bioRxiv preprint doi: https://doi.org/10.1101/2021.08.06.455362; this version posted August 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



# Cross-Individual Affective Detection Using EEG Signals with Audio-Visual Embedding

Zhen Liang[1,2,*], Xihao Zhang[1,2,†], Rushuang Zhou[1,2,‡], Li Zhang[1,2,§], Linling Li[1,2,¶], Gan Huang[1,2,‖], and Zhiguo Zhang[1,2,3,4,**]

[1]School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen, Guangdong 518060, China
[2]Guangdong Provincial Key Laboratory of Biomedical Measurements and Ultrasound Imaging, Shenzhen, Guangdong 518060, China
[3]Marshall Laboratory of Biomedical Engineering, Shenzhen, Guangdong 518060, China
[4]Peng Cheng Laboratory, Shenzhen, Guangdong 518055, China

Email: *janezliang@szu.edu.cn, †zhangxihao2019@email.szu.edu.cn, ‡2018222087@szu.edu.cn, §lzhang@szu.edu.cn, ¶lilinling@szu.edu.cn, ‖huanggan@szu.edu.cn, **zgzhang@szu.edu.cn

***Abstract*—EEG signals have been successfully used in affective detection applications, which could directly capture brain dynamics and reflect emotional changes at a high temporal resolution. However, the generalized ability of the model across individuals has not been thoroughly developed yet. An involvement of other data modality, such as audio-visual information which are used for emotion triggering, could be beneficial to estimate intrinsic emotions in video content and solve the individual differences problem. In this paper, we propose a novel deep affective detection model, named as EEG with audio-visual embedding (EEG-AVE), for cross-individual affective detection. Here, EEG signals are exploited to identify the individualized patterns and contribute the individual preferences in affective detection; while audio-visual information is leveraged to estimate the intrinsic emotions involved in the video content and enhance the reliability of the affective detection performance. For EEG-based individual preferences prediction, a multi-scale domain adversarial neural network is developed to explore the shared dynamic, informative, and domain-invariant EEG features across individuals. For video-based intrinsic emotions estimation, a deep audio-visual feature based hypergraph clustering method is proposed to examine the latent relationship between semantic audio-visual features and emotions. Through an embedding model, both estimated individual preferences and intrinsic emotions are incorporated with shared weights and further are used together to contribute to affective detection across individuals. We conduct cross-individual affective detection experiments on MAHNOB-HCI database for model evaluation and comparison. The results show our proposed EEG-AVE model achieves a better performance under a leave-one-individual-out cross-validation individual-independent evaluation protocol, with the accuracies of 90.21% and 85.59% for valence and arousal using aggregated labels and 71.13% and 66.47% for valence and arousal using non-aggregated labels. Hence, EEG-AVE is an effective model with good generalizability, which makes it a power tool for cross-individual emotion detection in real-life applications.**

***Index Terms*—Electroencephalography; Individual Differences; Affective Detection; Audio-Visual Embedding; Deep Domain Adaptation.**

## I. Introduction

ELECTROENCEPHALOGRAPHY (EEG) provides a nature way to record human brain activities and has been widely used in the affective intelligence studies [1]–[5]. In recent years, deep neural network learning methods have provided an effective and efficient approach to characterize informative deep features from EEG data and have achieved promising results in EEG-based affective detection applications. For example, a novel dynamic graph convolutional neural network (DGCNN) was proposed in [1] to learn the discriminant and hidden EEG characteristics in a non-linear approach for solving the multi-channel EEG based emotion decoding problem. Jirayucharoensak *et al*. [6] adopted a stack of several autoencoder structures to perform EEG-based emotion decoding and showed the deep learning network outperformed the traditional classification models such as support vector machine (SVM) and naïve Bayes classifiers. The valid, useful and optimal EEG information can be explored in a deep belief network (DBN) structure, which was demonstrated to be beneficial to the decoding performance [7]. Cui *et al*. [8] proposed an end-to-end regional-asymmetric convolutional neural network (RACNN) to capture the discriminant EEG features covering temporal, regional and asymmetric information. Based on a series of pretrained state-of-the-art CNN architectures, Cimtay and Ekmekcioglu [4] improved the feature extraction performance and classification capability based on raw EEG signals. The existing literature has showed deep learning is a powerful tool in EEG processing, which captures the abstract representations and disentangle the semantic gap between EEG signals and emotion states.

However, due to the problem of individual differences, the stability and generalizability of EEG-based affective detection models are of great challenge. Especially, EEG data are very weak signals and easily susceptible to interference from undesired noises, making it different to distinguish individual-specific and meaningful EEG patterns from noise. The key to solve the individual differences problem is to minimize the discrepancy in feature distributions across individuals. To improve model generalization to the variance of individual characteristics, transfer learning methods have been introduced and a fruitful line of prior studies has been explored [2], [9]–[11]. Based on feature distribution and classifier parameters learning, Zheng and Lu [10] developed two types of subject-to-subject transfer learning approaches and showed a significant increase in emotion recognition accuracy (conventional generic

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.06.455362; this version posted August 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



classifier: 56.73%; the proposed model: 76.31%). Lin and Jung [11] proposed a conditional transfer learning framework to boost a positive transfer for each individual, where the individual transferability was evaluated and effective data from other subjects were leveraged. Li *et al*. [2] developed a multi-source transfer learning method, where two sessions (calibration and subsequent) were involved and the data differences were transformed by the style transfer mapping and integrated classifier. Among various transfer learning strategies, domain adaptation is a popular way to learn common feature representations and make the feature representations invariant across different domains (source and target domains). Ganin *et al*. [12] proposed an effective domain-adversarial neural network (DANN) to align the feature distributions between source domain and target domain and also maintain the information of the aligned discriminant features which are predictive of the labels of source samples. Instead of the conventional domain adaptation methods that adapted a well-trained model based on a specific domain to another domain, DANN could well learn the shareable features from different domains and maintain the common knowledge about the given task. Inspired by this work, Li *et al*. [13] proposed a bi-hemisphere domain adversarial neural network (BiDANN) model for emotion recognition using EEG signals, in which a global and two local domain discriminators worked adversarially with an emotion classifier to improve the model generalizability. Li *et al*. [3] proposed a domain adaptation method through simultaneously adapting marginal and conditional distributions based on the latent representations and demonstrated an improvement of the model generalizability across subjects and sessions.

On the other hand, with the great development and application of the internet and multimedia nowadays, there are many approaches to characterize audio-visual content and embed the conveying information with other feature modalities for emotion detection. For example, based on traditional hand-crafted audio and visual features, Wang *et al*. [14] investigated several kernel based methods to analyze and fuse audio-visual features for bimodal emotion recognition. Mo *et al*. [15] proposed Hilbert-Huang Transform (HHT) based visual and audio features for time-frequency-energy description of videos and introduced cross-correlation features to indicate the dependencies between the visual and audio signals. Furthermore, the recent success of deep learning methods in computer vision brings new insights into video-content based affective study. Acar *et al*. [16] utilized CNNs to learn mid-level audio-visual feature representations for affective analysis of music video clips. Zhang *et al*. [17] proposed a hybrid deep model to characterize a joint audio-visual feature representation for emotion recognition, where CNN, 3D-CNN and DBN were integrated with a two-stage learning strategy.

In general, current affective computing models can be mainly categorized into two streams. One stream is to predict individual preferences through analyzing a user's spontaneous physiological responses (i.e. EEG signals) while watching the videos [18], [19]. Another stream is to estimate intrinsic emotions from video content itself by integrating visual and audio features in an either feature-level fusion or decision fusion and building a classifier for distinguishing emotions [17], [20]. However, spontaneous response based individual preferences prediction would be sensitive to individual differences and fail to achieve a reliable performance in affective detection across individuals; while video content based intrinsic emotions estimation could achieve a stable emotion detection performance, but fail to consider the deviations of individuals in emotion perceiving. An appropriate embedding of individual preferences and intrinsic emotions would be helpful to learn reliable affective features from video content and also individual preferences from the recording spontaneous responses. This motivates us to study the underlying associations among emotions, video content and brain responses, where video content functions as a stimulation clue indicating what kind of emotions would possibly be elicited and brain responses reveals individual emotion perceiving process showing how we exactly feel the emotions.

