CONTINUOUS EMOTION DETECTION USING EEG SIGNALS AND FACIAL EXPRESSIONS

Mohammad Soleymani[1], *Sadjad Asghari-Esfeden*[2], *Maja Pantic*[1,3], *Yun Fu*[2]

[1] Imperial College London, UK, [2] Northeastern University, USA, [3]University of Twente, Netherlands
{m.soleymani, m.pantic}@imperial.ac.uk {sadjad, yunfu}@ece.neu.edu

ABSTRACT

Emotions play an important role in how we select and consume multimedia. Recent advances on affect detection are focused on detecting emotions continuously. In this paper, for the first time, we continuously detect valence from electroencephalogram (EEG) signals and facial expressions in response to videos. Multiple annotators provided valence levels continuously by watching the frontal facial videos of participants who watched short emotional videos. Power spectral features from EEG signals as well as facial fiducial points are used as features to detect valence levels for each frame continuously. We study the correlation between features from EEG and facial expressions with continuous valence. We have also verified our model's performance for the emotional highlight detection using emotion recognition from EEG signals. Finally the results of multimodal fusion between facial expression and EEG signals are presented. Having such models we will be able to detect spontaneous and subtle affective responses over time and use them for video highlight detection.

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

1. INTRODUCTION

Multimedia content is made to induce emotions and be emotionally expressive. From drama to comedy, different genres of multimedia induce different emotions and are appealing to their audience in different mood and context. Affective features of multimedia are therefore an invaluable source of information for multimedia indexing and recommendation [1]. Given the difficulty of collecting emotional self-report to multimedia from users, emotion recognition is an effective way of collecting users' emotional feedback in response to multimedia for the purpose of multimedia indexing. In this paper, we focus on continuous emotion recognition in response to videos. The continuous emotion detection in response to videos will enable us to detect the emotional highlights of a video. The highlights and emotional moments can be used for video summarization [2], movie rating estimation [3] and affective indexing. For example, a user wants to retrieve the funny moments in a movie will be able to do so based on the continuous profile provided by this technique from the spontaneous responses of other users.

Different emotional representation models, including discrete and dimensional have been proposed by psychologists [4]. Discrete emotions, e.g., sadness, joy, fear, are easier to understand but are limited in describing the whole spectrum of emotions in different languages. Dimensional models of emotions represent emotions in different dimensions where an emotion can be mapped into a point in that space. One of the most adopted dimensional emotion representation model is valence and arousal space. Arousal ranges from calm to excited/activated, and valence ranges from unpleasant to pleasant [5].

The contributions presented in this paper are as follows. First, to the best of our knowledge, this is the first attempt in detecting continuous emotions, in both time and dimension, using EEG signals. Second, we detect continuous emotions from facial expressions and provide the multimodal fusion results. Third, we study the correlation between the EEG power spectral features that we used for emotion recognition and continuous valence annotations to look for the possible effect of muscular artifacts. Finally, we apply 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. In this work, the emotional responses visible in the frontal camera capturing facial expressions are annotated continuously in time and valence dimension by five annotators. The averaged annotations served as a ground truth to be detected from facial expression analysis and EEG signals. Different regression models, utilized in the similar state of the art studies [6], were tested, and the performance of continuous emotion detection was evaluated using a 10-folding cross validation strategy.

2. BACKGROUND

Wollmer et al. [7] suggested abandoning the emotional categories in favor of dimensions and applied it on emotion recognition from speech. Nicolaou et al. [8] used audio-visual

The work of Soleymani is supported by Marie Curie Fellowship: Emotional continuous tagging using spontaneous behavior (EmoTag). The work of Pantic is supported in part by the EU Community's 7th Framework Programme (FP7/2007-2013) under the grant agreement no 231287 (SSPNet). The work of Asghari Esfeden and Fu is supported in part by the NSF CNS award 1314484, Office of Naval Research award N00014-12-1-1028, Air Force Office of Scientific Research award FA9550-12-1-0201.

modalities to detect valence and arousal on the SEMAINE database [9]. 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. One of the major attempts in advancing the state of the art in continuous emotion detection is the Audio/Visual Emotion Challenge (AVEC) 2012 [10]. The 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. For a comprehensive review of continuous emotion detection, we refer the reader to [6].

