MULTIMEDIA IMPLICIT TAGGING USING EEG SIGNALS

Mohammad Soleymani[*]

Imperial College London, UK
m.soleymani@imperial.ac.uk

Maja Pantic[†]

Imperial College London, UK
University of Twente, the Netherlands
m.pantic@imperial.ac.uk

ABSTRACT

Electroencephalogram (EEG) signals reflect brain activities associated with emotional and cognitive processes. In this paper, we demonstrate how they can be used to find tags for multimedia content without users' direct input. Alternative methods for multimedia tagging is attracting increasing interest from multimedia community. The new portable EEG helmets are paving the way for employing brain waves in human computer interaction. In this paper, we demonstrate the performance of EEG for tagging purposes using two different scenarios on MAHNOB-HCI database. First, an emotional tagging and classification using a reduced set of electrodes is presented. The emotional responses of 24 participants to short video clips are classified into three classes on arousal and valence. We show how a reduced set of electrodes based on previous studies can preserve and even enhance the emotional classification rate. We then demonstrate the feasibility of using EEG signals for tag relevance tasks. A set of images were shown to participants first, without any tag and then with a relevant or irrelevant tag. The relevance of the tag was assessed based on the EEG responses of the participants in the first second after the tag was depicted. Finally, we demonstrate that by aggregating multiple participants' responses we can significantly improve the tagging accuracy.

Index Terms— EEG, affect, implicit tagging, tag relevance, multimedia

1. INTRODUCTION

Online multimedia repositories are becoming the media of choice. Classic content providers such as British Broadcasting Corporation (BBC) [1] are converting their whole archive into digital and using online interface, e.g., BBC iPlayer, to broadcast their content. Any large scale data will remain unused without proper indexing. Multimedia tags are any form of metadata that can be used to index the content to facilitate its finding and re-finding. Tags can come in different forms including semantic tags, affective tags and geotags [2]. In contrast to classic tagging schemes where users direct input is mandatory, human-centered implicit tagging was proposed [3] to gather tags and annotations without any effort from users. The main idea behind this passive tagging strategy is to use users' spontaneous reactions to a given content to identify tags. The resulting tags are called "implicit" since there is no need for users' direct input as reactions to multimedia are displayed spontaneously.

Multimedia indexing has focused on generating characterizations of content in terms of events, objects, etc. The judgment relies on cognitive processing combined with general world knowledge and is considered to be objective due to its reproducibility by users with a wide variety of backgrounds. An alternative to this approach to indexing is taking the affective aspects into account. Here, affect refers to the intensity and type of emotion that is evoked in a user while watching/listening to multimedia content [4]. Affective characteristics of multimedia are important features for describing multimedia content and can be presented by relevant emotional tags. Challenges and difficulties in using users' direct input or their self reported emotions [5] make implicit tagging a suitable alternative for recognizing emotional tags.

EEG signals reflect cognitive as well as affective processes in brain. Neuropsychologists have recognized certain emotional circuits in the brain that are activated during an emotional experience [6, 7]. There are also certain Event Related Potentials (ERP) which are associated with different types of stimuli. Koelstra et al [8] showed how N400 event-related potential appears in case of mismatch between the tag and videos. This ERP was observed in response to different types of stimuli [9, 10]. The contributions presented in this paper are two-fold. In the first part, we are verifying the finding by Koelstra et al. [11] on the correlation of certain electrodes with emotional dimensions. The second contribution is proposing an image tag relevance assessment system based on the aggregation of EEG responses. To the best of our knowledge this is the first study with classification results using EEG signals. We demonstrate how EEG signals can be

[*]The work of Soleymani is supported by the European Research Council under the FP7 Marie Curie Intra-European Fellowship: Emotional continuous tagging using spontaneous behavior (EmoTag).

[†]The work of Pantic is supported in part by the European Community's 7th Framework Programme (FP7/2007-2013) under the grant agreement no 231287 (SSPNet) and ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB).

used for tagging on one of the publicly available databases [12, 13, 14].