Besides, comparing to unimodal analysis, multimodal fusion could provide more details, compensate the incomplete information from another modality and develop advanced intelligent affective systems [21]. Recently, Wang *et al*. [22] incorporated video information and EEG signals to improve the video emotion tagging performance. This study characterized a set of traditional visual and audio features, including brightness, color energy and visual excitement for visual features, and average energy, average loudness, spectrum flow, zero crossing rate (ZCR), standard deviation of ZCR, 13 Mel-Frequency Cepstral Coefficients (MFCC) and the corresponding standard deviations for audio features. The proposed hybrid emotion tagging approach was realized on a modified SVM classifier, and the corresponding performance was improved from 54.80% to 75.20% for valence and from 65.10% to 85.00% for arousal after a fusion of multi-modality data. Inspired by the success of the embedding protocol across different data modalities, this study proposes a novel affective information detection model (termed as EEG-AVE) to learn transferable features from EEG signals and embed affective-related multimedia characteristics for better affective detection across individuals. The proposed EEG-AVE model is shown in Fig. 1, which is composed of three parts: EEG-based individual preferences prediction, audio-visual based intrinsic emotions estimation, and multimodal embedding.

(1) **EEG-based individual preferences prediction:** In this part, we propose a multi-scale domain adversarial neural network (termed as MsDANN hereinafter) based on DANN [12] to enhance the generalization ability of EEG feature representation across individuals and boost the model performance on individual preferences prediction. Specifically, EEG data from different individuals are treated as domains, where the source domain refers to the existing individuals and the target domain refers to the newcoming individual(s). Based on the input multi-scale feature representation, the feature extractor network, task classification network and discriminator network are designed to make the source and target domains share similar and close latent distribution so as to work with the same prediction model. As the mining emotional informative and sensitive features from EEG signals is

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.06.455362; this version posted August 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



still a great challenge, this study introduces a multi-scale feature representation to improve feature efficacy and model adaptability to complex and dynamic emotion cases. Comparing to single-scale feature representation, pioneer studies have showed EEG signals analysis with such a coarse-grain procedure could be beneficial to emotion studies [23]–[25].

(2) **Audio-visual based intrinsic emotions estimation:** To alleviate the individual differences problem, affective audio-visual content analysis is conducted to digest the intrinsic emotion information involved the videos which could be used as a supplementary information for individual affective detection. Due to the well-known "semantic gap" or "emotional gap" that the traditional handcrafted features may fail to sufficiently discriminate emotions, we develop a deep audio-visual feature based hypergraph clustering method (termed as DAVFHC) for characterizing high-level audio-visual features that could effectively distinguish emotions. Here, two pretrained CNN architectures (VGGNet [26] and VGGish [27]) are adopted to explore affective-related characteristics in audio-visual content and the most optimal audio-visual features are fused through a hypergraph theory.

(3) **Multimodal embedding:** The final affective detection result is determined by an embedding model where the predicted individual preferences from EEG signals and the estimated intrinsic emotions from audio-visual content are fused at a decision level. The compensation information from different modalities contributes together to tackle the individual differences problems in affective detection.

The major novelties of this work are as follows.

- We propose a novel affective detection model (EEG-AVE) to tackle individual differences problem in EEG-based emotion recognition through incorporating spontaneous brain responses and stimulation clues in a hybrid embedding strategy. Both EEG and audio-visual information are exploited to digest different dimensions of emotions, and the compensation relationships among different data modalities on the affective detection study are examined.
- We introduce an effective domain adaption method (MsDANN) to detect individual preferences from EEG signals, where the impact of individual differences is diminished through a transfer learning approach.
- We present an efficient method (DAVFHC) to estimate the intrinsic emotions in audio-visual materials. For characterizing emotion related features, semantic audio-visual features are extracted by using deep learning methods and the complex and latent relationships of deep audio-visual features with emotion labels are measured with hypergraph theory.

The remainder of this paper is structured as following. In Section II and III, the proposed MsDANN model for individual preferences prediction from EEG signals and the proposed DAVFHC model for intrinsic emotions estimation from audio-visual content are respectively introduced in details. Section IV presents the embedding model for final affective detection from individuals. Section V explains the database and experimental conditions used in this study. In Section VI, the model evaluation and comparison are conducted using a leave-one-individual-out cross-validation individual-independent protocol and the results are discussed. Section VII summarizes our findings and draws a conclusion of this paper.

## II. Individual Preferences Prediction

### A. MsDANN Model

In this section, we propose a new transfer learning based neural network, MsDANN, to address the individual differences problem in EEG based emotion detection. In this network, a multi-scale feature representation is incorporated to capture a series of rich feature characteristics of EEG signals and maximize the informative context for predicting a diverse set of individual preferences in emotions. Specifically, we extract the differential entropy (DE) features [28] from the defined frequency sub-bands (refer to Table I) at different frequency/scale resolutions (1 Hz, 0.5 Hz, and 0.25 Hz), and build respective domain adaptation models with domain adversarial training methods. In the proposed MsDANN, the common features from different individuals are learnt; at the same time, the relationships between the learnt common features and the related emotion information are preserved.

The network structure of MsDANN is shown in Fig. 2, which is composed of three parts: the generator (feature extractor network) for deep feature extraction, the classifier (task classification network) for emotion label prediction, and the discriminator (discriminator network) for real or fake data distinguishing. Here, the generator and classifier could be considered as a standard feed-forward architecture, while the generator and discriminator are trained based on a gradient reversal layer to ensure the feature distributions of two domains as indistinguishable as possible. In this study, the EEG data with emotion labels are treated as the source domain to train the generator, classifier and discriminator; while the EEG data without emotion labels are utilized to train the generator and discriminator. Through this multi-scale deep framework, a set of transferable features involving affective information could be characterized, the cross-domain discrepancy could be bridged, and the classification performance could be effectively improved in both source and target domains.

To learn a shared common feature space between the source and target domains and also guarantee the learnt feature representation involving enough information for revealing the emotion states, the loss objective function is designed below. Suppose that the source and target domains are denoted as $\mathbb{S}$ and $\mathbb{T}$. In the domain learning, the EEG data with emotion labels in $\mathbb{S}$ are given as $x^l = \{x_1^l, ..., x_{N_S}^l\}$ and $y = \{y_1, ..., y_{N_S}\}$, where $x_i^l$ is the input EEG data at $l$th scale feature representation and $y_i$ is the corresponding emotion label of $x_i^l$. $N_S$ is the sample size of $x^l$. On the other hand, the unlabeled EEG data in $\mathbb{T}$ is denoted as $z^l = \{z_1^l, ..., z_{N_T}^l\}$, where $z_i^l$ is the input EEG data at $l$th scale feature representation and $N_T$ is the corresponding sample size of $z^l$. We denote the generator, classifier, and discriminator as $r_\theta$, $c_\sigma$, $d_\mu$ with the parameters of $\theta$, $\sigma$ and $\mu$. To ensure the learnt features by

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.06.455362; this version posted August 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



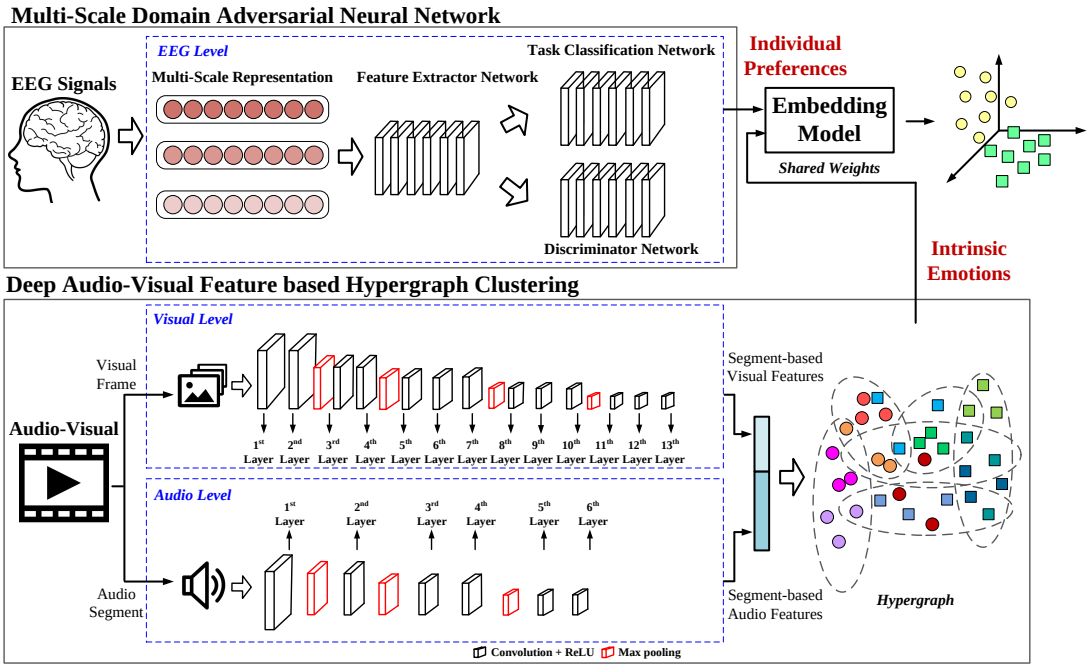


Fig. 1: The proposed EEG-AVE model.

