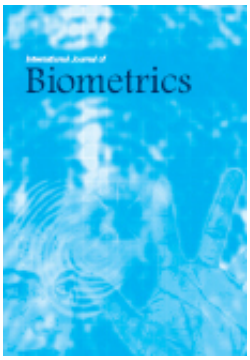
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Biometrics

ISSN online: 1755-831X - ISSN print: 1755-8301
https://www.inderscience.com/ijbm

Recent trends and challenges in human computer interaction using automatic emotion recognition: a review

Sukhpreet Kaur, Nilima Kulkarni

DOI: 10.1504/IJBM.2024.10053960

Article History:

Received:	24 August 2022
Last revised:	25 August 2022
Accepted:	23 December 2022
Published online:	01 December 2023

Recent trends and challenges in human computer interaction using automatic emotion recognition: a review

Sukhpreet Kaur* and Nilima Kulkarni

Department of Computer Science and Engineering,
MIT Art, Design and Technology University,
Pune, India
Email: sukhpreet.kaur@mituniversity.edu.in
Email: nilima.kulkarni@mituniversity.edu.in
*Corresponding author

Abstract: Automatic emotion recognition (AER) using facial expressions and electroencephalogram (EEG) signals is an interesting and booming area of research in the field of human computer interaction This paper aims to identify the key state-of-the-art methodologies, understand the standard workflow pipeline and know the existing findings. Different machine learning and deep learning approaches used recently for information pre-processing, feature extraction, feature classification and fusion schemes have also been explored. Furthermore, the purpose of this review work is to discuss the aspects motivating researchers to move from unimodal to multimodal AER systems. Also, this surveyed information is summarised in tabular forms to investigate the recent methods used and the results obtained. This comprehensive literature survey identifies the key points for inclusion of facial expressions and EEG signals over other channels, also the benefits of automated features, which are being leveraged over hand crafted features for building improved real time emotion recognition systems. This review work provides new research directions, open challenges and existing state-of-the-art methods in the field of AER using facial expressions and EEG signals which can be used as benchmark studies for researchers.

Keywords: emotion recognition; human computer interaction; affective computing; facial expressions; EEG signals; multimodal system.

Reference to this paper should be made as follows: Kaur, S. and Kulkarni, N. (2024) 'Recent trends and challenges in human computer interaction using automatic emotion recognition: a review', *Int. J. Biometrics*, Vol. 16, No. 1, pp.16–43.

Biographical notes: Sukhpreet Kaur received her Bachelor of Technology in Computer Science and Engineering from Punjab Technical University, India and Master of Technology in Computer Science from SGGSWU, India. Currently, she is a PhD research scholar at MIT Art, Design and Technology University, Pune, India. Also, she is working as an Assistant Professor in the Department of Computer Science and Engineering at MIT Art, Design and Technology University, Pune, India. Her areas of research interests include artificial intelligence, image processing, computer vision, deep learning and human-computer interaction.

Nilima Kulkarni is currently working as an Associate Professor in the Department of Computer Science and Engineering at MIT Art, Design and Technology University, Pune, India. She holds a PhD in Computer Science and Engineering from Amrita Vishwa Vidyapeetham (NCRF Ranked-8th), received Master's and Bachelor's in Computer Science Technology from SRT Marathwada University, Nanded, India. She has several years of working experience in teaching and research. Her area of research interest includes artificial intelligence, machine learning, computer vision, image processing analysis and human computer interaction.

1 Introduction

Automatic emotion recognition (AER) is the study of recognising one's emotions or sentiments using various computer science techniques. Emotions are the reflection of one's cognitive behaviour that is expressed in the form of happiness, sadness, joy, anger, fear, surprise, disgust. Along with emotions, other factors such as valence (polarity of emotion i.e. positive or negative), arousal (excitement or calmness level of an emotion) and dominance (control over the emotion) also play prominent roles to understand one's cognitive attitude, known as sentiment analysis (Kaur and Kulkarni, 2021b). Emotions are recognised through one's physiological or non-physiological channels, e.g., electroencephalogram (EEG) signals, facial expressions, voice, textual information, body movements, etc. The contribution of facial expression information is 55%, vocal part is 33% and semantic content is 7% for emotion recognition (Salama et al., 2021). So, it signifies the importance of the contribution of the multiple channels to recognise one's emotions. Multimodallty takes more than one modality or channel that may be a combination of visual, textual or physiological signal information for emotion recognition (Kaur and Kulkarni, 2021a). Therefore, AER is considered as a multifaceted area having different perspectives such as affective computing, artificial intelligence and behavioural or cognitive sciences (Mauss et al., 2009). Affective computing is a diverse field that targets emotion recognition using the disciplines in AI, cognitive sciences, physiological/non-physiological channels and sentiment analysis as shown in Figure 1. It enables hands free human computer interactions by understanding and interpreting one's emotions. Further, AI uses computer vision, signal processing and natural language processing to interpret the information (Koole, 2009; Poria et al., 2017; Sharmila, 2021). Cognitive processes are responsible for voluntary and involuntary body movements or activities in response to some stimuli (Singh and Kumar, 2021). It is categorised into the peripheral nervous system and autonomic nervous system e.g. facial expressions, limbs movements, hand/body gestures and heart rate, brain activities, blood pressure respectively.