The rest of this paper is organized as follows. In Section 2, a background on implicit tagging is given. The data collection protocols are presented in Section 3. The method for emotional tagging and tag relevance assessment using EEG signals are given in Section 4 and Section 5 respectively. Finally, the paper is concluded in Section 6.

2. BACKGROUND

Pantic and Vinciarelli define implicit tagging as using non-verbal spontaneous behavior to find relevant keyword or tags for multimedia content [3]. Implicit tagging research has recently attracted researchers' attention, and a number of studies have been published [15]. Implicit tagging has been used in image annotation, video highlight detection, topical relevance detection and retrieval result re-ranking. The existing literature can be divided into two categories, one dealing with using emotional reactions to tag the content with the expressed emotion, e.g., laughter detection for hilarity [16], and the second group of studies using the spontaneous reactions for information retrieval or search results, e.g., eye gaze for relevance feedback [17].

The users' behavior and spontaneous reactions to multimedia data can provide useful information for multimedia indexing with the following scenarios: (i) direct assessment of tags: users spontaneous reactions will be translated into emotional keywords, *e.g.*, funny, disgusting, scary [18, 13, 16, 19]; (ii) assessing the relevance of explicit tags or topic relevance, *e.g.*, agreement or disagreement over a displayed tag or the relevance of the retrieved result [8, 12, 20, 21]; (iii) user profiling: a user's personal preferences can be detected based on her reactions to retrieved data and be used for re-ranking the results; (iv) content summarization: highlight detection is also possible using implicit feedbacks from the users [22, 23].

There have been multiple attempts into recovering the affective response to multimedia using EEG signals. Soleymani et al. [13] used EEG signals and eye gaze responses to detect emotion in response to videos. They showed that multimodal fusion can improve the detection performance. Koelstra et al. [11] used EEG signals to detect emotional tags for music videos. In a recent study, Koelstra and Patras [14] fused facial expressions analysis and EEG signals to detect two classes of arousal, valence and dominance on MAHNOB-HCI database [12].

Users respond differently to the expected, relevant and mismatching, irrelevant tags. Koelstra et al. [8] found significant differences between N400 ERP responses between relevant and irrelevant tags displayed on short videos. Jiao and Pantic [21] used facial expressions to assess the relevance of tags, displayed on images.

3. DATASET

The experimental data was collected from 30 healthy volunteers, comprising 12 male and 16 female between 19 to 40 years old. The subjects had normal or corrected to normal vision. The experiment was controlled by the Tobii studio software. EEG signals were recorded from 32 active electrodes on 10-20 international system using a Biosemi Active II system. The experiments were conducted in an environment with controlled temperature and illumination. The synchronization method, hardware setup and the database details are given in [12]. MAHNOB-HCI is a publicly available database for multimedia implicit tagging [1]. The data recorded from some participants were not analyzed due to technical problems, poor signal quality and unfinished data collection. Hence, the analysis results of this paper are only based on the responses recorded from 24 participants for the emotional tagging and 26 participants for tag relevance assessment.

3.1. Emotional Tagging Experiment

In the first experiment on emotional tagging, participants were shown 20 short videos, between 34.9s to 117s long ($M = 81.4s,\ SD = 22.5s$), to elicit different emotions. Participants were asked to report their felt emotions by reporting the felt arousal (ranging from calm to excited/activated) and valence (ranging from unpleasant to pleasant) on nine points scales. To simplify the interface a keyboard was provided with only nine numerical keys and the participant could answer each question by pressing one of the nine. In a separate study 9 participants were asked to rate the videos on 9 points scale for arousal and valence dimensions. The ground truth is defined based on the separate group's ratings [13].