Physiological signals have been used to detect emotions with the goal of implicit emotional tagging. Soleymani et al. [11] proposed an affective characterization for movie scenes using peripheral physiological signals. Eight participants watched 64 movie scenes and self-reported their emotions. A linear regression trained by relevance vector machines (RVM) was utilized to estimate each clip's affect from physiological features. A similar approach was taken using a linear ridge regression for emotional characterization of music videos. Arousal, valence, dominance, and like/dislike rating was detected from the physiological signals and video content [12]. Koelstra et al. [13] used electroencephalogram (EEG) and peripheral physiological signals for emotional tagging of music videos.

3. DATA SET AND ANNOTATIONS

The dataset, which is used in this study, is the first experiment from MAHNOB-HCI database [14], which is a publicly available database for multimedia implicit tagging[1]. The experiments were conducted to record the participants' emotional responses to short videos with the goal of emotional tagging. The participants were shown 20 short videos, between 34.9s to 117s long ($M = 81.4s$, $SD = 22.5s$), to elicit different emotions. The short videos were movie scenes from famous commercially produced movies as well as some semi-professional, user-generated content from the Internet. The experimental data was collected from 28 healthy volunteers, comprising 12 male and 16 female between 19 to 40 years old. EEG signals were recorded from 32 active electrodes on 10-20 international system using a Biosemi Active II system. A frontal view video was captured at 60 frames per second with the goal of recording the facial expressions . The synchronization method, hardware setup and the database details are given in [14]. A subset of 239 responses containing visible facial expressions is selected to be analyzed in this study. The rest of the trials, except 10 trials that we used in the verification study, were discarded since the annotators were unable to annotate the expressions without any visible expression. Valence from the frontal videos were annotated continuously using a software implemented based on feeltrace [15] and a joystick. Unlike the SEMAINE database [9] where the participants are 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.

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

4. METHODS

4.1. EEG signals

EEG signals were available at 256Hz sampling rate. The unwanted artifacts, trend and noise were reduced by preprocessing the signals. EEG signals were re-referenced to the average reference to enhance the signal-to-noise ratio.

The spectral power of EEG signals in different bands was found to be correlated with emotions [16]. The power spectral densities were extracted from 1 second time windows with 50% overlapping. Koelstra et al. [13] studied the correlation between emotional dimensions, i.e., valence, arousal, dominance and EEG spectral power from different bands. They found the power spectral densities (PSD) of signals from following electrodes to be significantly correlated: Fp1, T7, CP1, Oz, Fp2, F8, FC6, FC2, Cz, C4, T8, CP6, CP2, PO4. Therefore, we only used these 14 electrodes for EEG feature extraction. The logarithms of the PSD from theta ($4Hz < f < 8Hz$), slow alpha ($8Hz < f < 10Hz$), alpha ($8Hz < f < 12Hz$), beta ($12Hz < f < 30Hz$) and gamma ($30Hz < f$) bands were extracted from all 14 electrodes to server as features. In addition to power spectral features, the difference between the spectral power of all the possible symmetrical pairs on the right and left hemisphere was extracted to measure the possible asymmetry in the brain activities due to the valence of emotional stimuli [17]. The asymmetry features were extracted from all 5 mentioned bands. The features from the selected electrodes comprised of 5 power spectral bands of 14 electrodes in addition to 3 symmetric pairs, i.e., T7-T8, Fp1-Fp2, and CP1-CP2. The total number of EEG features of a trial for 14 electrodes and 3 corresponding asymmetric features is $14 \times 5 + 3 \times 5 = 85$ features. These features were available at 2Hz temporal resolution due to the short time Fourier transform (STFT) window size ($w = 256$).