We contribute a review of AER using EEG physiological signals and non-physiological facial expressions over other channels, recent advancements, datasets, different machine learning deep learning approaches used and remarks for better AER systems in future.

EEG signals are electrophysiological signals generated by the brain when some activity happens in the brain itself. Activity may be the mood change (emotional stimuli) or spontaneous reaction to any external action. Key points for inclusion of EEG signal as one of the channels are:

- Human's real emotional state can be recognised through EEG signals more effectively than facial expressions and voice based features. Another feasibility measure of research is easy availability EEG signal information acquisition equipment e.g. EEG electrode headset, EmotivEPOC device. It is a cost efficient and non-invasive data acquisition method (Zhang et al., 2017).
- Physiological expressions can be faked by changing expressions or tone of voice to hide his/her real emotions whereas spontaneous mood change or emotion change gives direct impact on EEG signals (Yang et al., 2018).
- EEG signals are preferred for cognitive research over other physiological signals such as EOG, ECG, EMG because they have better temporal resolution (good in recognising time variant features) and higher accuracy than other physiological signals (Lin et al., 2017).
- Physiological signals can be continuous and real time monitoring methods to control electrical activity of the brain that are non-invasive, using the electrodes placed over the scalp (Pampouchidou et al., 2019; Yang et al., 2018).
    a *Brain regions and their functioning in AER:* Human brain consists of four parts:
        1 frontal
        2 parietal
        3 temporal
        4 occipital, also known as lobes.

        Lobe channels are responsible for generating certain kinds of frequency waves against the specific emotion and channel locations on the brain can be seen in Tables 1 and 2 (Pampouchidou et al., 2019; Bhatti et al., 2019; Wang et al., 2018).

As mentioned in Fdez et al. (2021), there is a standard set of 10–20 system locations for placement of electrodes over the head as shown in Figure 2. A good question answered by this 10–20 system is how placement of electrodes can be made fixed over the head as head size varies for everyone. So the answer to it is electrodes are placed in a circular form 10% above the nasion and anion and 20% distance is maintained between the electrodes. This is the reason, it's called the 10–20 system.

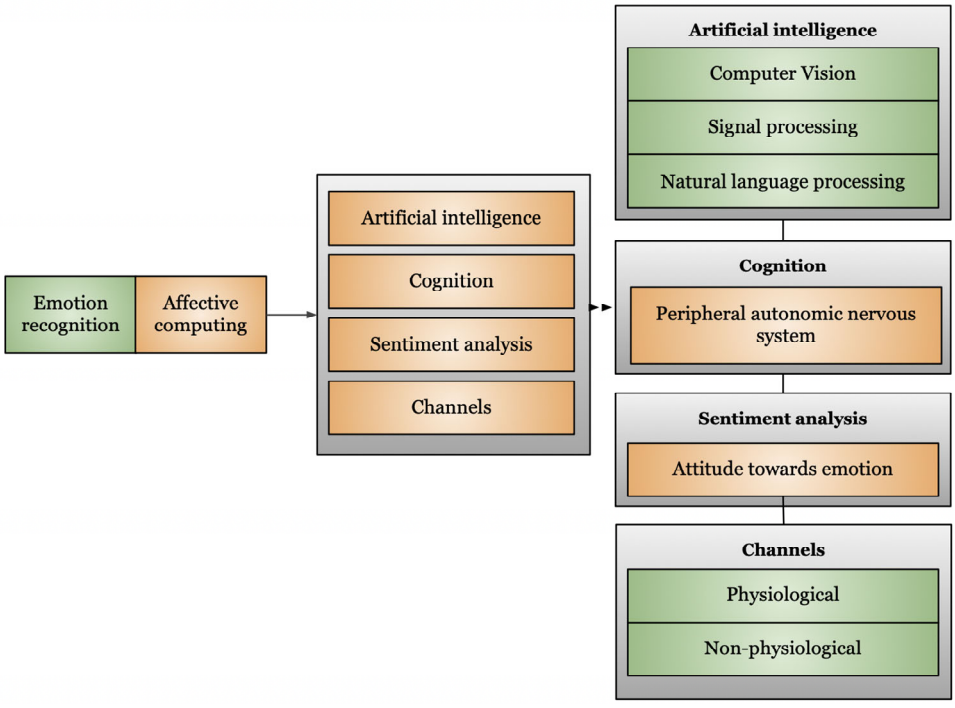
Table 1 EEG signals frequency bands

Frequency band	*Frequency range*	*State of mind*	*Consciousness*
Delta	0.5-4 Hz	Deep sleep	Lower
Theta	4–8 Hz	Light sleep	Low
Alpha	8–12 Hz	Relax state	Medium
Beta	12–30 Hz	Active thinking, focussed	High
Gamma	30–100 Hz	Cross-modal sensory processing	Higher