TABLE I: The defined frequency sub-bands for DE feature characterization.

| | $\theta$ | $\alpha 1$ | $\alpha 2$ | $\beta 1$ | $\beta 2$ | $\beta 3$ | $\gamma 1$ | $\gamma 2$ | $\gamma 3$ |
|---|---|---|---|---|---|---|---|---|---|
| frequency band (Hz) | 4-8 | 8-10 | 10-13 | 13-16 | 16-20 | 20-28 | 28-34 | 34-39 | 39-45 |

$r_\theta$ from source domain or target domain are indistinguishable, the domain adversarial training objective function is given as

$$\min_{\theta} \max_{\mu} \mathbf{E}_{(x^l,z^l)\sim(\mathbb{S},\mathbb{T})} \mathcal{L}_D(\mu,\theta,x^l) + \mathcal{L}_D(\mu,\theta,z^l), \quad (1)$$

where $\mathcal{L}_D$ is a binary cross-entropy loss for the discriminator to be trained to distinguish $\mathbb{S}$ and $\mathbb{T}$, defined as

$$\begin{aligned}\mathcal{L}_D(\mu,\theta,x^l) = &-\mathbb{I}[x^l \sim \mathbb{S}] \log(d_\mu \circ r_\theta(x^l)) \\ &- \mathbb{I}[x^l \sim \mathbb{T}] \log(1 - d_\mu \circ r_\theta(x^l)).\end{aligned} \quad (2)$$

Here, $\mathbb{I}$ is an indicator function. Based on Eq. 1, we add another loss function $\mathcal{L}_T$ for the classifier part as

$$\begin{aligned}&\min_{\sigma,\theta} \max_{\mu} \mathbf{E}_{x^l\sim\mathbb{S}}[\mathcal{L}_T(\sigma,\theta,x^l)] + \\ &\lambda \mathbf{E}_{(x^l,z^l)\sim(\mathbb{S},\mathbb{T})}[\mathcal{L}_D(\mu,\theta,x^l) + \mathcal{L}_D(\mu,\theta,z^l)],\end{aligned} \quad (3)$$

where $\mathcal{L}_T(\sigma,\theta,x^l)$ is the classification loss in the source domain, determined by $\sum Loss(c_\sigma \circ r_\theta(x^l), y)$. $\lambda$ is a balance parameter during the learning process, given as

$$\lambda = \frac{2}{1 - \exp(-\gamma p)} - 1, \quad (4)$$

where $\gamma$ is a constant value and $p$ is a factor of epoch. Eq. 3 is the final objective function for MsDANN model training.

The proposed MsDANN model is an end-to-end framework for cross-individual emotion prediction based on EEG signals, combining the feature learning adaptation and emotion classification into a unified deep model. Based on the input data with multi-scale DE feature representation, the domain adaption and classification loss are exploited to guide the generator to learn effective feature representations across individuals via the gradient reversal layer and efficiently tackle the individual differences problem in EEG data processing.

### B. MsNN Model

We also introduce a baseline method, multi-scale neural network (termed as MsNN), for model comparison under the condition with or without deep domain adaption. As illustrated in Fig. 3, the network is trained on the source domain without feature adaption and tested on the target domain.

## III. Intrinsic Emotions Estimation

At present, a number of well trained deep CNN models have been successfully applied to multimedia processing, such as AlexNet [29], GoogLeNet [30] and VGG [26] for visual content, and VGGish [27] for audio content. The deep features could bridge the semantic gap and improve semantic interpretation performance. In this section, we develop a DAVFHC method to learn and decode the semantic features from audio-visual content for intrinsic emotions estimation.

### A. Deep Visual Feature Characterization

At the visual level, a pretrained VGGNet network [26] is utilized to process frame-based visual information and characterize effective visual features. The training and testing data

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.06.455362; this version posted August 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



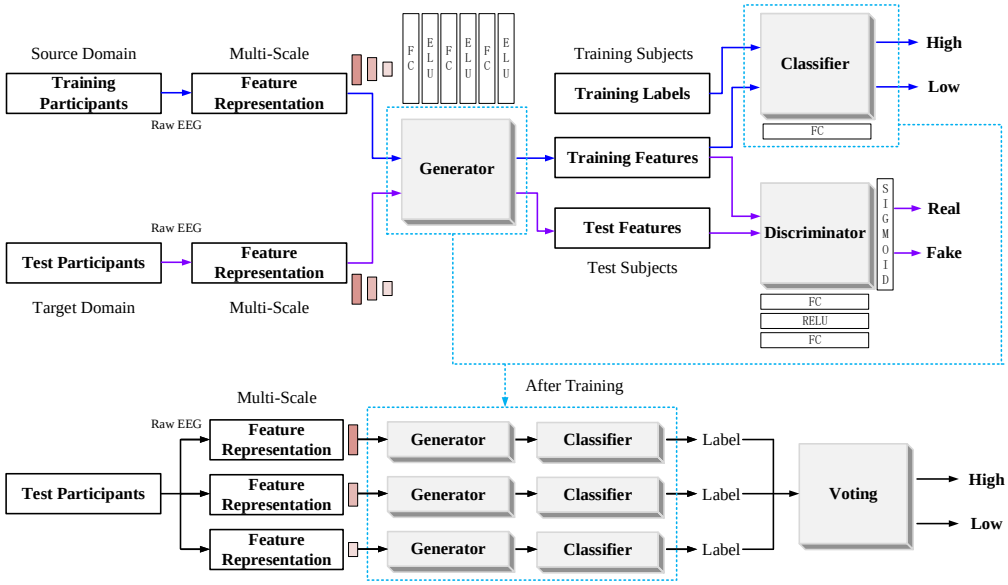


Fig. 2: The proposed MsDANN model.

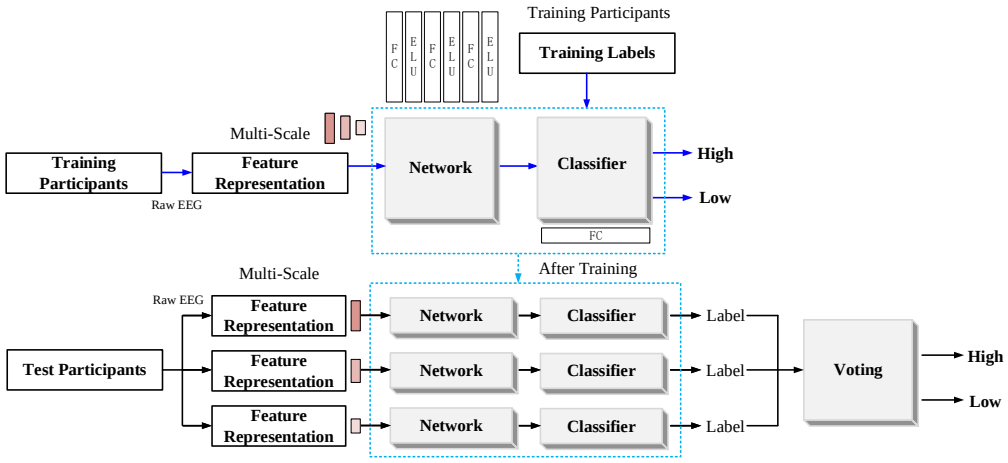


Fig. 3: The baseline method of MsNN model.

sets were based on ILSVRC-2012, with 1.3M training pictures, 50K test pictures, and 100K validation pictures. The network was trained by optimizing a polynomial logistic regression objective function with a smallest batch-based gradient descent momentum. Considering the balance of layer depth and performance, VGG16 is utilized in this paper to characterize the frame-based visual features. It consists of 13 convolutional layers and 3 fully connected layers. The corresponding number of convolution kernels at each layer are 64, 64, 128, 128, 256, 256, 256, 512, 512, 512, 512, 512, and 512, and the kernel size is $3 \times 3$.

