Feature Extraction for Emotion Recognition using Facial Videos and Physiological Signals*

Bhargav Ramudu Manam
M.Sc. Data Science and Artificial Intelligence

Université Côte d'Azur, 06410 Sophia Antipolis, France

Abstract. The role of human computer interaction (HCI) is omnipresent in the current era of highly digitized society. Anyone having access to computers, smartphones, internet of things (IOT) devices, bio-trackers or even smartwatches can generate meaningful data that can be used to solve a variety of complex problems through HCI. One such area of HCI is affective computing (AC) or emotion recognition (ER) which has wide applications in medical diagnosis, education, fraud detection and so on. In this report, the focus is on exploring the usage of multiple modalities from the data containing visual and physiological information (EEG, ECG, PPG, GSR, ...) in multi-modal emotion recognition (MER) systems and also to explore current state-of-the-art feature extraction techniques. A review has been made on MAHNOB and DEAP databases and it is found that multi-modal analysis achieves superior performance over single modality. Most of the feature extraction and feature selection methods from the literature are identified to be either based on domain specific knowledge (hand crafted features) or deep learning (DL) and discussed briefly.

Keywords: human computer interaction (HCI) · affective computing (AC) · multi-modal emotion recognition (MER) · physiological signals · feature extraction · feature selection · deep learning (DL) · MAHNOB · DEAP.

1 Introduction

Affective computing (AC) or Emotion recognition (ER) has been a key area of focus within the scope of human computer interaction (HCI) for many years. As the name suggests, ER involves identifying and classifying the emotional state of an individual into various categories (happy, sad, neutral, ...), by means of a machine or an algorithm. This is achieved by extracting features (from various types of data such as speech, voice, facial expression or biological signals) and analyzing their relationships in identifying the emotions by training a classifier.

Despite of a lot of research in this field, still ER or AC remains an open problem due to the complexity of human emotions. It is very hard to classify or

* Supervised by Francois Bremond, Laura Ferrari and Valeriya Strizhkova of STARS Team, INRIA, Sophia Antipolis.

model human emotions as the boundaries are not well defined because different people express emotions differently in their own way. Also emotion perception by humans themselves can be very subjective depending on who you ask. Evaluating emotions based on a single modality is most common but this is often misleading since there is a good chance of loss of information (context) due to the absence of other modalities (see figure (1)). Emotion prediction with multiple modalities fare better than that of with single modality as different modalities complement each other.

ER using facial expressions or videos is drawing a lot of interest not only due to the advancements in computer vision (CV) or DL but also due to their large variety of applications. A real-time ER system can accurately identify emotions, which is very crucial for HCI systems as they provide important cues such as feedback on user experience, the effectiveness of user-interface or responsiveness of an impaired patient to a treatment. Video based ER systems also has several use cases such as live sentiment analysis and integration into virtual reality headset to display emotion specific content. ER using visual data is challenging because often the quality of videos (in the databases) are not consistent i.e., lighting conditions are bad or significant portion of the face is covered by some object (occlusion). Also, lack of large and diverse databases for videos makes the training hard and the ER model prone to bias.

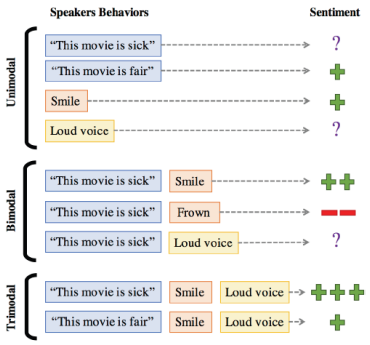

Fig. 1. Uni-modal, bi-modal and tri-modal interaction in MER systems, Fig. from (22)

Physiological or bio signals are widely used to study ER in addition to common modalities (video, audio and text), as they greatly enhance the model performance. This is because physiological signals are regulated by the central nervous system (CNS) and autonomous nervous system (ANS) and are difficult to manipulate. Signals from ANS are involuntary, this makes them uncontrollable. Hence, bio signals are less deceptive than other modalities. Also, cannon's theory (2) suggests that when there is an emotion change, the CNS and ANS impact the physiological responses, therefore, it is crucial to analyse the biological signals.