Table 2 Brain lobes, channels their functionality

Lobe	*Channels*	*Functionality*
Frontal	Fp1, Fp2, F3, F4, F7, F8, Fz	Personality, emotion, focus, planning, voluntary action and problem solving
Parietal	P3, P4, Pz	Orientation, recognition, perception to stimuli, proprioception
Temporal	T3, T4, T5, T6	Memory, speech and recognition of auditory stimuli
Occipital (smallest lobe)	O1, O2	Visual processing

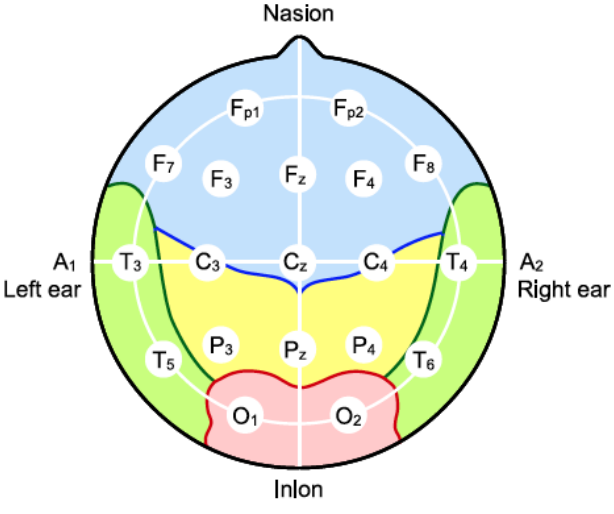
Figure 1 Affective computing taxonomy (see online version for colours)



Facial expressions are one of the most powerful, universal and non-verbal modes of communication among people which is not constrained by use of words or language. Also, human feelings at heart are directly reflected on the face. Computer vision and machine learning methods aid to interpret and encode one's feelings through his/her facial expression information. Facial expression recognition methods are categorised into two parts: geometric features and appearance based features (Maheshwariet al., 2021). As shown in Figure 3, Geometric features locate the positions of specific parts of the face such as mouth, lips, eyebrows etc. whereas appearance based features are based on either the entire face or some region in it. Key points for inclusion of facial expression information as one of the channels are (Yang et al., 2018; Bhattacharyya et al., 2020; Chaudhari et al., 2021):

- Seven basic emotions i.e. happiness, sadness, joy, anger, surprise, fear, disgust can be recognised accurately through the geometric locations of face parts like eyes, lips, mouth, nose, furrows etc as in figure. Where eyes and mouth contribute the most accurately for emotion recognition.
- Since the face is the most exposed part of the human body, computer vision systems visualise facial features handily, may be through images or real time capture in the form of video or image.
- Visual (facial expressions) are natural, easy to observe, and can also be captured using low priced equipment e.g. canon cameras.
- Facial emotion recognition has a big advantage in the field of behavioural science and medical rehabilitation centres. Henceforth emotion recognition using facial features becomes a practical and fast way for detecting one's mood through various intelligent systems.
    a *Types of facial features:* For the applicability of AER using facial expression information in the real world practices, diverse feature information is taken into consideration to build accurate systems. According to Kim et al. (2017), visual features are classified into three categories as shown in Table 3.

Figure 2 10–20 electrode placement over head (see online version for colours)



Source: Fdez et al. (2021)

Figure 3 Geometric and appearance based facial features (see online version for colours)

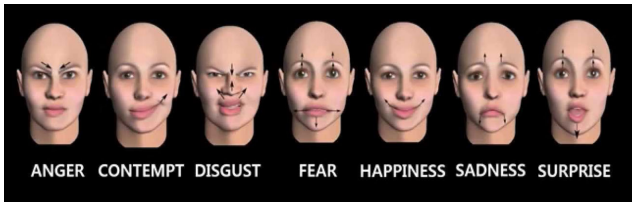


Table 3 Types of visual features

Low level features	*Mid level features*	*High level features*
Visual features derived through facial expressions in the form of colour and texture, also known as handcrafted features.	Mid level features are balance, contrast, harmony, variance, gradation and movement etc.	High level features are detailed semantic content in the image, also known as deep features.

1.1 Organisation

This paper identifies recent machine learning and deep learning methods used for AER using facial and EEG expressions, also opportunities for better work in future. It is organised as Section 2 and 3 provide the detailed information of standard pipeline architecture steps followed by AER using facial expressions, EEG signals. Also, this information is summarised in the tabular form individually. Section 4 presents the recent work done in multimodal AER using facial expressions and EEG signals. Again, a table is provided to investigate the steps followed by multimodal AER. Section 5 discusses the challenges and gives the future directions to researchers in the same field. Finally, Section 6 concludes the overall paper.

2 AER using facial expressions

This section describes the common steps of the standard workflow pipeline used by AER using facial expressions as shown in Figure 6.

2.1 Facial expression databases used

The review work has used following available datasets which have been built under constrained and unconstrained in wild environments for recognising deliberate and spontaneous facial expressions. Below databases (as shown in Table 4) vary from facial images with posed and real expressions to multimodal databases.

