music," *Emotion*, vol. 11, no. 5, pp. 1068–1072, 2011.

[2] M. Soleymani, M. Larson, T. Pun, and A. Hanjalic, "Corpus development for affective video indexing," *IEEE Transactions on Multimedia*, vol. 16, no. 4, pp. 1075–1089, 2014.

[3] H. Joho, J. Staiano, N. Sebe, and J. Jose, "Looking at the viewer: analysing facial activity to detect personal highlights of multimedia contents," *Multimedia Tools and Applications*, vol. 51, no. 2, pp. 505–523, 2010.

[4] F. Silveira, B. Eriksson, A. Sheth, and A. Sheppard, "Predicting audience responses to movie content from electro-dermal activity signals," in *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2013, UbiComp '13, pp. 707–716.

[5] N. Malandrakis, A. Potamianos, G. Evangelopoulos, and A. Zlatintsi, "A supervised approach to movie emotion tracking," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. May 2011, pp. 2376–2379, IEEE.

[6] K. R. Scherer, "What are emotions? And how can they be measured?," *Social Science Information*, vol. 44, no. 4, pp. 695–729, 2005.

[7] J. A. Russell, "Culture and the Categorization of Emotions," *Psychological Bulletin*, vol. 110, no. 3, pp. 426–450, 1991.

[8] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of Research in Personality*, vol. 11, no. 3, pp. 273–294, 1977.

[9] 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.

[10] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: Journal of the Econometric Society*, pp. 424–438, 1969.

[11] M. Soleymani, S. Asghari-Esfeden, M. Pantic, and Y. Fu, "Continuous emotion detection using eeg signals and facial expressions," in *Multimedia and Expo, 2014. ICME 2014. IEEE International Conference on*, 2014.

[12] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, "Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies.," in *INTERSPEECH*, 2008, pp. 597–600.

[13] M. Nicolaou, H. Gunes, and M. Pantic, "Continuous Prediction of Spontaneous Affect from Multiple Cues and Modalities in Valence-Arousal Space," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 92–105, 2011.

[14] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, 2012.

[15] M. A. Nicolaou, H. Gunes, and M. Pantic, "Output-associative rvm regression for dimensional and continuous emotion prediction," *Image and Vision Computing, Special Issue on The Best of Automatic Face and Gesture Recognition 2011*, vol. 30, no. 3, pp. 186–196, 2012.

[16] B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic, "AVEC 2012: the continuous audio/visual emotion challenge," in *ACM International Conference on Multimodal Interaction (ICMI)*, 2012, pp. 449–456.

[17] T. Baltrusaitis, N. Banda, and P. Robinson, "Dimensional affect recognition using continuous conditional random fields," in *IEEE Int' Conf. Automatic Face Gesture Recognition (FG)*, 2013, pp. 1–8.

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

[19] X. Bao, S. Fan, A. Varshavsky, K. Li, and R. Roy Choudhury, "Your reactions suggest you liked the movie: Automatic content rating via reaction sensing," in *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, New York, NY, USA, 2013, UbiComp '13, pp. 197–206, ACM.

[20] D. McDuff, R. El Kaliouby, and R. W. Picard, "Crowdsourcing facial responses to online videos," *Affective Computing, IEEE Transactions on*, vol. 3, no. 4, pp. 456–468, 2012.

[21] D. McDuff, R. el Kaliouby, D. Demirdjian, and R. Picard, "Predicting online media effectiveness based on smile responses gathered over the internet," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, 2013, pp. 1–7.

[22] H. Joho, J. M. Jose, R. Valenti, and N. Sebe, "Exploiting facial expressions for affective video summarisation," in *Proceeding of the ACM International Conference on Image and Video Retrieval*, New York, NY, USA, 2009, CIVR '09, ACM.

[23] C. Chênes, G. Chanel, M. Soleymani, and T. Pun, "Highlight detection in movie scenes through inter-users, physiological linkage," in *Social Media Retrieval*, N. Ramzan, R. van Zwol, J.-S. Lee, K. Clüver, and X.-S. Hua, Eds., Computer Communications and Networks, pp. 217–237. Springer London, 2013.

[24] J. Fleureau, P. Guillotel, and I. Orlac, "Affective Benchmarking of Movies Based on the Physiological Responses of a Real Audience," in *Affective Computing and Intelligent Interaction and Workshops, 2013. ACII 2013. 3rd International Conference on*, September 2013, pp. 73–77.