3.2. Tag Relevance Experiment

In the second experiment, 28 images depicting human actions (e.g. handshake) were subsequently shown on their own and accompanied by a word tag that is either relevant or irrelevant to the shown action. Images were downloaded from Flickr[2] and were cropped and resized to 1280×695 pixels to be displayed on a display size of $51.9 \times 32.45cm$ with a resolution of 1280×800 pixels. The space under and above the image was filled with black pixels. The tags were overlaid under the image (see Fig. 1). For each image a correct and an incorrect tag was displayed in the total of 54 trials in random order. For each trial, the following procedure was taken. First, the untagged images were displayed for 5 seconds. This allowed the subject to get to know the content of the image. Second, the same image was displayed with a tag for 5 seconds. The subjects' behavior in this period contained their reaction to the displayed tag. Third, a question was displayed on the screen to ask whether the subject agreed with the suggested tag. The

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

[2] http://www.flickr.com

length of each trial was about 11 seconds. In this study, the trials in which the subjects' responses contradict the true tag relevance were discarded as confusing examples, e.g., a trial in which a subject agreed with an irrelevant tag was discarded.

Fig. 1. Example image depicting a human action including relevant tag ('Sit Down') as shown to the subjects. Part of the recorded eye gaze fixation and scan path of one subject is overlaid in red.

4. EEG SIGNALS FOR EMOTIONAL TAGGING

EEG signals were originally recorded with a 1024Hz sampling rate and later downsampled to 256Hz to reduce the memory and processing costs. The unwanted artifacts, trend and noise were reduced prior to extracting the features from EEG data by pre-processing the signals. Drift and noise reduction were done by applying a 4-45Hz band-pass filter. Biosemi active electrodes record EEG signals referenced to common mode sense electrode (CMS) as a part of its feedback loop. In order to gain the full common-mode rejection ratio (CMRR) at 50Hz, EEG signals should be re-referenced to another reference. EEG signals were thus re-referenced to the average reference to maximize signal to noise ratio.

The spectral power of EEG signals in different bands was found to be correlated with emotions [24, 25, 26]. Power spectral densities (PSD) from different bands were computed using Fast Fourier transform (FFT) and Welch algorithm [27]. In this method, the signal is split into overlapping segments and the PSD is estimated by averaging the periodograms. The averaging of periodograms results in smoother power spectrum. The PSD of each electrode's EEG signals was estimated using 15s long windows with 50% overlapping.

Koesltra et al. [11] studied the correlation between emotional dimensions, i.e., valence, arousal, dominance and EEG spectral power from different bands and found the following electrodes to be significantly correlated: Fp1, T7, CP1, Oz, Fp2, F8, FC6, FC2, Cz, C4, T8, CP6, CP2, PO4. We conducted the emotion detection based on both the full set, i.e., 32 electrodes, and the reduced set of the electrodes, i.e., 14 electrodes.

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 32 electrodes of 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 an emotional stimuli [28]. The asymmetry features were extracted from all 5 mentioned bands. In the full set, the total number of EEG features of a trial for 32 electrodes and 14 corresponding asymmetric features from 14 symmetric pairs is $32 \times 5 + 14 \times 5 = 230$ features. In the reduced set, we only used the signals of 14 electrodes out of 32 electrodes. 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.

4.1. Classification

We performed a participant independent emotion detection for emotional tagging. A leave-one-participant-out was subsequently used as a cross validation strategy. The most popular emotional class or tag can satisfy a larger population of viewers in a video retrieval scenario. Prior to the classification, features were normalized by subtracting their mean and dividing them by the standard deviation. The LIBSVM [29] implementation of a support Vector Machine (SVM) classifier with a Radial Basis Function (RBF) kernel was employed for classification. The kernel and classifier parameters were determined by a grid search based on the best average F1 score on a 10-folding cross validation on the training set. No feature selection was performed for this classification.

4.2. Results

F1 scores were calculated for the detection of three classes on arousal and valence (see Table 1). The results of the features extracted from 14 electrodes are equal to the full set and discarding 18 electrodes did not reduce the performance. These results are about 5% higher than the previously reported results on the same database [13]. In average the arousal detection results are superior to the valence detection results. This is in line with the previous results on the same database. The detection rate for the medium valence is lower than the low and high valence classes. This can be due to the smaller sample size for the medium valence class.

In order to aggregate the classification results from different participants, the confidence scores provided by SVM classifier was summed over different participants' results, on

Table 1. The F1 scores of emotion recognition for different dimensions.