FER2013 is a database, built on a large scale under an unconstrained environment for recognising spontaneous emotions. All images have been collected through Google search APIs. Count of training images is 28,709, validation and testing images is 3,589 with seven universal emotions. This database can be used for building real time application systems (Nguyen et al., 2019; Minaee et al., 2021; Melinte and Vladareanu, 2020). Facial expression recognition group (FERG) is an animated dataset generated using MAYA software. It contains 55,767 annotated facial images, purposely designed for cartoonish AER models (Minaee et al., 2021). The Japanese Female Facial Expression (JAFFE) contains 213 posed expressions from 10 Japanese females. It has six basic expressions and one neutral expression (Akhand et al., 2021) extended Cohn-Kanade (CK+) is a laboratory controlled database, containing facial expression frames ranging from neutral expressions to peak expressions (Wei et al., 2020; Minaee et al., 2021; Melinte and Vladareanu, 2020). Abstract dataset contains 279 images (out of which 228 are affective images) of abstract paintings that show both positive and negative emotions in the form of anger, disgust, amusement, contentment, excitement, fear, awe,

sadness. Emotion6 dataset contains 1980 images, retrieved from flickr database and shows six basic emotions. Where Image Emotion Social Net dataset (IESN) is also collected from flickr but a large scale dataset containing 1,012,901 images with more than 15 categorical emotion labels (Zhao et al., 2020). Deliberate expression dataset (MMI) consists of 205 deliberate expression images of front views of the face from 30 subjects with six universal emotions. The sequence of frames it used is onset – apex – offset.

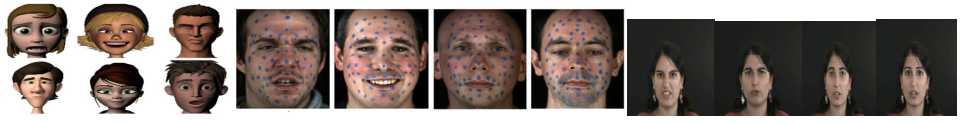
From spontaneous microexpressions, CASME II dataset can be used as it consists of 246 microexpression sequences, ranging from onset—> apex —-> offset frames (Kim et al., 2017). eNTERFACE contains 42 subjects, coming from 14 different nationalities. Subjects those who have participated are from different geographic locations. This dataset may assist to work on culture dependent AER applications (Noroozi et al., 2019).

Table 4 An overview of the facial expression datasets

Dataset	*Origin*	*Details*	*Paper*	*Availability*
Abstract	Machajdik and Hanbury (2010)	228 abstract painting images with combination of colour and texture without recognisable objects with 8 emotion categories. Limitation: Unbalanced no. of emotion classes.	Zhao et al. (2020)	YES
Emotion6	Peng et al. (2015)	1980 images collected from Flickr with 6 basic emotion categories with different distributions rather than only dominant one.	Zhao et al. (2020)	YES
IESN	Zhao et al. (2016)	Image-emotion-social-set, 1,012,901 images collected from text comments on Flickr, with 6 basic categorical emotions and dimensional emotions.	Zhao et al. (2020)	YES
MMI	Valstar and Pantic (2002)	2900 videos and 75 subjects' images, a deliberate dataset which is growing continuously online..	Kim et al. (2017)	YES
FER2013	Dhall et al. (2011), Lucey et al. (2010)	35,685 grey scale images of size 48X48 pixels with six basic emotion categories.	Nguyen et al. (2019), Melinte and Vladareanu (2020), Minaee et al. (2021)	YES
JAFFE	Lyons et al. (1998)	213 images (resolution 256 × 256) of 10 Japanese female subjects with 7 Posed Facial Expressions (6 basic facial expressions + 1 neutral	Melinte and Vladareanu (2020), Minaee et al. (2021), Akhand et al. (2021)	Restricted access
KDEF	Goeleven et al. (2008)	280 face images from 20 male and 20 female with six basic categories of emotions including neutrality.	Akhand et al. (2021), Melinte and Vladareanu (2020)	On request
CK+	Lucey et al. (2010)	593 video sequences from 123 subjects, posed & non-posed facial expressions built under controlled environment.	Ackermann et al. (2016), Melinte and Vladareanu (2020)	On request

Multimodal databases: acted facial expression in wild database (AFEW 7.0) is a multimodal database, collected from media channels. It has been built in different unconstrained environmental conditions (Nguyen et al., 2019). Surrey audio-visual expressed emotion (SAVEE) database consists of recordings from four male actors in seven different emotions, 480 British English utterances in total. The recordings were evaluated by ten subjects under audio, visual and audio-visual conditions (Noroozi et al., 2019).

Figure 4 Sample images from FERG, SAVEE, eNTERFACE (see online version for colours)



Source: Noroozi et al. (2019) and Minaee et al. (2021)