[25] M. Soleymani, G. Chanel, J. J. M. Kierkels, and T. Pun, "Affective Characterization of Movie Scenes Based on Content Analysis and Physiological Changes," *Int'l Journal of Semantic Computing*, vol. 3, no. 2, pp. 235–254, June 2009.

[26] J. J. M. Kierkels, M. Soleymani, and T. Pun, "Queries and tags in affect-based multimedia retrieval," in *ICME'09: Proceedings of the 2009 IEEE international conference on Multimedia and Expo*, Piscataway, NJ, USA, 2009, pp. 1436–1439, IEEE Press.

[27] M. Soleymani, S. Koelstra, I. Patras, and T. Pun, "Continuous emotion detection in response to music videos," in *IEEE Int' Conf. Automatic Face Gesture Recognition (FG)*, march 2011, pp. 803–808.

[28] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Y. Patras, "DEAP: A database for emotion analysis using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, pp. 18–31, 2012.

[29] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," *Affective Computing, IEEE Transactions on*, vol. 3, no. 2, pp. 211–223, 2012.

[30] M. K. Abadi, S. M. Kia, R. Subramanian, P. Avesani, and N. Sebe, "User-centric Affective Video Tagging from MEG and Peripheral Physiological Responses," in *Affective Computing and Intelligent Interaction and Workshops, 2013. ACII 2013. 3rd International Conference on*, September 2013, pp. 582–587.

[31] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Transactions on Affective Computing*, vol. 3, pp. 42–55, 2012.

[32] M. M. Bradley and P. J. Lang, "Measuring emotion: the Self-Assessment Manikin and the Semantic Differential.," *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 25, no. 1, pp. 49–59, March 1994.

[33] A. Schaefer, F. Nils, X. Sanchez, and P. Philippot, "Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers," *Cognition & Emotion*, vol. 24, no. 7, pp. 1153–1172, 2010.

[34] J. Rottenberg, R. D. Ray, and J. J. Gross, *Emotion elicitation using films*, pp. 9–28, Series in affective science. Oxford University Press, 2007.

[35] R. J. Barry and E. N. Sokolov, "Habituation of phasic and tonic components of the orienting reflex," *International Journal of Psychophysiology*, vol. 15, no. 1, pp. 39–42, 1993.

[36] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. Mcmahon, M. Sawey, and M. Schröder, "'feeltrace': an instrument for recording perceived emotion in real time," in *ISCA Workshop on Speech and Emotion*, 2000.

[37] R. J. Davidson, "Affective neuroscience and psychophysiology: toward a synthesis.," *Psychophysiology*, vol. 40, no. 5, pp. 655–665, 2003.

[38] Xuehan-Xiong and F. De la Torre, "Supervised descent method and its application to face alignment," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[39] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[40] I. Goncharova, D. McFarland, T. Vaughan, and J. Wolpaw, "EMG contamination of EEG: spectral and topographical characteristics," *Clinical Neurophysiology*, vol. 114, no. 9, pp. 1580–1593, Sept. 2003.

[41] E. Niedermeyer, "The normal EEG of the waking adult," *Electroencephalography: Basic principles, clinical applications, and related fields*, p. 167, 2005.

[42] S. Mariooryad and C. Busso, "Analysis and compensation of the reaction lag of evaluators in continuous emotional annotations," in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, Sept 2013, pp. 85–90.

[43] F. Weninger, F. Eyben, and B. Schuller, "On-line continuous-time music mood regression with deep recurrent neural networks," in *Proc. of ICASSP 2014, Florence, Italy*. 2014, p. 4 pages, IEEE, to appear.

[44] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *The Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[45] S. Koelstra and I. Patras, "Fusion of facial expressions and eeg for implicit affective tagging," *Image and Vision Computing*, vol. 31, no. 2, pp. 167–174, 2013.

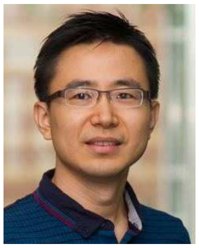
[46] J. Nicolle, V. Rapp, K. Bailly, L. Prevost, and M. Chetouani, "Robust continuous prediction of human emotions using multiscale dynamic cues," in *ACM International Conference on Multimodal Interaction (ICMI)*, 2012, pp. 501–508.

[47] Y. Song, L.-P. Morency, and R. Davis, "Learning a sparse codebook of facial and body microexpressions for emotion recognition," in *ACM International Conference on Multimodal Interaction (ICMI)*, 2013, pp. 237–244.