Full set of 32 electrodes				
Dimension	**Low**	**Medium**	**High**	**Average**
Arousal	0.61	0.64	0.65	0.63
Valence	0.55	0.49	0.61	0.55
Reduced set of 14 electrodes				
Dimension	**Low**	**Medium**	**High**	**Average**
Arousal	0.61	0.65	0.66	0.64
Valence	0.59	0.47	0.63	0.56

the reduced electrodes set. In each aggregation step, a combination of $\binom{N}{n}$, $n \in \{2, ..., 24\}$, $N = 24$ participants were selected and their emotion detection results were combined. The resulting F1 scores from different combinations were then averaged. The aggregation of multiple participants can significantly improve the detection rate (see Fig. 2). The arousal and valence detection rates on three classes can go up to the average F1 scores of 0.89 and 0.83 respectively. Even using 5 participants is enough to increase the detection rate by more than 15%.

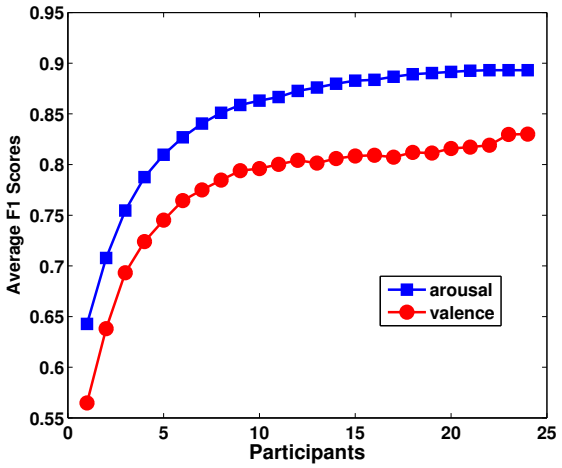

Fig. 2. Aggregation of multiple participants' emotion detection significantly improves the detection for implicit tagging. Arousal detection rate was superior to valence detection rates.

5. EEG SIGNALS FOR TAG RELEVANCE ASSESSMENT

Similarly to the last Section, we used the downsampled EEG signals at 256Hz. EEG drift was removed by subtracting the moving averaged signal with a 5 seconds (1280 point) window. This time, the noise reduction was done by applying a low-pass filter with the cut-off frequency of 10Hz, since the ERP responses are low frequency [30]. EEG signals were again re-referenced to the average reference.

We expected the ERP responses to appear in 400ms to 600ms after showing the tag. Therefore, the EEG signals of the one second period after displaying overlaid tags under the images were downsampled 16 times and used as features. As a result, we had $32 \times 16 = 512$ features for every trial. The trials in which the participants' responses contradict the relevant tag were discarded as confusing examples, e.g., a participant agreed with an irrelevant tag.

5.1. Classification

For this study, we conducted a participant dependent or intra-participant classification. For each participant, one of the 54 trials was taken as test set and the rest were used as the training set in the leave-one-out cross validation strategy. The size of feature vectors was reduced by applying Principal Component Analysis (PCA) and keeping the basis with 79% of the variance. This threshold was chosen based on the average classification performance. We used a Linear Discriminant Analysis (LDA) classifier to discriminate between the responses to relevant and non-relevant tags. In order to aggregate multiple participants' responses their EEG signals were averaged after normalization. The resulting features were classified using the same classifier and validated by the same cross validation strategy.

5.2. Results

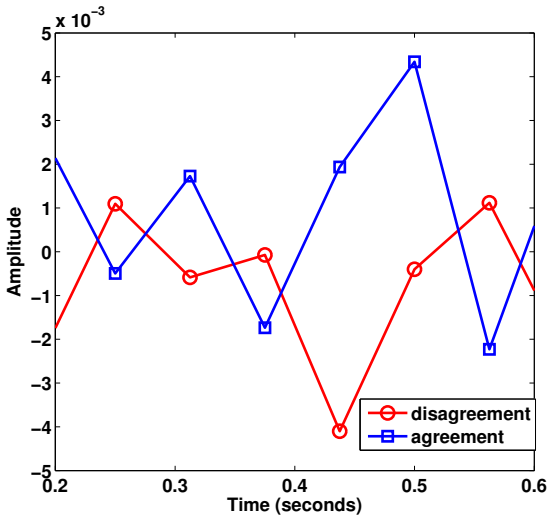

Fig. 3. Average downsampled EEG responses over all electrodes and all the participants. P300 ERP response for relevant and N400 ERP for irrelevant tags are visible.