2.2 Preprocessing

As a prerequisite to feature extraction, preprocessing helps to learn meaningful information by removing irrelevant information from input visual content (facial expression images). It is performed by data normalisation, data cleaning, removing the noisy data, data alignment and data augmentation. For input sequence of images, preprocessing is performed by aligning the image using multiple detectors for landmark estimation of visual points. To overcome subjective emotional variances among different images, authors proposed a shared sparse learning (SSL) approach. Here, subjectivity challenges are personal, cultural, personality or situational factors that affect the perception of one's emotions from an image. It used an SSL based discrete probability distribution (DPD) approach to predict dominating as well as participating emotions (Zhao et al., 2020). Next, to avoid illumination variances among input static images, diverse network input (LBP for low level representation robust to illumination variations...), multitask networks, cascaded networks, histogram equalisation, linear mapping and generative adversarial networks (GAN) are used for fine tuning of input data. Whereas Deep FER based dynamic image sequences: frame aggregation is an important task which is performed at decision level feature level. Dynamic image sequences are: non-peak – peak – neutral expressions, captured by deeper cascaded peak networks (Li and Deng, 2020), clustering based strategy works on the clusters of key frames extracted from the videos having similar keyframes altogether based on the distance from centroid, tiny face detector to capture the frames and major frames are clustered together using it (Noroozi et al., 2019; Nguyen et al., 2019). To reduce expression state intraclass variations, normalised cross correlations on onset, apex & offset frames of dynamic video frames were applied (Kim et al., 2017).

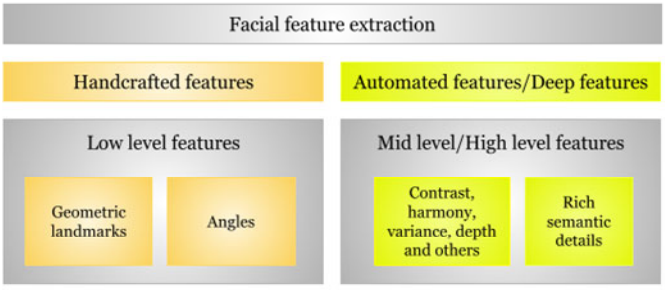
Further for data augmentation using normalisation, mirroring, scaling are applied that also aids to avoid data overfitting. To overcome overfitting issues, pre-trained CNN models, GAN can be applied to augment data. Altogether, to compensate for the challenges like data overfitting, data preprocessing, fine-tuning or other and bringing the proposed methods into practical use, the concept of transfer learning plays a vital role (Melinte and Vladareanu, 2020).

2.3 Feature extraction

Here, key features to be extracted from preprocessed data have been identified as shown in Figure 5. It also explains the state-of-the-art methods used for feature extraction.

- *Handcrafted features:* Majority of the traditional methods (e.g. local binary patterns, geometric feature extraction etc.) were used to extract handcrafted features from facial expression data, also known as shallow learning. Handcrafted features help in recognising categorical emotions (seven basic emotions anger, happiness, sadness, disgust, surprise, fear, joy) only. For geometric visual features, Euclidean distances were calculated for eyes and mouth regions. Further the landmarks obtained, were normalised in order to make it scale invariant. Ten angles and 60 landmark features were extracted in the form of a feature vector (Noroozi et al., 2019). Here, another approach is to learn optimal weight features using weighted multi feature SSL, applied on DPDs extracted from preprocessing phase by Zhao et al. (2020). This method is for combining handcrafted learning features to form multi-features. After using ML methods, researchers suggest switching to neural networks for time dependent data and to increase computational accuracy.
- *Automated features:* Due to existing challenges using shallow learning, trend has been moving to deep features for detailed semantic information for the applicability in real world scenarios. Zhao et al. (2020) suggested applying deep networks [autoencoders, recurrent neural networks (RNN), GAN] for feature learning with softmax loss function. Spatial features of onset, apex offset frames are learnt using CNN and temporal features were learnt using stacked LSTM layered architecture on deliberate and uncontrolled dataset both as by Kim et al. (2017). It was found that it's easier to learn deliberate features than spontaneous ones. For mid level high level feature extraction, multi-level CNN was used with four convolutional layers, two pooling layers and softmax classifier have been designed. Where each convolutional block is connected to a fully connected layer being specific to the level of feature that forms ensemble MLCNN. It also works well for temporal data or time series data (Nguyen et al., 2019). For forming an end to end pipeline, beginning from the camera, face detection continuing to facial emotion recognition, Resnet, inception V3 with Adam optimiser based SSD and faster RCNN models were applied by Melinte and Vladareanu (2020). End to end learning models can be used in various applications, suggested by the authors.

Figure 5 Types of facial features (see online version for colours)



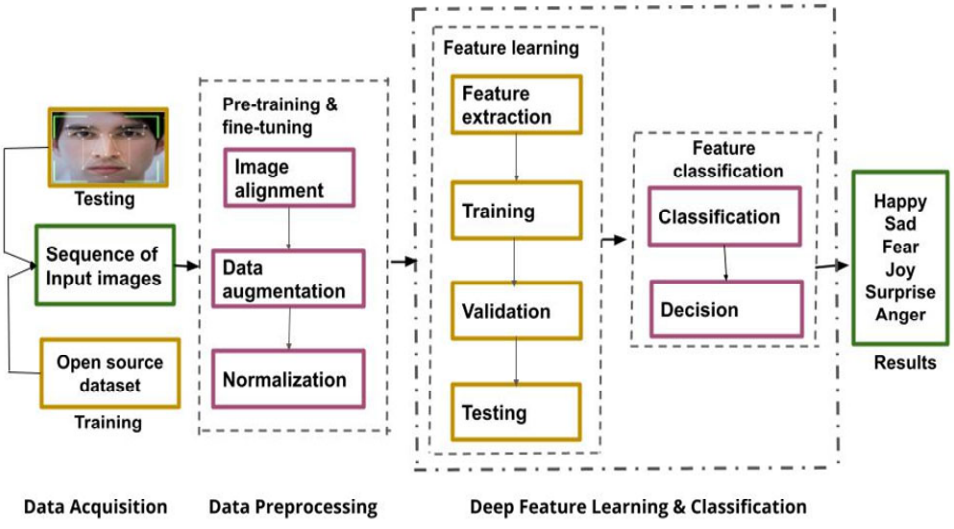
This paper (Akhand et al., 2021) contributes to overcome the existing challenges such as

1 standard CNNs only look after certain features in high resolution images.