2 Literature Review

2.1 Databases

After reviewing the literature, MAHNOB (18) and DEAP (9) databases are identified to be the most popular among the scientific community, for a MER Task, with visual and physiological signals. Studies based on these two databases

are mentioned and analysed in the next section. Both of these databases are developed, using similar procedures and protocols, in a controlled lab setting. The subjects are made to watch set of videos, each eliciting a particular emotion and at the end of each video they are asked to answer a set of questions. The questions include rating the videos, on a given scale based on arousal, valence or choosing an emotion category, etc. During this process, subjects facial movements are captured using cameras while their physiological data has been recorded with the help of a multi-sensor setup. A brief overview of the databases is presented in the table (1).

Database	**MAHNOB**	**DEAP**
# Subjects	27	32
# Sessions per subject	20	40
Session length	Varied (34.9 - 117 sec)	Fixed (60 sec)
Induced or Natural Emotion	Induced	Induced
Posed or Spontaneous Emotion	Spontaneous	Spontaneous
Modalities	Audio (2 channels), Visual (6 views), EEG (32 channels), ECG (3 channels), GSR, Respiration Amplitude, Skin Temperature, Eye-Gaze data	EEG(32 channels), PPG, GSR, EOG (4 channels), EMG (4 channels), Visual (for 22 subjects), Respiration Amplitude, Skin Temperature
Sampling rate for bio-signals	256 Hz	256 Hz
Metrics assessed	Arousal, Valence, Emotion, Dominance, Predictability	Arousal, Valence, Dominance, Liking, Familiarity
Rating scales	Discrete scale: 1 - 9	Continuous scale: 1 - 9, Discrete scale: 1 - 5 (Familiarity)
Assessment	Self-Assessment	Self-Assessment
Baseline data (Fixation cross)	30 sec	5 sec

Table 1: Overview of MAHNOB and DEAP databases

The bio-signals collected include Electroencephalogram (EEG), Electrocardiogram (ECG), Photoplethysmogram (PPG), Electrooculogram (EOG), Electromyogram (EMG) and Electrodermal activity (EDA) or Galvanic Skin Response (GSR). All except PPG are electrical signals, whereas PPG is an optic signal. EEG is used to measure the brain activity that occurs due to ionic current among the neurons. EEG reflects the changes in voltage fluctuations across the pair of electrodes placed over the scalp. The location of these electrodes is determined by an expert. Usually, EEG signal is recorded in several channels across the scalp simultaneously. ECG signal is related to heart activity. Every heart beat is comprised of three components: P wave, QRS complex and the T wave. PPG is a plethysmogram optically deduced to detect blood volume changes in the micro vascular bed of a target tissue. Both ECG and PPG are used to monitor and detect Heart Rate (HR) and Heart Rate Variability (HRV). EOG signal detects the potential differences across the eye and useful in obtaining eye movement and blinks. EMG produces a signal that can detect neurological activity through the body muscles. EDA or Galvanic Skin Response (GSR) measures the skin conductance that arise due to autonomic sympathetic changes. EDA has direct applications in stress monitoring and lie detection as it can infer ANS activity which are closely associated with emotional and cognitive states.

2.2 State-of-the-art (SOA) Review

Table (2) presents the SOA review on MAHNOB and DEAP databases (top 4 models are in bold). MAHNOB and DEAP databases are selected for the review due to the variety of input modalities. The declared model accuracies are based on binary classification tasks (random chance of 50%) on arousal and valence with each having two classes i.e., arousal- High and low and valence- High and low. Only the first four results from (18) are based on 3-class classification (random chance $\approx 33.33\%$).