4.2. Analysis of facial expressions

An active appearance model face tracker was employed to track 40 points [18] (see Figure 1). The facial points were extracted after registering the face to a normalized face and correcting the head pose. A reference point was generated by averaging the inner corners of eyes and points on the subjects' nose which assumed to be stationary. The distances of 33 point including eyebrows, eyes, lips and iris to the reference point were calculated and averaged to be used as features. The first derivative of these distances was also calculated as

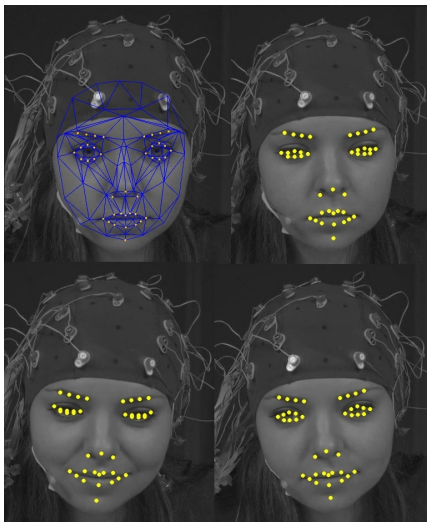

Fig. 1. Examples of the recorded camera view including tracked facial points. The top left image shows the active appearance model that is fit to the face.

the features reflecting the dynamics of facial expressions. Finally the angles between the horizontal line, the line connecting the inner corners of eyes, and outer corner of eyebrows as well as the angles between corners of lips were also calculated as features. We also extracted the distance between the central points of the upper lip and the lower lip as a measure of mouth openness. In total 271 features were extracted from facial points.

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), conditional random fields (CCRF) [19], and long short-term memory recurrent neural networks (LSTM-RNN) [20].

4.3.1. Long Short Term Memory Neural Networks

LSTM-RNN have shown to achieve top performances in emotion recognition studies for audio-visual modalities [8, 6]. 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. The number of hidden neurons were set to the half 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 overfitting. 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.

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

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) [19] are developed to extend the CRFs for 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:

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

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:

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

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

The neighbor similarity (Equation 5) shows the connection of one output with its neighbors and the distance similarity, (Equation 6), controls the relation 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. EXPERIMENTAL RESULTS

5.1. Analyses of features

There is often a strong interference of facial muscular activities and eye movements in the EEG signals. We hence expect that the contamination from the facial expressions in the EEG signals to contribute to the effectiveness of the EEG signals for valence detection. The facial muscular artifacts are usually more present in the peripheral electrodes and higher frequencies. To study this assumption, the correlation between the EEG features and continuous valence were calculated. The topographs in Figure 2 show that the higher frequency components from electrodes positioned on the frontal, parietal and occipital lobes have higher correlation with valence measurements. The location of the correlated electrodes undermines the assumption that the correlation between the EEG features and valence were due to the contamination from the electromyogram (EMG) signals, facial muscular activities, on the peripheral electrodes. However, the strong correlation from the higher frequency bands, beta and gamma, supports this assumption. We think the correlation is caused by a combination of the effect from the facial expression and brain activities.

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 correlations were related to the mouth points, e.g, ranked by their averaged correlation were: lower lip angle ($\rho = -0.15$), left lip corner distance ($\rho = -0.13$), and right lip corner distance ($\rho = -0.13$).

5.2. Continuous emotion detection

All the features and annotations were re-sampled to 4Hz from their original sampling rate. This enabled us to perform multimodal fusion on different levels. All the features were normalized by removing the average of the features in the training set and dividing by their standard deviation. The results were evaluated in a 10-folding cross validation. 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) were taken as the training set and the rest were used as the validation set. For the multi-linear regression (MLR) and support vector regression (SVR), only the training sets were used to train the regressors and the validation sets were not used. A linear ϵ-SVR with L2 regularization, from Liblinear library [21], was used and its hyper-parameters were found based on a grid search on the training set. We used the validation sets in the process of training the LSTM-RNN to avoid over-fitting. 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 [19].