2 existing systems are capable of recognising emotions from the frontal views of images, ignoring the profile views and different angles.

AUthors have used a very deep CNN that uses the pretrained models where deep CNN blocks are further pre-trained to attain maximum possible accuracy. Here, the concept of transfer learning also makes the system practically adaptable. Proposed system used the VGG-16 model, pretrained with imagenet dataset. Existing VGG-16 model was again fine tuned with a cleaned emotion dataset and dense layers are replaced by a new dense layer for classification of emotions. The VGG-16 model has five convolutional layers and one fully connected layer where the existing dense layer has been replaced with the new dense layer for classification of emotions in seven categories. Since proposed research is working on profile views (only one eye, one side of face or an ear), it can lead to a number of real time applications.

Figure 6 Standard workflow pipeline for AER using facial expressions (see online version for colours)



2.4 Feature classification

Feature classification is the final step of AER to classify the input testing data into specific emotion categories. After feature learning is performed using ML methods, researchers have used global weighting, k-nearest neighbour, softmax regression, softmax classifier (Zhao et al., 2020; Nguyen et al., 2019; Minaee et al., 2021) to classify the results into different emotion spaces. DL based feature classification sometimes, being independent of additional classifiers as authors have used softmax loss function to calculate the class probabilities and others used support vector machine (SVM) and random forest (RF) to reduce the loss as part of training network architecture itself (Kim et al., 2017; Akhand et al., 2021; Melinte and Vladareanu, 2020). These methods are

deployed at a fully connected layer of network architecture to support end to end learning and achieve significant accuracies. So better to say, deep nets are also used for feature classification along with SVM and softmax classifiers (Lin et al., 2017). Where some deep nets, not supporting end to end learning, used additional classifiers as SVM, RF and softmax classifiers, Gaussian classifiers (Minaee et al., 2021). Here, Multiclass classifier helps to generate confidence matrices for both geometric features and CNN (GoodleNet) visual features (Noroozi et al., 2019).

3 AER using EEG signals

This section describes the common steps of the standard workflow pipeline used by AER using EEG signals as shown in Figure 8.

Table 5 An overview of EEG signal datasets

Dataset	*Origin*	*Details*	*Paper*	*Availability*
DEAP	Koelstra et al. (2012)	32 participants' physiological data and 22 participants' frontal face data on scale of valence, arousal, dominance, liking and familiarity	Zhao et al. (2020)	YES
DREAMER	Katsigiannis and Ramzan (2018)	23 participants' EEG and ECG signals with self assessment ratings on scale of valence, dominance and arousal	Tao et al. (2020), Maheshwari et al. (2021)	On request
DASPS	Baghdadi et al. (2021)	23 participants' anxiety elicited EEG signals	Maheshwari et al. (2021)	YES
SEAD	Liu et al. (2021)	10 males' and 10 females' EEG data with positive, negative and neutral	Fdez et al. (2021)	On request
OpenBCI	https://openbci.com/community/publiclyavailable-eegdatasets/)	Publicly available datasets including DEAP, SEED, motor imagery signals and others	Wang (2018), Bhatti et al. (2019)	...
MAHNOB-HCI	https://mahnob-db.eu/hcitagging/	Multimodal audio, video, EEG and gaze data from 30 participants	....	On request

3.1 EEG databases used

The review work has used available datasets which have been recorded under different environments as shown in Table 5. A 32 channel, DEAP database consists of EEG signals and physiological signals, collected from 32 subjects by showing them different emotional music videos. Recorded emotions are arousal, valence, liking and dominance from level 1 to 9. A multimodal database, DREAMER database contains EEG, ECG signals from 23 subjects (both male female) in the form of valence, arousal and dominance emotions. EEG signals were recorded by an emotiv EPOC device by showing

them 4s video clips (Tao et al., 2020; Ackermann et al., 2016; Li et al., 2016; Huang et al., 2017; Lin et al., 2017; Yang et al., 2018). A 64 channel, MI-EEG Open BCI dataset contains EEG signals from 109 subjects. It was collected through imagination based five brain activities. It's mentioned as the largest dataset among all existing EEG databases (Zhang et al., 2017). A 62 channel SEED dataset consists of EEG signals from 15 subjects by having three sessions on the same film clips. Participants indicated their emotions as positive, negative and neutral. Fdez et al. (2021) mentioned that SEED is superior to DEAP dataset as it contains high quality data information. Maheshwari et al. (2021) used DASPS which is a multi-channel EEG database, collected from 23 participants by exposing them to six different stimuli where some situation is recited first and then recalled. Self assessment is made by participants in given time and classified in terms of valence, arousal and dominance.