Reference	**Modality**	**Extracted Features**	**Accuracy (%)**
		Database: MAHNOB (18)	Arousal, Valence
Modalities: EEG, ECG, GSR, Respiration (R), Visual (V), Skin Temperature (ST), Eye gaze			
(18)	All	Hand crafted features and selection through ANOVA test	46.2, 45.5
(18)	EEG	Hand crafted features and selection through ANOVA test	52.4, 57.0
(18)	Eye gaze	Hand crafted features and selection through ANOVA test	63.5, 68.8
(18)	EEG, Eye gaze	Hand crafted features and selection through ANOVA test	67.7, 76.1
(10)	EEG	Hand crafted features and selection through ICA	66.0, 71.5
(10)	V	Features based on Action Units (AUs) and selection through ICA	65.0, 64.5
(10)	EEG, V	Hand crafted features, AUs and selection through ICA	68.0, 72.5
(16)	EEG	PSD, CE, from power spectrum images using pre-trained VGG-16 and selection through PCA	**80.42, 80.77**
(16)	ECG	HR, HRV, from spectogram images using pre-trained VGG-16 and selection through PCA	78.76, 78.76
(16)	GSR	Statistical, from spectogram images using pre-trained VGG-16 and selection through PCA	81.84, 78.98
(16)	V	Based on AUs (each frame) and statistical (all frames)	**82.15, 83.04**
(16)	V	Using pre-trained VGG-16 and selection through PCA	**81.57, 85.13**
(16)	EEG, ECG, GSR	Fusion of individual features mentioned above	80.61, 80.36
(16)	EEG, V	Fusion of individual features mentioned above	**82.93, 85.49**
		Database: DEAP (9)	Arousal, Valence
Modalities: EEG, PPG, GSR, EOG, Visual (V)			
(11)	EEG	High level features through 2 layer DBN	64.3, 58.4
(16)	EEG	PSD, CE, from power spectrum images using pre-trained VGG-16 and selection through PCA	72.58, 71.09
(16)	PPG	HR, HRV, from spectogram images using pre-trained VGG-16 and selection through PCA	71.09, 70.86
(16)	GSR	Statistical, from spectogram images using pre-trained VGG-16 and selection through PCA	71.64, 70.70
(16)	V	Based on AUs (each frame) and statistical (all frames)	72.21, 71.08
(16)	V	Using pre-trained VGG-16 and selection through PCA	74.47, 72.28
(16)	EEG, V	Fusion of individual features mentioned above	74.13, 73.94
(16)	EEG, V	Fusion of individual features mentioned above (by taking mean PSD images and facial images per every sec) and then passing them through a LSTM to account for time dependency	**78.34, 79.52**
(16)	EEG, PPG, GSR	Fusion of individual features mentioned above	73.05, 71.87
(12)	EEG, EOG	High level features using BDAE	**80.5, 85.2**
(21)	All	Hand crafted features and passed through a SAE network separately and fused in the end	**84.18, 83.04**
(19)	All	DE and Time-Domain and dimension reductionality using a BDDAE network	**83.23, 83.82**

Table 2: Review of MAHNOB and DEAP databases

Inferences:

1. Multi-modal models classify arousal and valence better than those using single-modality.
2. Overall, for a given modality, classification using visual modality achieves higher accuracy followed by EEG.
3. Single modality ER model can perform better than MER model if the modalities are not chosen well.
4. Models using DL based feature extraction strategies fare better than those using rule based or hand crafted features, though the difference is not significantly large.

Remark 1. In the table, ANOVA: Analysis of Variance, CE: Conditional Entropy, ICA: Independent Component Analysis, DBN: Deep Belief Networks, BDAE: Bi-directional Auto Encoder, BDDAE: Bimodal Deep Denoising Auto Encoder.

3 Proposed Methodology for MER Systems

In this section, a common methodology (refer fig. (2)), based on the literature, is proposed for MER systems.

3.1 Data Acquisition

There are several databases containing visual and physiological data: DEAP (9), AMIGOS (15) and MAHNOB-HCI (18). Data for analysis can be acquired easily since most of these databases are either publicly available or can be accessed freely for research purposes by taking necessary permissions.