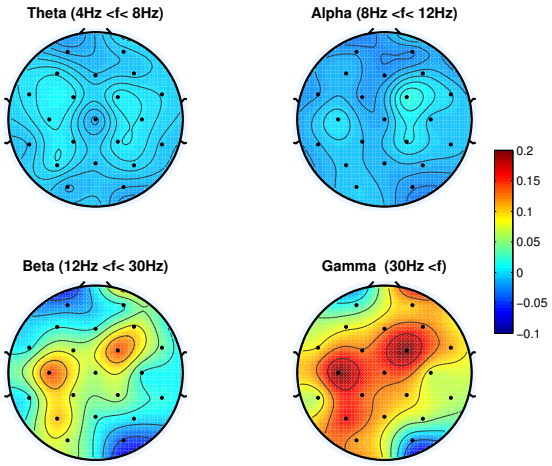

Fig. 2. The correlation maps between PSD and continuous valence for theta ,alpha, beta, and gamma bands. The correlation values are averaged over all sequences. In these topographs the frontal lobe, the nose, is positioned on top.

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 1. The LSTM-RNN achieved the best performance. In general, facial expression and EEG modalities performed similarly, even though the ground truth is heavily under the influence of the participants' facial expressions. This further confirms the finding of Koelstara and Patras [22], who showed that in case of single trial emotion recognition based on participants' self report, EEG signals outperform facial expressions. Regarding the correlation, although the highest average correlation is higher for the fusion of facial expressions and EEG modalities of CCRF, the average correlation resulting from decision level fusion of LSTM-RNN is very different with lower standard deviation. The lowest linear error is achieved with the LSTM-RNN and decision level fusion. Therefore, we conclude that the LSTM-RNN performed the best in this setting and with the goal of continuous valence detection. Although direct comparison of

the performance is not possible with the other works due to the difference in the nature of the databases, the best achieved correlation is in the same range as the result of [23], the winner of AVEC 2012 challenge, on valence and superior to the correlation value reported on valence in a more recent work, [24]. 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.

5.3. Emotional highlight detection

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 highlight moments, the church scene from "Love Actually", and took the EEG responses of the 10 participants who did not show any significant facial expression while watching that video clip. Since these responses did not include any visible facial expressions, they were not used in 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 regression models and averaged the output curves. The regression models were trained on all the available data which had annotations. The resulting valence detection are shown in Figure 3. The results show that despite the fact that the participants did not express any visible facial expressions and likely did not have very strong emotions, the valence detected from their EEG responses can still detect the highlight moments and the valence trend in the video. The CCRF provides a smoother profile compared to the other methods whereas all the methods are resulting fairly similar profiles. The snapshots in Figure 3, 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.

6. CONCLUSIONS

We presented a study of continuous detection of valence using EEG signals and facial expressions. Promising results are obtained from EEG signals. 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. However, the results from LSTM-RNN showed that EEG modality performance is not far inferior to the one of facial expressions. 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 and the informative features

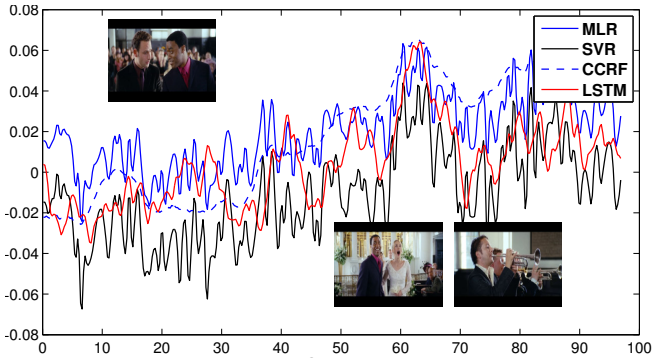

Fig. 3. The average valence curve, emotional pleasantness profile, resulted from the EEG signals of 10 participants who did not show any facial expressions while watching a scene from Love Actually. The joyful moments and a highlights are still detectable from the curve and its trend.