As illustrated in Fig. 4, the visual feature extraction procedure includes three steps. **1. Frame-based visual feature extraction.** The video frames are input to the pretrained VGG16 and the corresponding feature maps are characterized at each convolutional layer. For each layer, an average feature map is then calculated and converted into a feature vector. **2. Segment-based visual feature extraction.** Instead of direct averaging all the frame-based features in one segment, we introduce an adaptive key frame detection step to detect a key frame from every segment based on the feature distribution. Suppose that one segment is composed of $k$ frames with the corresponding extracted features, denoted as $B^\iota = \{b_1^\iota, ..., b_k^\iota\}$, where $\iota = 1, ..., N_\iota$ refers to the convolutional layer. The key frame detection is illustrated as follows. (1) All frames are grouped into one cluster in terms of $B^\iota$; (2) The cluster center $c^\iota$ is computed; (3) The distance between each frame $b_i^\iota$ $(i \in [1, k])$ and the cluster center $c^\iota$ is calculated, denoted as $\{d_1^\iota, ..., d_k^\iota\}$; (4) the frame which is the closest to $c^\iota$ is selected as the key frame of the segment, termed as $k^* = \arg\min\{d_1^\iota, ..., d_k^\iota\}$. Then, the corresponding feature of the key frame $b_{k^*}^\iota$ is treated as the segment-based feature

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.06.455362; this version posted August 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



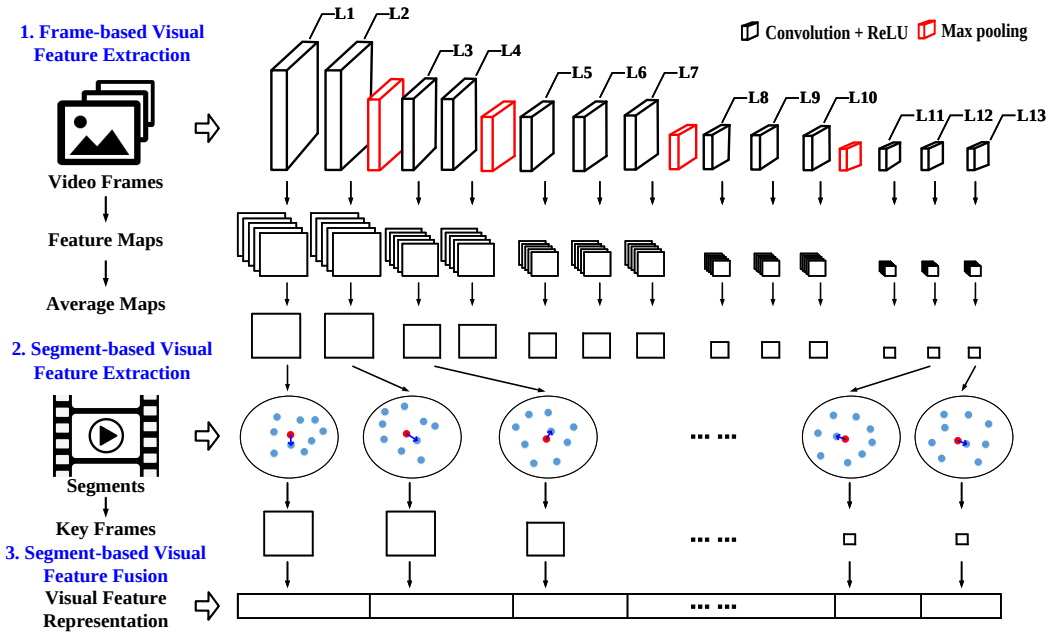


Fig. 4: The visual feature extraction procedure.

representation. **3. Segment-based visual feature fusion.** The characterized segment-based features at each single convolutional layer ($b^{\iota}_{k^*}, \iota \in [1, N_\iota]$) are then fused by concatenation. Empirically, the segment length is set to 1s. To get the semantic features, only the characterized features at the last two convolutional layers ($\iota = 12$ and 13) are used as visual features ($\Psi_V$) in the proposed DAVFHC method.

### B. Deep Audio Feature Characterization

At the audio level, a pretrained CNN network, VGGish [27], is adopted to characterize effective audio features. VGGish is a deep network model trained on a Youtube-8M database (training/validation/test: 70M/20M/10M), which has been proved to be capable of extracting effective and efficient deep auditory features in various applications [31]–[33]. The network contains 6 convolutional layers, and the corresponding numbers of convolution kernels are 64, 128, 256, 256, 512, and 512, respectively. The kernel size is $3 \times 3$.

Same as the visual feature extraction process, the audio features are also characterized at the segment level. **1. Data preparation.** The audio signals are detected from the emotional clips and then partitioned into a number of segments with a fixed length. **2. Data preprocessing.** The segment-based audio data is preprocessed following the procedures presented in [27]. **3. Deep audio feature characterization.** For each segment, the logarithmic melspectrum is characterized and input to the VGGish. The deep feature maps are extracted at each convolutional layer and averaged into one feature map. **4. Deep audio feature fusion.** For each segment, the feature map at each single convolutional layer is converted into a feature vector. The converted feature vectors across different convolutional layers are then fused by concatenation. Empirically, the segment length is set to 1s (same as the visual data. To get the semantic features, only the feature vectors extracted from last two layers (5th and 6th) are used as audio features ($\Psi_A$) in the proposed DAVFHC method.

### C. Feature Fusion and Emotion Estimation

The characterized segment-based visual and audio features are concatenated and formed into a segment-based audio-visual feature vector termed as $\Psi_M = [\Psi_V, \Psi_A]$. The complex relationships among all the segments from the emotional clips are constructed with a hypergraph which has been widely recognized as an effective approach for complex hidden data structure description. For the traditional graph, only pairwise relationships between any two vertices are considered, which would lead to the information loss [34]. In the hypergraph, one edge (termed as hyperedge in the hypergraph) could connect more than two vertices and the complex relationship among a group of vertices could be well described. In the paper, the segments are the vertices denoted as $V$, and the connections among the segments are the hyperedges denoted as $E$. One hypergraph could be represented as $G = (V, E)$, where the vertices and hyperedges are denoted as $V = \{v_1, v_2, ..., v_{|V|}\}$ and $E = \{e_1, e_2, ..., e_{|E|}\}$, respectively. The vertices belong to one hyperedge $e_k \in E$ is termed as $\{v_1^{e_k}, v_2^{e_k}, ..., v_{|e_k|}^{e_k}\}$. To define the vertices and hyerpedges relationships, the similarity between any two vertices (the emotional clip segments denoted as $\Psi_M^{v_i} = \{\psi_{M,1}^{v_i}, ..., \psi_{M,N_M}^{v_i}\}$ and $\Psi_M^{v_j} = \{\psi_{M,1}^{v_j}, ..., \psi_{M,N_M}^{v_j}\}$, with the feature dimensionality of $N_M$) are measured as

$$a(\Psi_M^{v_i}, \Psi_M^{v_j}) = \frac{1}{1 + \xi_{\Psi_M^{v_i}, \Psi_M^{v_j}}}, \tag{5}$$

where $\xi_{\Psi_M^{v_i}, \Psi_M^{v_j}}$ is the calculated distance, given as

$$\xi_{\Psi_M^{v_i}, \Psi_M^{v_j}} = \sum_{t=1,...,N_M} \frac{(\psi_{M,t}^{v_i} - \psi_{M,t}^{v_j})^2}{\psi_{M,t}^{v_i} + \psi_{M,t}^{v_j}}. \tag{6}$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.06.455362; this version posted August 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



Based on the measured similarity matrix $A = \{a(\Psi_M^{v_i}, \Psi_M^{v_j})\}_{i,j=1}^N$ ($N$ is the sample size), an incident matrix $H \in |V| \times |E|$ is formed, in which the connection relationships between the vertices $V$ and the hyperedges $E$ is described as

$$h(v_i, e_k) = \begin{cases} 1 & \text{if } v_i \in e_k \\ 0 & \text{if } v_i \notin e_k \end{cases}. \tag{7}$$

The hyperedge weight matrix $W$ is a diagonal matrix indicating the weights of all the hyperedges $E$ in the hypergraph $G$. The weight $w(e_k)$ of one hyperedge $e_k \in E$ is computed based on the calculated similarities among the vertices that belong to $e_k$, given as

$$w(e_k) = \frac{\sum_{v_i, v_j \in e_k, v_i \neq v_j} a(\Psi_M^{v_i}, \Psi_M^{v_j})}{\tau}, \tag{8}$$

where $a(\Psi_M^{v_i}, \Psi_M^{v_j})$ is the similarity value between the vertices of $v_i$ and $v_j$, given in Eq. 5. $\tau$ is the total number of vertices connected to the hyperedge $e_k$. As $w(e_k)$ is a measurement of all the similarity relationships among the vertices that belong to one hyperedge, a higher $w(e_k)$ value indicates a strong connection of homogeneous vertices of the hyperedge and a lower $w(e_k)$ refers to a weak connection of the hyperedge in which the connected vertices share little similar properties. In other words, the hypergraph structure could well describe the relationships of the audio-visual segments in terms of properties. The vertex degree matrix ($D_v$) is a diagonal matrix presenting the degree of all the vertices in the hypergraph $G$. The degree of one vertex $v_k \in V$ is calculated as the summation of all the hyperedge weights of the hyperedges that the vertex belong to, defined as

$$d(v_i) = \sum_{e \in E | v_i \in e} h(v_i, e) w(e). \tag{9}$$

The hyperedge degree matrix ($D_e$) is also a diagonal matrix showing the degree of all the hyperedges in the hypergraph $G$. The degree of one hyperedge $e_k \in E$ is calculated as the summation of all the vertices that connect to the hyperedge, given as