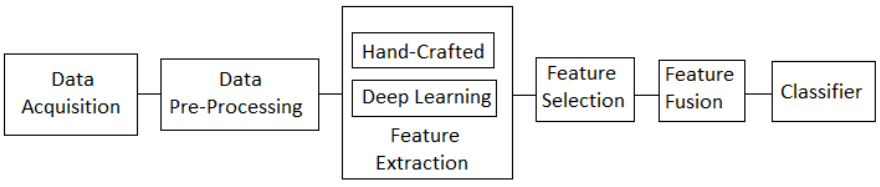

Fig. 2. Proposed methodology for feature extraction in a Multi-Modal Emotion Recognition (MER) Task

3.2 Data Pre-processing

The input data from different modalities may contain lot of unnecessary components like noise, artifacts, missing information for some sessions, outliers or some form of inconsistencies. Hence, the data needs to be processed before feature extraction step, in order to make it effective for extracting features or to feed into an end-to-end DL architecture.

Videos:

1. Segmentation: the input videos has certain sampling rate (frames per second) at which an image frame is generated. This should be set taking into considerations on both computation cost and sampling rate of other modalities (for parallel or combined feature extraction)
2. Face detection: image from each frame is processed to highlight or crop into the region containing only face for better feature extraction. Popularly, Viola-jones (20) algorithm or OpenCV (7) are used for face detection.
3. Resolution: each image is resized according to the input dimensions of the model.

Physiological signals:

1. Filtering and Emphasizing: unnecessary frequencies and artifacts occurred during data collection can be removed by applying low-pass, high-pass and band-pass filters respectively. Consequently, filtering can be used to emphasize the high frequencies as they usually have low amplitudes. This is beneficial during fourier transform computation.
2. Segmentation: similar to videos, bio signals are divided into windows (frames) and the length of the window is chosen as per requirement (keeping in mind the frame rate of videos). Consecutive windows are overlapped (around 50%) so that spectral information is not lost (non-stationary signals). Additionally, signal within each window is assumed to be stationary.
3. Hamming: Hamming window function is applied over the window for smoothing the signal.
   The Hamming window function is given by:
   $$w[n] = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right)$$
   where n is the sample index and N is the length of the window.
4. Discrete fourier transform (DFT): used to convert the signal from time domain to frequency domain.
5. Discrete Wavelet transform (DWT): used to remove noise from signals.

Remark 2. Data augmentation can be performed after data pre-processing step, to generate additional training data, if the size of available data is small. This also helps in better generalizability of the model.

3.3 Feature Extraction

Hand crafted features: these are the features extracted by experts, using the domain knowledge of the particular modality, to identify the most relevant features for a given task. Some times these can also be referred as Machine Learning (ML) or rule-based features since ML algorithms can be formulated to extract these features.

For Videos

1. Facial landmarks: the images in each frame are used for detecting relevant points on the face. These points are typically surrounded around eyes, nose and mouth. These points are then used to calculate certain distances (features) such as distances between eyes, lips and nose, etc,.
2. Facial Action Coding System (FACS) or Action Units (AUs): the features obtained from facial landmarks for a given session can be used to determine high level features called Action Units (6) based on (5)(AUs) or gestures (3) (lip, eye, head position/movements). In turn, these AUs or gestures are associated with a certain emotion (8). Chehra (1) is a SOA algorithm to extract AUs from facial landmark features.

For Physiological Signals

1. Time domain (TD): these are based on the time series data of the signal containing amplitude at a given time. Simple statistical features can be deduced based on time domain signal such as HR, HRV, etc,.
2. Frequency or spectral domain (FD): the signal in time domain can be converted into frequency domain and analysed. This can be done using DFT or short-time fourier transform (STFT). In (18) and (9), EEG signals in frequency domain are divided into, theta ($4Hz < f < 8Hz$), slow alpha ($8Hz < f < 10Hz$), alpha ($8Hz < f < 12Hz$), beta ($12Hz < f < 30Hz$), and gamma ($30Hz < f$), bands and features such as power spectral density (PSD) and spectral entropy (SE) are evaluated for all bands.
3. Time-frequency domain (TFD): these signify the change in frequency or spectral content over time. This can be achieved by either wavelet transform (WT) or spectrogram. Features such as wavelet entropy, wavelet energy and power spectral entropy are obtained from this domain.