from EEG signals are not completely due to the contamination from facial muscular activities. The continuous annotation of facial expressions suffers from the lag and the lack of synchronicity of the annotators. In future, the continuous annotations should be aligned to improve the ground truth.

7. REFERENCES

[1] M. Soleymani, M. Larson, T. Pun, and A. Hanjalic, "Corpus development for affective video indexing," *IEEE Trans. Multimedia*, 2014, in press.

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

[3] F. Silveira, B. Eriksson, A. Sheth, and A. Sheppard, "Predicting audience responses to movie content from electro-dermal activity signals," in *ACM UbiComp'13*, 2013, pp. 707–716.

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

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

[6] 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.

[7] 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.

Table 1. To evaluate the detection performances from different modalities and fusion schemes the Pearson correlation coefficient (ρ) and averaged linear error or distances (Dist.) are reported. The Dist. 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 expression (face), Feature Level Fusion (FLF) and Decision Level Fusion (DLF).

Model	**MLR**		**SVR**		**CCRF**		**LSTM-RNN**	
Metric	$\bar{\rho}$	Dist.	$\bar{\rho}$	Dist.	$\bar{\rho}$	Dist.	$\bar{\rho}$	Dist.
EEG	0.21±0.35	0.043±0.024	0.21±0.34	0.047±0.023	0.26±0.44	0.048±0.030	0.28±0.33	0.040±0.022
face	0.28±0.41	0.046±0.028	0.28±0.39	0.055±0.034	0.32±0.46	0.053±0.029	0.28±0.42	0.043±0.027
FLF	0.30±0.37	0.043±0.024	0.29±0.36	0.050±0.027	0.34±0.44	0.048±0.027	0.30±0.37	0.041±0.022
DLF	0.29±0.40	0.040±0.023	0.29±0.39	0.043±0.023	0.34±0.46	0.043±0.025	**0.33±0.38**	**0.038±0.023**

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

[9] 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 Trans. Affective Computing*, vol. 3, no. 1, pp. 5–17, 2012.

[10] B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic, "AVEC 2012: the continuous audio/visual emotion challenge," in *ACM ICMI*, 2012, pp. 449–456.

[11] 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 J. Semantic Computing*, vol. 3, no. 2, pp. 235–254, June 2009.

[12] 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.

[13] 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 Trans. Affective Computing*, vol. 3, pp. 18–31, 2012.

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

[15] 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.

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

[17] S. K. Sutton and R. J. Davidson, "Prefrontal Brain Asymmetry: A Biological Substrate of the Behavioral Approach and Inhibition Systems," *Psychological Science*, vol. 8, no. 3, pp. 204–210, 1997.

[18] J. Orozco, O. Rudovic, J. Gonzàlez, and M. Pantic, "Hierarchical On-line Appearance-Based Tracking for 3D Head Pose, Eyebrows, Lips, Eyelids and Irises," *Image and Vision Computing*, vol. 31, no. 4, pp. 322–340, 2013.

[19] 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.

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

[21] 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.

[22] 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.

[23] J. Nicolle, V. Rapp, K. Bailly, L. Prevost, and M. Chetouani, "Robust continuous prediction of human emotions using multiscale dynamic cues," in *ACM ICMI*, 2012, pp. 501–508.

[24] Y. Song, L.-P. Morency, and R. Davis, "Learning a sparse codebook of facial and body microexpressions for emotion recognition," in *ACM ICMI*, 2013, pp. 237–244.