$$d(e_k) = \sum_{e_k \in E | v \in e_k} h(v, e_k). \tag{10}$$

In this study, we introduce a spectral hypergraph partitioning method [35] to partition the constructed hypergraph into a number clusters corresponding to the emotion states (high or low). Thus, it is a two-way hypergraph partitioning problem which could be described as

$$Hcut(S, \bar{S}) = \sum_{e \in \partial S} w(e) \frac{|e \cap S||e \cap \bar{S}|}{d(e)}, \tag{11}$$

where $S$ and $\bar{S}$ are the partitions of the vertices $V$. For two-way partitioning, $\bar{S}$ is the complement of $S$. $\partial S$ is the partition boundary, given as $\partial S = \{e \in E | e \cap S \neq \varnothing \text{ and } e \cap \bar{S} \neq \varnothing\}$. $d(e)$ is the hyperedge degree defined in Eq. 10. To avoid unbalanced partitioning, $Hcut(S, \bar{S})$ is further normalized by

$$NHcut(S, \bar{S}) = Hcut(S, \bar{S})(\frac{1}{vol(S)} + \frac{1}{vol(\bar{S})}), \tag{12}$$

where $vol(S)$ and $vol(\bar{S})$ are the volumes of $S$ and $\bar{S}$, given as $vol(S) = \sum_{v \in S} d(v)$ and $vol(\bar{S}) = \sum_{v \in \bar{S}} d(v)$. The partitioning rule is to look for the weakest hyperedge $e$ between $S$ and $\bar{S}$, where the vertices in the same cluster should be tightly connected (high hyperedge weights) and the vertices in the different clusters should be weakly connected (low hyperedge weights). An optimal partitioning is given in Eq. 13 to find the weakest connection between two partitions, which is an NP-complete problem solved by a real-valued optimization method.

$$\begin{aligned} &\arg\min_f \frac{1}{2} \sum_{e \in E} \sum_{u,v \in V} \frac{w(e)h(u,e)h(v,e)}{d(e)} \\ &\times (\frac{f(u)}{\sqrt{d(u)}} - \frac{f(v)}{\sqrt{d(v)}})^2 \\ =&\arg\min_f \sum_{e \in E} \sum_{u,v \in V} \frac{w(e)h(u,e)h(v,e)}{d(e)} \\ &\times (\frac{f^2(u)}{\sqrt{d(u)}} - \frac{f(u)f(v)}{\sqrt{d(u)d(v)}}) \\ =&\arg\min_f \sum_{u \in V} f^2(u) \sum_{e \in E} \frac{w(e)h(u,e)}{d(u)} \sum_{v \in V} \frac{h(v,e)}{d(e)} \\ &- \sum_{e \in E} \sum_{u,v \in V} \frac{f(u)h(u,e)w(e)h(v,e)f(v)}{\sqrt{d(u)d(v)}d(e)} \\ =&\arg\min_f f^T(I - \Theta)f \end{aligned} \tag{13}$$

where $\Theta$ is given as

$$\Theta = D_v^{-(1/2)} H W D_e^{-1} H^T D_v^{-(1/2)}, \tag{14}$$

and $I$ is an identity matrix with the same size as $W$. The hypergraph Laplacian is denoted as

$$\Delta = I - \Theta. \tag{15}$$

The optimal solution is transformed to find the eigenvectors of $\Delta$ whose eigenvalues are the smallest. In other words, the optimal hypergraph partitioning results find the top eigenvectors with the smallest non-zeros eigenvalues in $\Delta$ and form an eigenspace for the subsequent vertex clustering with the K-means method. Through this approach, all the vertices are grouped into two clusters. The corresponding emotion state of each cluster is determined by the majority distribution of the involved vertices. If most of vertices are belong to high level, the cluster's emotion state is assigned as high; on the other hand, it is assigned as low. In practice, to avoid information leaking, the clusters' emotion states are only determined based on the training samples.

## IV. Embedding Model

Based on the aforementioned work, we incorporate the estimated intrinsic emotions based on deep audio-visual features and the predicted individual preferences from the collected simultaneous EEG signals, and conduct a decision-level information fusion for final affective prediction.

Specifically, we fuse EEG signals and audio-visual information in a decision level through shared weights. Suppose that

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.06.455362; this version posted August 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



the predicted emotional individual preferences based on EEG signals are denoted as $Y^{EEG} = \{y_1^{EEG}, ..., y_N^{EEG}\}$ and the estimated intrinsic emotions based on audio-visual content are denoted as $Y^{MUL} = \{y_1^{MUL}, ..., y_N^{MUL}\}$. The final detected affective results are determined by

$$y_i^{FUS} = \frac{w^{EEG} \times y_i^{EEG} + w^{MUL} \times y_i^{MUL}}{w^{EEG} + w^{MUL}}, \quad (16)$$

where $w^{EEG}$ and $w^{MUL}$ are the shared weights of EEG signals and audio-visual information in the fusion process. $Y^{FUS} = \{y_1^{FUS}, ..., y_N^{FUS}\}$ are the final affective detection results.

## V. Database and Experimental Conditions

A famous EEG-emotion database, the MAHNOB-HCI [36], is used for sentiment analysis and affective detection from both EEG signals and audio-visual content in this paper. In this database, a total of 30 participants (male/female: 13/17; age: 26.06±4.39) from different cultural backgrounds were invited to participate in the experiment and 20 commercial film clips (duration: from 34.9s to 117s, with an average of 81.4s and a standard deviation of 22.5s) were selected for emotional eliciting. The experimental paradigm is shown in Fig. 5. In one trial, to minimize the emotion evoking effect from other trials, a neutral clip that was randomly selected from the Stanford Psychophysiology Laboratory played first. Then, an emotional clip was randomly selected from the 20 commercial film clips and played for emotion eliciting. After the emotional clip playing, the participants were requested to give a subjective assessment about their emotions during watching the emotional clip using a score in the range of 1 to 9. During the experiment, EEG signals were simultaneously collected at a sampling rate of 256Hz, by using the Biosemi active II system with 32 Ag/AgCl electrodes placed according to the standard international 10-20 electrode system [37]. Due to the data incompleteness of participant 3, 9, 12, 15, 16 and 26, only 24 participants are used in this paper.

To cross-compare with the results presented in the other studies, we utilize a fixed threshold of 5 for scores (in the range of 1 to 9) to discretize the subjective feedback into high and low levels ($\geq 5$ high; $< 5$ low). Two performance metrics, detection accuracy $P_{acc}$ and F1-Score $P_f$, are used to validate the evaluation performance. $P_{acc}$ is an overall detection performance measurement and $P_f$ is a harmonic average of the precision and sensitivity which is less susceptible to the unbalanced classification problems. The corresponding definitions are given as

$$P_{acc} = \frac{n_{TN} + n_{TP}}{n_{TN} + n_{FN} + n_{TP} + n_{FP}} \times 100\%, \quad (17)$$

and

$$P_f = \frac{2 \times P_{pre} \times P_{sen}}{P_{pre} + P_{sen}} \times 100\%, \quad (18)$$

where $n_{TN}$ and $n_{TP}$ are the correctly predicted samples, and $n_{FN}$ and $n_{FP}$ are the incorrectly predicted samples. The precision $P_{pre}$ and sensitivity $P_{sen}$ are given as

$$P_{pre} = \frac{n_{TP}}{n_{TP} + n_{FP}}, \quad (19)$$

$$P_{sen} = \frac{n_{TP}}{n_{TP} + n_{FN}}. \quad (20)$$

To fully evaluate the validity and reliability of the model performance, a strict leave-one-out cross-validation is adopted. All the predicted individual preferences and the estimated intrinsic emotions are obtained in a cross-validation manner. For the proposed MsDANN model, the model training and testing are conducted on a leave-one-individual-out cross-validation. In one round of cross-validation, all the samples from 1 individual are treated as the test data, while the other samples from the remaining individuals (other 23 participants) are used as the training data. After repeating 24 times until each participant is treated as the test data once, the final result of MsDANN is a formation of all the obtained test results through the cross-validation rounds. For the developed DAVFHC method, the model training and testing are conducted on a leave-one-video-out cross-validation. In one round of cross-validation, all the samples from 1 video are used as test data and the other samples from the remaining videos (other 19 videos) are treated as training data. After repeating 20 times until each video is treated as the test data once, the final prediction result of DAVFHC is a formation of the obtained test results in all the cross-validation rounds. In other words, after obtaining all the test results of all EEG and video samples in the above mentioned cross-validation rounds, the final affective results are obtained by a decision fusion. As the affective detection on different emotion dimensions such as valence and arousal are independent, the corresponding hyperedge size and feature size in DAVFHC are set to different values empirically, where the values are set to 15 and 20 for valence and 21 and 25 for arousal.