Statistical features, such as mean, min., max., higher order moments, percentile, etc, can be calculated for all the three domains mentioned above.

Table (3) contains the information about several hand crafted features extracted on MAHNOB and DEAP databases.

Deep learning features: feature generation using DL is performed by propagating the input modalities through multiple layers of linear and non-linear

Signal (# features)	**Extracted hand crafted features**
EEG (216)	theta, slow alpha, alpha, beta, and gamma Spectral power for each electrode. The spectral power asymmetry between 14 pairs of electrodes in the four bands of alpha, beta, theta and gamma.
ECG or PPG (64)	Average and standard deviation of HR, HRV, and inter beat intervals, energy ratio between the frequency bands [0.04-0.15]Hz and [0.15-0.5]Hz, spectral power in the bands ([0.1-0.2]Hz, [0.2-0.3]Hz, [0.3-0.4]Hz), low frequency [0.01-0.08]Hz, medium frequency [0.08-0.15]Hz and high frequency [0.15-0.5]Hz components of HRV power spectrum.
GSR or EDA (20)	average skin resistance, average of derivative, average of derivative for negative values only (average decrease rate during decay time), proportion of negative samples in the derivative vs. all samples, number of local minima in the GSR signal, average rising time of the GSR signal, 10 spectral power in the [0-2.4]Hz bands, zero crossing rate of Skin conductance slow response (SCSR) [0-0.2]Hz, zero crossing rate of Skin conductance very slow response (SCVSR) [0-0.08]Hz, SCSR and SCVSR mean of peaks magnitude
Respiration amplitude (14)	band energy ratio (difference between the logarithm of pattern energy between the lower (0.05-0.25Hz) and the higher (0.25-5Hz) bands), average respiration signal, mean of derivative (variation of the respiration signal), standard deviation, range or greatest breath, breathing rhythm (spectral centroid), breathing rate, 10 spectral power in the bands from 0 to 2.4Hz, average peak to peak time, median peak to peak time
Skin temperature (4)	average, average of its derivative, spectral power in the bands ([0-0.1]Hz, [0.1-0.2]Hz)
EMG and EOG (4), only for DEAP	eye blinking rate, energy of the signal, mean and variance of the signal

Table 3: Hand crafted features extracted by the original authors of MAHNOB(18) and DEAP (9) databases

operations. DL architecture learns the high-level representation of data and generate features. These features are then used to classify the emotion present in the session. The entire network can be optimized for better performance, by training the parameters of the DL model through back propagation.

As we have already seen, time series inputs need to be discretized (DFT, WT) during signal processing and this results in loss of information. DL approaches have the advantage to avoid this loss as they are capable of taking continuous signals as inputs. Most generally used DL architectures for feature extraction are stacked autoencoders (SAE), convolution neural network (CNN), recurrent neural network (RNN) and RNN-LSTM (Long Short-Term Memory).

DL methods have the advantage of having end-to-end architecture, this saves the effort of curating the hand crafted features, which often is a tedious job.

Image frames can be feed into large-scale image-recognition DL networks like VGG network (17) for DL feature extraction. (16) used convolution-deconvolution network on several channels of an EEG signal to obtain spatial information such as salient brain regions corresponding to an emotion. Using similar networks, one can avoid using AUs by obtaining the class (emotion) activation maps, that are more appropriate.

Hybrid features: these features can be those that are obtained by applying DL architectures on the existing hand crafted features. For example, the spectrograms (images) generated from time-frequency domain can be fed into a DL based model (CNN) to further extract DL features using a pre-trained model like (17).

Remark 3. Feature normalization or standardization needs to be done once all the features are extracted to avoid biases and increase stability of the ER model.

3.4 Feature Selection

Feature Selection methods are necessary since most often the dimension of features space is very large compared to the sample of data available. There is a need to identify redundant and highly correlated features otherwise they will require high computational cost to analyse those features and also this means that model learns a lot of parameters. This can easily lead to over-fitting and poor model generalization. In addition, there is the problem of "Curse of dimensionality" and in very high dimensions the notion of distance and relative position seem to disappear (mathematically). (4) proposes using certain filtering methods based on Pearson's Correlation coefficient and Shannon's Entropy (based on mutual information among features) to eliminate similar and unwanted features. Classical methods, Principal Component Analysis (PCA) and ICA, are also widely used in dimensionality reduction.