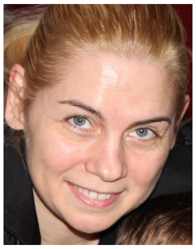
[48] D. J. McDuff, *Crowdsourcing affective responses for predicting media effectiveness*, Ph.D. thesis, Massachusetts Institute of Technology, 2014.

Mohammad Soleymani (S'04-M'12) received his PhD in computer science from the University of Geneva, Switzerland in 2011. From 2012 to 2014, he was a Marie Curie Fellow at the intelligent Behaviour Understanding Group (iBUG) at Imperial College London, where he conducted research on sensor-based and implicit emotional tagging. He is currently a Swiss SNF Ambizione fellow at the University of Geneva, Switzerland. His research interests include affective computing, multimedia information retrieval, and multimodal interactions. He is one of the founding organizers of the MediaEval benchmarking campaign. He has served as an associate editor and guest editor for the IEEE Transactions on Affective Computing and Journal of Multimodal User Interfaces. He has also served as an area chair, program committee member and reviewer for multiple conferences and workshops including ACM MM, ACM ICMI, ISMIR, and IEEE ICME.

Sadjad Asghari-Esfeden received the B.Sc. degree in computer engineering from the Department of Electrical and Computer Engineering, University of Tehran, Iran, in 2013. He is currently working toward his Ph.D. degree in computer engineering at Northeastern University, USA. His current research interests include applied machine learning, computer vision, and social media analysis.



Yun Fu (S'07-M'08-SM'11) received the B.Eng. degree in information engineering and the M.Eng. degree in pattern recognition and intelligence systems from Xi'an Jiaotong University, China, respectively, and the M.S. degree in statistics and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign, respectively. Prior to joining the Northeastern faculty, he was a Scientist working at BBN Technologies, Cambridge, MA, during 2008-2010. He holds a Part-Time Lecturer position in the Department of Computer Science, Tufts University, Medford, MA, in 2009. He was a tenure-track Assistant Professor of the Department of Computer Science and Engineering, State University of New York, Buffalo, during 2010-2012. He is an interdisciplinary faculty member affiliated with College of Engineering and the College of Computer and Information Science at Northeastern University since 2012. His research interests are Interdisciplinary research in Machine Learning and Computational Intelligence, Social Media Analytics, Human-Computer Interaction, and Cyber-Physical Systems. He has extensive publications in leading journals, books/book chapters and international conferences/workshops. He serves as associate editor, chairs, PC member and reviewer of many top journals and international conferences/workshops. He is the recipient of 5 best paper awards (SIAM SDM 2014, IEEE FG 2013, IEEE ICDM-LSVA 2011, IAPR ICFHR 2010, IEEE ICIP 2007), 3 young investigator awards (2014 ONR Young Investigator Award, 2014 ARO Young Investigator Award, 2014 INNS Young Investigator Award), 2 service awards (2012 IEEE TCSVT Best Associate Editor, 2011 IEEE ICME Best Reviewer), the 2011 IC Postdoctoral Research Fellowship Award, the 2010 Google Faculty Research Award, the 2008 M. E. Van Valkenburg Graduate Research Award, the 2007-2008 Beckman Graduate Fellowship, 2007 Chinese Government Award for Outstanding Self-Financed Students Abroad, the 2003 Hewlett-Packard Silver Medal and Science Scholarship, Edison Cups of the 2002 GE Fund Edison Cup Technology Innovation Competition, and the 2002 Rockwell Automation Master of Science Award. He is currently an Associate Editor of the IEEE Transactions on Neural Networks and Leaning Systems (TNNLS), and IEEE Transactions on Circuits and Systems for Video Technology (TCSVT). He is a Lifetime Member of ACM, AAAI, SPIE, and Institute of Mathematical Statistics, member of INNS and Beckman Graduate Fellow during 2007-2008.



Maja Pantic (M'98, SM'06, F'12) is a professor in affective and behavioral computing in the Department of Computing at Imperial College London, United Kingdom, and in the Department of Computer Science at the University of Twente, the Netherlands. She currently serves as the editor in chief of Image and Vision Computing Journal and as an associate editor for both the IEEE Transactions on Pattern Analysis and Machine Intelligence and the IEEE Transactions on Affective Computing. She has received various awards for her work on automatic analysis of human behavior, including the European Research Council Starting Grant Fellowship 2008 and the Roger Needham Award 2011.