3.2 Data preprocessing

EEG signals are recorded by the electrodes through specific channels over the head, and need to be resampled to lower frequency range. Data preprocessing phase, resampled signals are then converted into spectrograms. There are more chances in case of EEG data for the inclusion of noises, artifacts, eye blinking etc. in spectrograms or scalograms which are to be removed for accurate results and we call it artifact filtering.

For non-stationary EEG signals (real time signal processing), It used wavelet transform (sparse representation) to extract the compact structure and high level semantic information. On the other hand, STFT (short time fourier transform) was used to extract information from stationary signals. They had used DB-4 wavelets for decomposing raw channel signals, called CWT (continuous wavelet trans form). After CWT, one dimensional channel signal is transformed into wavelet coefficients based time scale representation. Further, wavelet coefficients are transformed into scalograms. Scalogram and it's energy simply reflects the cognitive process where

- Variations in spectral energy of the scalogram represent cognitive processes.
- It represents the spectral energy percentage against the frequency range of the wavelet.

Frame construction: from scalograms of multi channels, frames are constructed of dimension CxS where C is no. of channels and S is spectral energy contained in a particular time window as:

1 Length of time window is decided as if its 1 sec then for 60 Sec there'll be 60 frames.

2 Within a window, elements of scalogram are added together. That'll generate a vector, representing the energy distribution of scales within a window.

3 Obtained vectors are combined together to form a 2D frame in the current window.

Then all above steps are repeated for the next time window. Then, no. of scales are selected whose spectral energy is high. To check high spectral energy (EER – energy to shannon entropy ratio), shannon entropy was calculated. One having low Shannon entropy, gets high spectral energy, representing the cognitive processes (Yang et al., 2018). Next, for processing nonlinear non-complex signals into intrinsic mode functions,

empirical mode decomposition (EMD) was used. Each IMF component specifies different characteristics of signals at different time scales of the original signal. Further they have employed wavelet transform methods based on EMD to decompose the signals and autoregressive (AR) coefficients obtained to build feature vectors. Further, fast fourier transform (FFT) was used to calculate beta energy bands in order to get the highest energy appearing band time slot for feature extraction (Huang et al., 2017). For specific rhythm or band selection, butterworth fourth order bandpass filters were used by Maheshwari et al. (2021). It mentions that each rhythm has specific spatial and temporal features for every emotion.

Lin et al. (2017) inspired to use end to end learning EEG emotion recognition applications. In the proposed work, for given input data, preprocessing normalised the physiological signals, normalised EEG data was converted into frequency bands in terms of time and frequency domain information.

1 Database channels are firstly downsampled from 512 Hz to 128 Hz. Using bandpass filters, unwanted noises are removed and signals are split into 60s frames each using data normalisation.

2 Based on frequency bands of physiological signals sampling rate of each channel, six images are obtained for each signal generated grey images out of it. Generated images handcrafted features are given as input to the ALexnet model for feature extraction.

Further, Bhatti et al. (2019) emphasised the selection of optimal features and the concept of preserving all information by using overlapping frequency bands. Input EEG signals were cleaned using bandpass filters/averaging spatial filters by removing artifacts such as eye blinking or other undesired movements. Many variants of common spatial patterns (CSPs) indicate that rather than giving frequency gaps between different frequency bands, overlapping frequency bands can be used to preserve all the information. Since it does not remove noise from sub bands, sequential backward floating selection (SBFS) was used to eliminate the noise or feature selection of optimal features.

On the other hand, in paper by Wang et al. (2018), bandpass filtering may lead to loss of important information and result in misclassification. So raw EEG data is normalised using FFT with range from 8Hz–35Hz and segments are fed into 1d-AX rather than bandpass filtering. Further, 1d – aggregate approximation finds mean and slope of frequency values of segments' inter channel class and intra channel class that forms a feature vector out of the given channel. To avoid variations among EEG features of inter subjects for the same class of emotion, Fdez et al. (2021) introduced stratified normalisation over batch normalisation. Stratified normalisation helped in reducing inter subject variability and also outperformed other methods for participant independent tasks.

3.3 Feature extraction

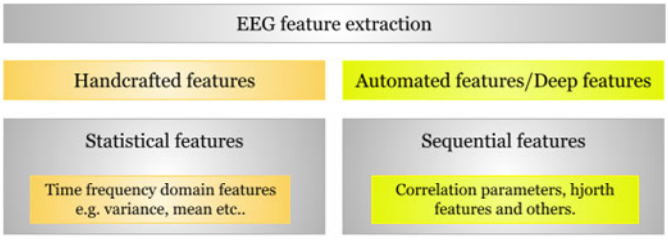
Here, key features to be extracted from preprocessed data have been identified as shown in Figure 7. It also explains the state-of-the-art methods used for feature extraction.