Also, autoencoders (AE) as opined by (14) can act as a very good feature selectors because of its rich representation of features into high level information (in latent space). They are also very good at detecting outliers. After training the AE for encoding and decoding, the decoding network is removed and features in the encoded space (reduced dimension) can be used as new features.

3.5 Feature Fusion

In order to account for the inter-dependency among the modalities, individual features obtained after feature selection are fused together with the help of a fusion mechanism. The two most popular ways of feature fusion are discussed below.

Feature level Fusion:

Individual features from each session are concatenated (early fusion) or represented by n-fold Cartesian space (22) (n: number of modalities) and then passed to a classifier for classification.

Decision level Fusion:

Here a decision (class of emotion) is already made by different classifiers using individual modalities and final decision is obtained by fusing all the individual decisions by some aggregation (weighted) strategy. The importance or weight assigned to each decision is to be treated as a hyper-parameter and is learned during training of the model.

3.6 Classifier

There are several possible classifiers one can use based on literature such as linear discriminant analysis (LDA) classifier, quadratic discriminant analysis (QDA), Support vector Machine (SVM) classifier or they can be a neural network based (multi-layer perceptron or CNN). The authors in (13) discuss several such classifiers for ER using EEG and presented guidelines for selection of a classifier.

Remark 4. If one uses a DL model with an end-to-end structure, all the steps after data pre-processing are handled within the DL architecture and the emotion output is generated. In these models, interpretability is often an issue.

4 Conclusion and Future Work

In this report, the emotion recognition task using visual and physiological modalities is discussed. Scientific literature based on MAHNOB and DEAP databases is explored and analysed. SOA review establishes the fact that emotion recognition using multi-modalities is more accurate and model performance depends on feature extraction strategies. Recent studies show that more researchers are giving preference to DL based feature extraction methods over traditional hand crafted features. A general methodology for MER task has been proposed based on the SOA review. Several possible domains of feature extraction and feature selection techniques are mentioned and discussed briefly.

To conclude, MER systems are classically solved using ML (model specific) algorithms but more recently researchers are moving towards DL models (model free) due to their superior performance and easy data handling. Feature Extraction and selection methods presented are more relevant in the context of ML models than the DL models, as DL models are capable of learning high level features themselves from the raw signals or input data due to their inherent non-linear structures. Additionally, there are several methods which are in-between (hybrid) trying to take the advantage of both approaches.

As a future work, the proposed methodology can be carried out on databases AMIGOS, DEAP and MAHNOB-HCI. Experiments can be run on several combinations of available modalities and results are compared against existing SOA. Also, due to similarity of these databases it can be interesting to train and test models by combining several databases and the resultant model can be used to make predictions or used as a pre-trained model on any other similar database (transfer learning). This can result in a much more generalized model that can be deployed real-time. Additionally, a similar framework for MER using more common mediums such as audio and text can be developed and integrated. Finally, as it was reported that DL frameworks were only performing marginally better than the traditional methods, there is a need to look into the architectures of the existing DL models and modify them or design a better model that can capture the maximum information from the input modalities and can extract rich features.

Bibliography

[1] Asthana, A., Zafeiriou, S., Cheng, S., Pantic, M.: Incremental face alignment in the wild. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1859–1866 (2014). https://doi.org/10.1109/CVPR.2014.240

[2] Cannon, W.B.: The james-lange theory of emotions: a critical examination and an alternative theory. by walter b. cannon, 1927. The American journal of psychology **100 3-4**, 567–86 (1987)

[3] Caridakis, G., Asteriadis, S., Karpouzis, K., Kollias, S.: Detecting human behavior emotional cues in natural interaction. In: 2011 17th International Conference on Digital Signal Processing (DSP). pp. 1–6 (2011). https://doi.org/10.1109/ICDSP.2011.6004962