The averaged features over all electrodes and all participants are shown in Fig. 3. The average ERP responses are

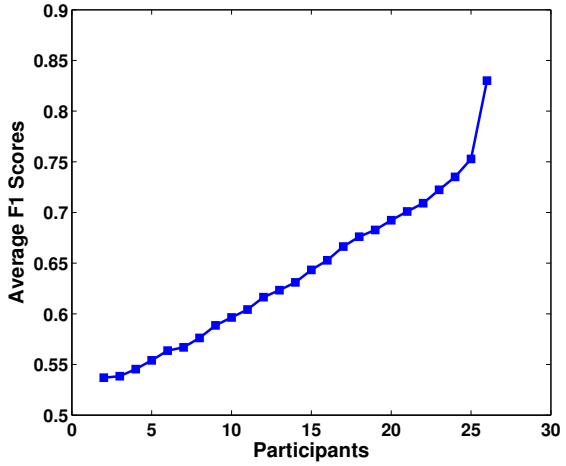

Fig. 4. Aggregation of multiple participants' ERP responses increases the signal to noise ratio and consequently the accuracy of tag relevance assessment.

different in case of relevant and irrelevant tags. The response to the relevant tags looks more like P300 response which is associated with the target tag in participant's mind. P300 responses are used for BCI spellers [31]. The response to the irrelevant tags is more like a delayed N400 response associated with mismatch. The F1 scores were calculated for every participant. The average F1 score was 0.56 with the standard deviation of 0.08 and maximum of 0.79. In order to verify the statistical significance of the results, a one sided one sample t-test with 0.0005 significance level was applied to the F1 scores. The t-test rejected the assumption that the detection rates were equal or inferior to the chance level.

In order to aggregate multiple participants' signal we simply average the features. Since the extracted features are downsampled and filtered EEG signals, we expect the averaging to improve the signal to noise ratio to detect the ERPs. In each aggregation step, a combination of different $\binom{N}{n}$, $n \in \{2, ..., 26\}$, $N = 26$ participants were selected and their features were averaged. The result of aggregation is given in Fig. 4. It is visible how the aggregation of multiple participants can increase the detection rate up to 0.83.

6. CONCLUSIONS

We presented two different scenarios of using EEG signals for human centered implicit tagging. In the emotional tagging approach, we showed how choosing a subset of electrodes preserved the detection rates compared to the full set of electrodes. Preserving the detection rate with a reduced set of electrodes enable us to use a simpler recording apparatus in future studies. We have also shown how aggregation leads into better emotion detection accuracy.

For the first time, we presented classification results on using EEG for tag relevance assessment. In case of tag relevance assessment, the performance varies between participants and only with aggregation or repeating trials we can expect good detection accuracy. This conclusion is in line with the Brain Computer Interfaces (BCI) which is based on ERPs. BCI protocols usually involve multiple stimuli to improve their accuracy [31]. The aggregation of multiple participants' responses provides a practical solution for the deployment of the current state of the art implicit tagging methods for real world applications.

7. REFERENCES

[1] J. Eggink and D. Bland, "A large scale experiment for mood-based classification of tv programmes," in *Multimedia and Expo (ICME), 2012 IEEE International Conference on*, July 2012, pp. 140–145.

[2] M. Larson et al., "Automatic tagging and geotagging in video collections and communities," in *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, New York, NY, USA, 2011, ICMR '11, pp. 51:1–51:8, ACM.

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

[4] A. Hanjalic and L.Q. Xu, "Affective video content representation and modeling," *Multimedia, IEEE Transactions on*, vol. 7, no. 1, pp. 143–154, 2005.