## VI. Results and Discussion

### A. Performance Comparison with Existing Methods

To improve the affective detection performance, both EEG signals and audio-visual information are embedded in the proposed EEG-AVE model. Here, we roughly estimate what kind of emotion could be triggered according to the audio-visual content itself (intrinsic emotions estimation), and detect the individual preferences for each individual through analyzing the recording EEG signals while he / she is watching the multimedia material (individual preferences detection). The contributions of EEG signals and audio-visual information through the affective detection process are considered equally important (the values of $w^{EEG}$ and $w^{MUL}$ are set to 0.5). The corresponding emotion decoding performance for valence and arousal are reported in Table II and Table III, respectively. We compare EEG-AVE model with the existing representative methods such as [36], [38], [39], [40], [41], [42], [43], and [22]. It is worth note that the experimental results presented in [22] were evaluated with the aggregated groundtruth (different participants watching one video are tagged with same emotion label). For the non-aggregated groundtruth, different participants watching one video are tagged with different emotion labels according to the corresponding subjective assessment. Even for a same emotional clip, different participants would have different emotional feelings due to the differences in

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.06.455362; this version posted August 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



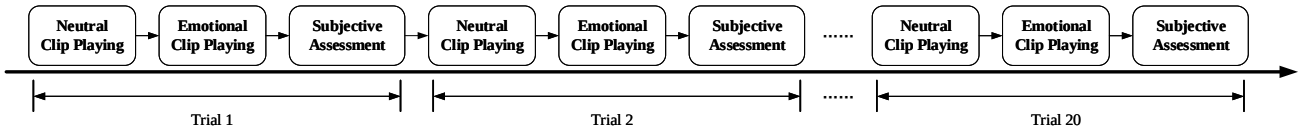


Fig. 5: The experimental paradigm of MAHNOB-HCI database.

TABLE II: Emotion recognition performance on MAHNOB-HCI database among various methods for valence.

| Methods | Groundtruth | $P_{acc}$ | $P_f$ |
|---|---|---|---|
| Soleymani *et al.* [36] | Non-Aggregated | 57.00 | 56.00 |
| Zhu *et al.* [38] | Non-Aggregated | 58.16 | 56.36 |
| Huang *et al.* [39] | Non-Aggregated | 62.13 | - |
| Rayatdoost and Soleymani [40] | Non-Aggregated | 71.25 | 62.08 |
| Huang *et al.* [41] | Non-Aggregated | 66.25 | - |
| Yin *et al.* [42] | Non-Aggregated | 69.93 | - |
| Zhang and Yin [43] | Non-Aggregated | 62.99 | 62.56 |
| Wang *et al.* [22] | Aggregated | 75.20 | 73.80 |
| **Proposed EEG-AVE model** | Aggregated | **90.21** | **90.45** |
| **Proposed EEG-AVE model** | Non-Aggregated | 71.13 | **66.83** |

TABLE III: Emotion recognition performance on MAHNOB-HCI database among various methods for arousal.

| Methods | Groundtruth | $P_{acc}$ | $P_f$ |
|---|---|---|---|
| Soleymani *et al.* [36] | Non-Aggregated | 52.40 | 42.00 |
| Zhu *et al.* [38] | Non-Aggregated | 61.35 | 63.08 |
| Huang *et al.* [39] | Non-Aggregated | 61.80 | - |
| Rayatdoost and Soleymani [40] | Non-Aggregated | 61.46 | 50.60 |
| Huang *et al.* [41] | Non-Aggregated | 71.88 | - |
| Yin *et al.* [42] | Non-Aggregated | 67.43 | - |
| Zhang and Yin [43] | Non-Aggregated | 65.21 | 61.82 |
| Wang *et al.* [22] | Aggregated | 85.00 | 82.40 |
| **Proposed EEG-AVE model** | Aggregated | **85.59** | **86.55** |
| **Proposed EEG-AVE model** | Non-Aggregated | 66.47 | **63.25** |

background, experience, religion, education, and so on. The non-aggregated groundtruth could be more capable of reflecting the emotion dynamics in individuals.

As shown in Table II, our proposed model outperforms the existing methods for valence, where the $P_{acc}$ and $P_f$ results are 90.21% and 90.45% for the aggregated groundtruth and 71.13% and 66.83% for non-aggregated groundtruth. For the results with non-aggregated groundtruth, even the obtained $P_{acc}$ values of our proposed EEG-AVE model and Rayatdoost and Soleymani [40]'s work are comparable, a better $P_f$ of our proposed EEG-AVE model is observed, where $P_f$ is 62.08% for Rayatdoost and Soleymani [40]'s work and 66.83% for our model (improved by 7.65%). For the results with aggregated groundtruth, our proposed EEG-AVE model increases the affective detection performance by 19.96% for $P_{acc}$ and 22.56% for $P_f$, comparing to Wang *et al.* [22]'s work. Similar promising emotion recognition performance are observed for arousal as shown in Table III, where the $P_{acc}$ and $P_f$ results are 85.59% and 86.55% for the aggregated groundtruth and 66.47% and 63.25% for non-aggregated groundtruth. For aggregated groundtruth, the proposed EEG-AVE model performs better than Wang *et al.* [22], especially for F1-score (increased by 5%). For non-aggregated groundtruth, EEG-AVE model gains better performance than most of existing methods on recognition accuracy, except for Huang *et al.* [41] and Yin *et al.* [42]. The obtained F1-score of EEG-AVE is the highest among all the methods. Besides, the above results show aggregated groundtruth leads a higher detection performance comparing to the non-aggregated groundtruth, as the individual differences in emotion feelings about the clip is not considered.

TABLE IV: Emotion recognition performance with different embedding strategies using aggregated and non-aggregated groundtruth for valence.

| Groundtruth | EEG+Visual+Audio | | EEG+Visual | | EEG+Audio | |
|---|---|---|---|---|---|---|
| | $P_{acc}$ | $P_f$ | $P_{acc}$ | $P_f$ | $P_{acc}$ | $P_f$ |
| Aggregated | **90.21** | **90.45** | 74.65 | 74.61 | 69.08 | 73.06 |
| Non-Aggregated | **71.13** | **66.83** | 67.75 | 61.79 | 58.57 | 58.27 |

TABLE V: Emotion recognition performance with different embedding strategies using aggregated and non-aggregated groundtruth for arousal.

| Groundtruth | EEG+Visual+Audio | | EEG+Visual | | EEG+Audio | |
|---|---|---|---|---|---|---|
| | $P_{acc}$ | $P_f$ | $P_{acc}$ | $P_f$ | $P_{acc}$ | $P_f$ |
| Aggregated | **85.59** | **86.55** | 77.28 | 78.57 | 68.55 | 72.20 |
| Non-Aggregated | **66.47** | **63.25** | 63.26 | 59.27 | 54.91 | 53.64 |

### *B. Performance Evaluation of Embedding Strategy*

We compare the affective detection performance when different embedding strategies are adopted. Here are three embedding strategies: EEG+Visual+Audio (the proposed EEG-AVE model), EEG+Visual (only visual information embedded with EEG signals) and EEG+Audio (only audio information embedded with EEG signals). The corresponding affective detection performances for valence and arousal with aggregated and non-aggregated groundtruth are summarized in Table IV and Table V.

The results show that EEG based affective detection with an embedding of both visual and audio information achieve the best performance for both valence and arousal. For EEG+Visual strategy, the affective detection performance for valence decreases to 74.65% (aggregated) and 67.75% (non-aggregated) for $P_{acc}$ and 74.61% (aggregated) and 61.79% (non-aggregated) for $P_f$; while the affective detection performance for arousal decreases to 77.28% (aggregated) and 63.26% (non-aggregated) for $P_{acc}$ and 78.57% (aggregated) and 59.27% (non-aggregated) for $P_f$. The average decrease

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.06.455362; this version posted August 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



rates of valence and arousal are 11.76% and 7.51%, respectively. For EEG+Audio embedding strategy, the affective detection performance for valence decreases from 90.21% to 69.08% for $P_{acc}$ and from 90.45% to 73.06% for $P_f$ when aggregated groundtruth is utilized; while it decreases from 71.13% to 58.57% for $P_{acc}$ and from 66.83% to 58.27% for $P_f$ when non-aggregated groundtruth is used. A similar decrease pattern is also observed on the affective detection performance for arousal, where it decreases from 85.59% to 68.55% for $P_{acc}$ and from 86.55% to 72.20% for $P_f$ when aggregated groundtruth is adopted; while it decreases from 66.47% to 54.91% for $P_{acc}$ and from 63.25% to 53.64% for $P_f$ when non-aggregated groundtruth is utilized. The average decrease rates of valence and arousal are 18.28% and 17.27%, respectively. The comparison results with different embedding strategies reveal that an embedding of both visual and audio information has a capability to reach better affective detection performance, compared to only visual or audio embedded. In addition, we find only visual embedded outperforms only audio embedded, which suggests that visual information plays a more critical role on emotion perceiving, especially in film clips.