[4] Chandrashekar, G., Sahin, F.: A survey on feature selection methods. Comput. Electr. Eng. **40**, 16–28 (2014)

[5] Ekman, P., Freisen, W.V., Ancoli, S.: Facial signs of emotional experience. Journal of Personality and Social Psychology **39**, 1125–1134 (1980)

[6] Ekman, P., Friesen, W.V.: Facial action coding system: a technique for the measurement of facial movement (1978)

[7] Itseez: The OpenCV Reference Manual, 2.4.9.0 edn. (April 2014)

[8] Kleinsmith, A., Bianchi-Berthouze, N.: Affective body expression perception and recognition: A survey. IEEE Transactions on Affective Computing **4**(1), 15–33 (2013). https://doi.org/10.1109/T-AFFC.2012.16

[9] Koelstra, S., Muhl, C., Soleymani, M., Lee, J.S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., Patras, I.: Deap: A database for emotion analysis ;using physiological signals. IEEE Transactions on Affective Computing **3**(1), 18–31 (2012). https://doi.org/10.1109/T-AFFC.2011.15

[10] Koelstra, S., Patras, I.: Fusion of facial expressions and eeg for implicit affective tagging. Image and Vision Computing **31**(2), 164–174 (2013)

[11] Li, X., Zhang, P., Song, D., Yu, G., Hou, Y., Hu, B.: Eeg based emotion identification using unsupervised deep feature learning. In: SIGIR2015 Workshop on Neuro-Physiological Methods in IR Research (August 2015), http://oro.open.ac.uk/44132/

[12] Liu, W., Zheng, W.L., Lu, B.L.: Emotion recognition using multimodal deep learning. vol. 9948 (10 2016). https://doi.org/10.1007/978-3-319-46672-9$_5$8

[13] Lotte, F., Bougrain, L., Cichocki, A., Clerc, M., Congedo, M., Rakotomamonjy, A., Yger, F.: A review of classification algorithms for eeg-based brain–computer interfaces: a 10 year update. Journal of neural engineering **15**(3), 031005 (2018)

[14] Martinez, H.P., Bengio, Y., Yannakakis, G.N.: Learning deep physiological models of affect. IEEE Computational Intelligence Magazine **8**(2), 20–33 (2013). https://doi.org/10.1109/MCI.2013.2247823

[15] Miranda-Correa, J.A., Abadi, M.K., Sebe, N., Patras, I.: Amigos: A dataset for affect, personality and mood research on individuals and

groups. IEEE Transactions on Affective Computing **12**(2), 479–493 (2021). https://doi.org/10.1109/TAFFC.2018.2884461

[16] Siddharth, Jung, T., Sejnowski, T.J.: Utilizing deep learning towards multi-modal bio-sensing and vision-based affective computing. CoRR **abs/1905.07039** (2019), http://arxiv.org/abs/1905.07039

[17] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014). https://doi.org/10.48550/ARXIV.1409.1556, https://arxiv.org/abs/1409.1556

[18] Soleymani, M., Lichtenauer, J., Pun, T., Pantic, M.: A multimodal database for affect recognition and implicit tagging. IEEE Transactions on Affective Computing **3**(1), 42–55 (2012). https://doi.org/10.1109/T-AFFC.2011.25

[19] Tang, H., Liu, W., Zheng, W.L., Lu, B.L.: Multimodal emotion recognition using deep neural networks. pp. 811–819 (10 2017). https://doi.org/10.1007/978-3-319-70093-9$_8$6

[20] Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001. vol. 1, pp. I–I (2001). https://doi.org/10.1109/CVPR.2001.990517

[21] Yin, Z., Zhao, M., Wang, Y., Yang, J., Zhang, J.: Recognition of emotions using multimodal physiological signals and an ensemble deep learning model. Computer Methods and Programs in Biomedicine **140**, 93–110 (03 2017). https://doi.org/10.1016/j.cmpb.2016.12.005

[22] Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.: Tensor fusion network for multimodal sentiment analysis. CoRR **abs/1707.07250** (2017), http://arxiv.org/abs/1707.07250