[5] R.W. Picard and S.B. Daily, "Evaluating Affective Interactions: Alternatives to Asking What Users Feel," in *CHI Workshop on Evaluating Affective Interfaces: Innovative Approaches*, 2005.

[6] R. Adolphs, D. Tranel, and A.R. Damasio, "Dissociable neural systems for recognizing emotions," *Brain and Cognition*, vol. 52, no. 1, pp. 61–69, June 2003.

[7] A. R. Damasio et al., "Subcortical and cortical brain activity during the feeling of self-generated emotions," *Nature Neuroscience*, vol. 3, no. 10, pp. 1049–1056, October 2000.

[8] S. Koelstra, C. Muhl, and I. Patras, "Eeg analysis for implicit tagging of video data," in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*. IEEE, 2009, pp. 1–6.

[9] G. Orgs, K. Lange, J.H. Dombrowski, and M. Heil, "Is conceptual priming for environmental sounds obligatory?," *International Journal of Psychophysiology*, 2007.

[10] S. Koelsch, E. Kasper, D. Sammler, K. Schulze, T. Gunter, and A.D. Friederici, "Music, language and meaning: brain signatures of semantic processing," *Nature neuroscience*, vol. 7, no. 3, pp. 302–307, 2004.

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

[12] 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.

[13] 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.

[14] 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.

[15] M. Soleymani and M. Pantic, "Human-centered implicit tagging: Overview and perspectives," in *Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on*, oct. 2012, pp. 3304–3309.

[16] S. Petridis and M. Pantic, "Is this joke really funny? judging the mirth by audiovisual laughter analysis," in *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, 2009, pp. 1444 –1447.

[17] D.R. Hardoon and K. Pasupa, "Image ranking with implicit feedback from eye movements," in *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, New York, NY, USA, 2010, ETRA '10, pp. 291–298, ACM.

[18] 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*, 2009, pp. 1436–1439.

[19] S. Koelstra, A. Yazdani, M. Soleymani, C. Mühl, J-S. Lee, A. Nijholt, T. Pun, T. Ebrahimi, and I. Patras, "Single Trial Classification of EEG and Peripheral Physiological Signals for Recognition of Emotions Induced by Music Videos," in *Brain Informatics*, Yao et al, Ed. Springer, 2010.

[20] I. Arapakis, I. Konstas, and J.M. Jose, "Using facial expressions and peripheral physiological signals as implicit indicators of topical relevance," in *Proceedings of the ACM international conference on Multimedia*, New York, NY, USA, 2009, MM '09, pp. 461–470, ACM.

[21] J. Jiao and M. Pantic, "Implicit image tagging via facial information," in *Proceedings of the 2nd international workshop on Social signal processing*. ACM, 2010, pp. 59–64.

[22] 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, October 2010.

[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 et al., Eds., Computer Communications and Networks, pp. 217–237. Springer London, 2013.

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

[25] L. I. Aftanas, N. V. Reva, A. A. Varlamov, S. V. Pavlov, and V. P. Makhnev, "Analysis of evoked EEG synchronization and desynchronization in conditions of emotional activation in humans: temporal and topographic characteristics.," *Neuroscience and behavioral physiology*, vol. 34, no. 8, pp. 859–867, October 2004.

[26] G. Chanel, J.J. M. Kierkels, M. Soleymani, and T. Pun, "Short-term emotion assessment in a recall paradigm," *International Journal of Human-Computer Studies*, vol. 67, no. 8, pp. 607–627, August 2009.

[27] P. D. Welch, "The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging Over Short, Modified Periodograms," *IEEE Transactions on Audio and Electroacoustics*, vol. 15, pp. 70–73, 1967.

[28] 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.

[29] C. Chang and C. Lin, "LIBSVM: a Library for Support Vector Machines," 2001.

[30] F. Lotte and C. Guan, "Learning from other subjects helps reducing brain-computer interface calibration time," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 614–617.

[31] A. Bashashati, M. Fatourechi, R.K. Ward, and G.E. Birch, "A survey of signal processing algorithms in brain–computer interfaces based on electrical brain signals," *Journal of Neural engineering*, vol. 4, no. 2, pp. R32, 2007.