### C. Performance Evaluation of Domain Adaptation Effect

To analyze the domain adaptation effect in solving the individual differences problem, we evaluate our proposed EEG-AVE model with a baseline method MsNN (as illustrated in Fig. 3). In MsNN based EEG-AVE model, no feature adaption or transfer learning is adopted in EEG processing, and the EEG based emotional individual preferences prediction is trained and tested on source domain and target domain separately.

The corresponding affective detection performance of MsDANN and MsNN based EEG-AVE model for valence is reported in Table VI. For aggregated groundtruth, comparing to MsDANN based EEG-AVE model under EEG+Visual+Audio strategy, the detection performance of MsNN based EEG-AVE model decreases by 9.10% and 7.19% in terms of $P_{acc}$ and $P_f$, respectively. Comparing MsDANN and MsNN based EEG-AVE model performance under EEG+Visual and EEG+Audio embedding strategies, both $P_{acc}$ and $P_f$ values also have similar decrease patterns. The $P_{acc}$ value decreases from 74.65% to 70.04% for EEG+Visual, and from 69.08% to 65.33% for EEG+Audio. The $P_f$ value declines from 74.61% to 73.03% for EEG+Visual, and from 73.06% to 71.64% for EEG+Audio. When non-aggregated groundtruth is used, cross-comparing MsDANN and MsNN based model performance under an embedding strategy of EEG+Visual+Audio, it is found that the decoding performance significantly decreases from 71.13% to 63.58% (decreased by 10.61%) in terms of $P_{acc}$ and from 66.83% to 62.35% (decreased by 6.7%) in terms of $P_f$. Similar trends are also observed in the other embedding strategies. The corresponding detection accuracies decrease to 60.98% ($P_{acc}$) and 59.20% ($P_f$) for EEG+Visual embedding strategy, and to 53.38% ($P_{acc}$) and 56.34% ($P_f$) for EEG+Audio embedding strategy.

The corresponding affective detection performance of MsDANN and MsNN based EEG-AVE model for arousal is summarized in Table VII. For aggregated groundtruth, MsDANN based EEG-AVE model outperforms MsNN based EEG-AVE model across all three different embedding strategies in terms of both $P_{acc}$ and $P_f$. For EEG+Visual+Audio, EEG+Visual and EEG+Audio embedding strategies, the corresponding improvement rates from MsNN to MsDANN are 9.18%, 7.30% and 4.93% for $P_{acc}$, and that are 6.30%, 3.71% and 1.08% for $P_f$. For non-aggregated groundtruth, similar patterns are observed. Better results are achieved when MsDANN based EEG-AVE model is adopted. Here, the improvement rates for three different embedding strategies are 14.60%, 14.73% and 10.73% for $P_{acc}$ and 8.73%, 7.90%, and 2.72% for $P_f$.

The above results show, comparing to MsNN, MsDANN is much more powerful in the proposed EEG-AVE model to deal with the individual differences problem in EEG signal processing. It provides a reliable and useful way to adaptively learn the shared emotion-related common and discriminant feature representation across individuals, and demonstrates the validity of domain adaptation method in EEG-based affective detection applications.

### D. Performance Evaluation of Multimedia Affective Representation

In this study, audio-visual information is represented by deep features characterized from two pretrained networks. We further verify the effectiveness of the deep feature representation and compare with the performance using more traditional handcrafted features. Inspired from the previous video affective studies [22], [44]–[46], the commonly used handcrafted features are extracted and compared here. For visual information representation, the adopted handcrafted features include lighting key features, color information and shadow portions in the HSL and HSV spaces. For audio information representation, the used traditional audio features include energy, loudness, spectrum flux, zero crossing rate (ZCR), Mel-frequency cepstral coefficients (MFCCs), log energy, and the standard deviations of the above ZCR, MFCC and log energy. The affective analysis of multimedia content with different feature representations is conducted and the corresponding comparison results of valence and arousal are summarized in Table VIII and Table IX. The results show, for both valence and arousal, a significant improvement on affective detection performance is obtained when deep feature representation is used instead of handcrafted features. It reveals that, comparing to the traditional handcrafted feature representation, deep feature representation is a better and richer affective representation for understanding and perceiving the multimedia content.

## VII. Conclusion

In this paper, we propose a novel affective detection model (EEG-AVE) with an embedding protocol, where both EEG based emotional individual preferences and audio-visual based intrinsic emotions are incorporated to tackle the individual differences problem in EEG processing. The multimodal information are analyzed and compensated to realize efficient and effective EEG-based affective detection. The experimental results show that the proposed EEG-AVE model achieves

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.06.455362; this version posted August 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



TABLE VI: Emotion recognition performance of MsDANN and MsNN based EEG-AVE model with different embedding strategies using aggregated and non-aggregated groundtruth for valence.

| Groundtruth | EEG Model | EEG+Visual+Audio | | EEG+Visual | | EEG+Audio | |
|---|---|---|---|---|---|---|---|
| | | $P_{acc}$ | $P_f$ | $P_{acc}$ | $P_f$ | $P_{acc}$ | $P_f$ |
| Aggregated | MsDANN | **90.21** | **90.45** | **74.65** | **74.61** | **69.08** | **73.06** |
| | MsNN | 82.00 | 83.95 | 70.04 | 73.03 | 65.33 | 71.64 |
| Non-Aggregated | MsDANN | **71.13** | **66.83** | **67.75** | **61.79** | **58.57** | **58.27** |
| | MsNN | 63.58 | 62.35 | 60.98 | 59.20 | 53.38 | 56.34 |

TABLE VII: Emotion recognition performance of MsDANN and MsNN based EEG-AVE model with different embedding strategies using aggregated and non-aggregated groundtruth for arousal.

| Groundtruth | EEG Model | EEG+Visual+Audio | | EEG+Visual | | EEG+Audio | |
|---|---|---|---|---|---|---|---|
| | | $P_{acc}$ | $P_f$ | $P_{acc}$ | $P_f$ | $P_{acc}$ | $P_f$ |
| Aggregated | MsDANN | **85.59** | **86.55** | **77.28** | **78.57** | **68.55** | **72.20** |
| | MsNN | 78.39 | 81.42 | 72.02 | 75.76 | 65.33 | 71.43 |
| Non-Aggregated | MsDANN | **66.47** | **63.25** | **63.26** | **59.27** | **54.91** | **53.64** |
| | MsNN | 58.00 | 58.17 | 55.14 | 54.93 | 49.59 | 52.22 |

TABLE VIII: Performance comparison of audio-visual affective representation using aggregated and non-aggregated groundtruth for valence.

| Groundtruth | Modality | Deep Features | | Handcrafted Features | |
|---|---|---|---|---|---|
| | | $P_{acc}$ | $P_f$ | $P_{acc}$ | $P_f$ |
| Aggregated | Visual | **73.81** | **72.31** | 57.91 | 57.03 |
| | Audio | **69.09** | **71.99** | 60.83 | 55.09 |
| Non-Aggregated | Visual | **67.60** | **59.06** | 46.70 | 35.39 |
| | Audio | **58.28** | **56.04** | 56.42 | 39.26 |

TABLE IX: Performance comparison of audio-visual affective representation using aggregated and non-aggregated groundtruth for arousal.

| Groundtruth | Modality | Deep Features | | Handcrafted Features | |
|---|---|---|---|---|---|
| | | $P_{acc}$ | $P_f$ | $P_{acc}$ | $P_f$ |
| Aggregated | Visual | **78.52** | **78.21** | 70.83 | 71.52 |
| | Audio | **69.34** | **71.51** | 65.49 | 70.92 |
| Non-Aggregated | Visual | **61.35** | **53.27** | 54.00 | 46.86 |
| | Audio | **53.75** | 49.60 | 52.39 | 53.70 |

promising affective detection results, comparing to the state-of-the-art methods. Besides, aiming at characterizing dynamic, informative, and domain-invariant EEG features across individuals, we develop a deep neural network with a transfer learning method (MsDANN) to solve the individual differences problem in the EEG data processing and investigate the performance variants with different neural network architectures (with or without domain adaptation). Our analysis demonstrates a superior cross-individual result is achieved under an evaluation of leave-one-individual-out cross-validation individual-independent method. Furthermore, we utilize two well-known pretrained CNNs for semantic audio-visual feature extraction and introduce hypergraph theory to decode deep visual features, deep auditory features, and deep audio-visual fusion features for intrinsic emotions estimation. The possibility of affective detection using the multimedia materials is verified and the benefit of the proposed embedding strategy is examined. These results show both EEG signals and audio-visual information play important and helpful roles in affective detection, and the proposed EEG-AVE model could be applied to boost the development of affective brain-computer interface in the real applications.

## VIII. Conflicts of Interest

The authors declare that they have no conflicts of interest.

## IX. Acknowledgments

This study was supported by National Natural Science Foundation of China (No.61906122 and No.91859122).

## References

[1] T. Song, W. Zheng, P. Song, and Z. Cui, "Eeg emotion recognition using dynamical graph convolutional neural networks," *IEEE Transactions on Affective Computing*, vol. 11, no. 3, pp. 532–541, 2018.

[2] J. Li, S. Qiu, Y.-Y. Shen, C.-L. Liu, and H. He, "Multisource transfer learning for cross-subject eeg emotion recognition," *IEEE transactions on cybernetics*, vol. 50, no. 7, pp. 3281–3293, 2019.

[3] J. Li, S. Qiu, C. Du, Y. Wang, and H. He, "Domain adaptation for eeg emotion recognition based on latent representation similarity," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 12, no. 2, pp. 344–353, 2019.

[4] Y. Cimtay and E. Ekmekcioglu, "Investigating the use of pretrained convolutional neural network on cross-subject and cross-dataset eeg emotion recognition," *Sensors*, vol. 20, no. 7, p. 2034, 2020.

[5] Y. Yin, X. Zheng, B. Hu, Y. Zhang, and X. Cui, "Eeg emotion recognition using fusion model of graph convolutional neural networks and lstm," *Applied Soft Computing*, vol. 100, p. 106954, 2021.

[6] S. Jirayucharoensak, S. Pan-Ngum, and P. Israsena, "Eeg-based emotion recognition using deep learning network with principal component based covariate shift adaptation," *The Scientific World Journal*, vol. 2014, 2014.

[7] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks," *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 3, pp. 162–175, 2015.

[8] H. Cui, A. Liu, X. Zhang, X. Chen, K. Wang, and X. Chen, "Eeg-based emotion recognition using an end-to-end regional-asymmetric convolutional neural network," *Knowledge-Based Systems*, vol. 205, p. 106243, 2020.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.06.455362; this version posted August 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.




[9] W.-L. Zheng, Y.-Q. Zhang, J.-Y. Zhu, and B.-L. Lu, "Transfer components between subjects for eeg-based emotion recognition," in *2015 international conference on affective computing and intelligent interaction (ACII)*. IEEE, 2015, pp. 917–922.

[10] W.-L. Zheng and B.-L. Lu, "Personalizing eeg-based affective models with transfer learning," in *Proceedings of the twenty-fifth international joint conference on artificial intelligence*, 2016, pp. 2732–2738.

[11] Y.-P. Lin and T.-P. Jung, "Improving eeg-based emotion classification using conditional transfer learning," *Frontiers in human neuroscience*, vol. 11, p. 334, 2017.

[12] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[13] Y. Li, W. Zheng, Y. Zong, Z. Cui, T. Zhang, and X. Zhou, "A bi-hemisphere domain adversarial neural network model for eeg emotion recognition," *IEEE Transactions on Affective Computing*, 2018.

[14] Y. Wang, L. Guan, and A. N. Venetsanopoulos, "Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition," *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 597–607, 2012.

[15] S. Mo, J. Niu, Y. Su, and S. K. Das, "A novel feature set for video emotion recognition," *Neurocomputing*, vol. 291, pp. 11–20, 2018.

[16] E. Acar, F. Hopfgartner, and S. Albayrak, "A comprehensive study on mid-level representation and ensemble learning for emotional analysis of video material," *Multimedia Tools and Applications*, vol. 76, no. 9, pp. 11 809–11 837, 2017.

[17] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, "Learning affective features with a hybrid deep model for audio–visual emotion recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 3030–3043, 2017.

[18] J. Cheng, M. Chen, C. Li, Y. Liu, R. Song, A. Liu, and X. Chen, "Emotion recognition from multi-channel eeg via deep forest," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 2, pp. 453–464, 2020.

[19] S. Kim, H.-J. Yang, N. A. T. Nguyen, S. K. Prabhakar, and S.-W. Lee, "Wedea: A new eeg-based framework for emotion recognition," *IEEE Journal of Biomedical and Health Informatics*, 2021.

[20] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari, "Audio-visual emotion recognition in video clips," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 60–75, 2017.

[21] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.

[22] S. Wang, S. Chen, and Q. Ji, "Content-based video emotion tagging augmented by users' multiple physiological responses," *IEEE Transactions on Affective Computing*, vol. 10, no. 2, pp. 155–166, 2017.

[23] Y. Tonoyan, D. Looney, D. P. Mandic, and M. M. Van Hulle, "Discriminating multiple emotional states from eeg using a data-adaptive, multiscale information-theoretic approach," *International journal of neural systems*, vol. 26, no. 02, p. 1650005, 2016.

[24] K. Michalopoulos and N. Bourbakis, "Application of multiscale entropy on eeg signals for emotion detection," in *2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE, 2017, pp. 341–344.

[25] A. Martínez-Rodrigo, B. García-Martínez, R. Alcaraz, P. González, and A. Fernández-Caballero, "Multiscale entropy analysis for recognition of visually elicited negative stress from eeg recordings," *International journal of neural systems*, vol. 29, no. 02, p. 1850038, 2019.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[27] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2017, pp. 131–135.

[28] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu, "Differential entropy feature for eeg-based emotion classification," in *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 2013, pp. 81–84.

[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[30] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[31] L. Shi, K. Du, C. Zhang, H. Ma, and W. Yan, "Lung sound recognition algorithm based on vggish-bigru," *IEEE Access*, vol. 7, pp. 139 438–139 449, 2019.

[32] S. Kurada and A. Kurada, "Poster: Vggish embeddings based audio classifiers to improve parkinson's disease diagnosis," in *2020 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*. ACM, 2020, pp. 9–11.

[33] W. Han, T. Jiang, Y. Li, B. Schuller, and H. Ruan, "Ordinal learning for emotion recognition in customer service calls," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6494–6498.

[34] A. Ducournau, S. Rital, A. Bretto, and B. Laget, "A multilevel spectral hypergraph partitioning approach for color image segmentation," in *2009 IEEE International Conference on Signal and Image Processing Applications*. IEEE, 2009, pp. 419–424.

[35] D. Zhou, J. Huang, and B. Schölkopf, "Learning with hypergraphs: Clustering, classification, and embedding," *Advances in neural information processing systems*, vol. 19, pp. 1601–1608, 2006.

[36] 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, no. 1, pp. 42–55, 2011.

[37] H. H. Jasper, "The ten-twenty electrode system of the international federation," *Electroencephalogr. Clin. Neurophysiol.*, vol. 10, pp. 370–375, 1958.

[38] Y. Zhu, S. Wang, and Q. Ji, "Emotion recognition from users' eeg signals with the help of stimulus videos," in *2014 IEEE international conference on multimedia and expo (ICME)*. IEEE, 2014, pp. 1–6.

[39] X. Huang, J. Kortelainen, G. Zhao, X. Li, A. Moilanen, T. Seppänen, and M. Pietikäinen, "Multi-modal emotion analysis from facial expressions and electroencephalogram," *Computer Vision and Image Understanding*, vol. 147, pp. 114–124, 2016.

[40] S. Rayatdoost and M. Soleymani, "Cross-corpus eeg-based emotion recognition," in *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2018, pp. 1–6.

[41] Y. Huang, J. Yang, S. Liu, and J. Pan, "Combining facial expressions and electroencephalography to enhance emotion recognition," *Future Internet*, vol. 11, no. 5, p. 105, 2019.

[42] Z. Yin, L. Liu, J. Chen, B. Zhao, and Y. Wang, "Locally robust eeg feature selection for individual-independent emotion recognition," *Expert Systems with Applications*, vol. 162, p. 113768, 2020.

[43] W. Zhang and Z. Yin, "Eeg feature selection for emotion recognition based on cross-subject recursive feature elimination," in *2020 39th Chinese Control Conference (CCC)*. IEEE, 2020, pp. 6256–6261.

[44] M. Soleymani, G. Chanel, J. J. Kierkels, and T. Pun, "Affective ranking of movie scenes using physiological signals and content analysis," in *Proceedings of the 2nd ACM Workshop on Multimedia Semantics*, 2008, pp. 32–39.

[45] M. Soleymani, J. J. Kierkels, G. Chanel, and T. Pun, "A bayesian framework for video affective representation," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. IEEE, 2009, pp. 1–7.

[46] A. Yazdani, K. Kappeler, and T. Ebrahimi, "Affective content analysis of music video clips," in *Proceedings of the 1st international ACM workshop on Music information retrieval with user-centered and multimodal strategies*, 2011, pp. 7–12.