- *Handcrafted features:* Out of all frontal, temporal and central areas, researchers, Huang et al. (2017) used F3, F4 channels only to recognise EEG based emotions. Firstly, for feature extraction, empirical wavelets were applied on segments as bandpass filters. Further, using wavelet transform equation coefficient parameters

were generated. As a second part, the AR model was employed on generated parameters to calculate coefficients that form a feature vector further for classification of results. Here, researchers (Bhatti et al., 2019) have worked on left right brain motor imagery channels. For the same, features are extracted using CSP features from attained sub bands. Then, linear discriminant analysis score is calculated using the cost function of CSP features. LDA score for each band is used as a feature and further optimised. SBFS is used to extract an optimal set of bands (features attained as LDA score) by keeping a desired length of features in the feature vector for classification ahead.

- *Automated features:* C-RNN was implemented on EEG data received from the preprocessing step by Li et al. (2016). Here, C-RNN generated the labels, depending upon the output of every layer unlike many to one approach. It led to the consideration of the entire experience of participants in the trial. Here, customised size of filters checked for the correlation among different channels as well as scales. They had worked to find correlation among different channels by setting appropriate filter sizes and pooling sizes. To feed the feature vectors into LSTM, it's flattened by getting it into one dimensional vector. At RNN stage- RNN is specifically having the ability to learn features from time series data. Here, LSTM is used as an RNN unit. LSTM with RNN aids to extract contextual information from feature sequence and output of LSTM at every layer. It resulted in 12 hand engineered features such as mean, variance, root mean square, peak to peak value, hjorth features(mobility, complexity and activity), correlation parameter and shannon entropy.

Figure 7 Types of EEG features (see online version for colours)



Next, Channel weighting is applied (on two feature vectors i.e. mean and variance) as a group of spatial filters mounted in hidden layers to extract useful information gained from Wang et al. (2018). Now, these two channel weighted feature vectors are given as an input into back propagation based LSTM. It proved superiority among other existing approaches. Also, it concluded that faster the real time processing of data, less the overfitting will be. Further by Fdez et al. (2021), feature extraction is performed using differential entropy with alpha, beta and gamma frequency bands. Delta was not included because it is only related to the sleeping state of the brain. Gaussian distribution was used in differential entropy to extract features of EEG signals by knowing continuous complexity of a signal. It retained the feature information layer to layer of the neural network so also assisted in participant identification. Accuracy achieved for binary classification was higher than ternary classification, though it overpassed the existing approaches for the case of ternary classification too (Fdez et al., 2021).

Deep CNN has been used where a number of filters are used at every layer to optimise the extracted features (Maheshwari et al., 2021). It evaluated the results using 10 cross fold validation, also calculated specificity, accuracy, sensitivity, F1-score and kappa values for all the channels, saying multichannel, alpha, beta, delta, gamma and theta rhythms of a signal used.

Some authors have worked using transfer learning as Lin et al. (2017) used multimodal deep CNN (pre trained Alexnet) with five convolutional layers, one fully connected layer (500 hidden units) as feature learning and proved the effectiveness of deep CNN over other existing work.

3.4 Feature classification

Classification of features is performed to achieve the results for 2D (valence arousal) or 3D (valence, arousal and dominance). As per the threshold value fixed, results can be seen in the form of low/high e.g. low valence, high valence, low arousal, and so on. End to end learning of image EEG samples (fed into CNN as input) was performed by Lin et al. (2017), Fdez et al. (2021), and Maheshwari et al. (2021). Fdez et al. (2021) have chosen binary (positive, negative) and ternary (positive, negative and neutral) classification emotion classes and resulted in accuracy achieved for binary classification was higher than ternary classification. Further, they used the k-means clustering algorithm for achieving classes' k-mean value against one subject. Along with, they also deployed a softmax function at a fully connected layer. That falls under high/low categories of valence and arousal. And Maheshwari et al. (2021) worked on rhythm specific emotion recognition using Deep CNN, gave an accuracy of 98% for selected rhythms in all three categories as low or high for valence, arousal and dominance. Bhatti et al. (2019) used radial basis function neural network (RBFNN) as a 2-layer network where the hidden layer was deployed using gaussian radial mathematical regression function as classifier. Further, for EEG emotion classification, softmax regression was used by Wang et al. (2018). Experimentation purpose, proposed 1d-AX, channel weighting with LSTM (2 hidden layers) and softmax regression were compared with no. of CSPs and other deep learning approaches, henceforth, it proved superiority among other existing approaches. In other papers by practitioners (Zhang et al., 2017; Huang et al., 2017; Yang et al., 2018), SVM checks for the different classes using linear or nonlinear hyperplanes with gaussian kernel function to classify the emotions.

In summary (as shown in Figure 8), the recorded data in the form of EEG signals is preprocessed using different approaches as mentioned in part (3.2) of AER using EEG signals to produce different frequency bands which are free from noise and artifacts. Next, preprocessed data is fed into the feature extraction phase of AER network architecture. Further, machine learning or deep learning approaches are used for extracting hand engineered or deep spatiotemporal (spatial and temporal) features respectively. Henceforth, feature classification of data is performed using different types of classifiers [as mentioned in part (3.4) of AER using EEG signals] to achieve the results in the form of 2D (valence and arousal) or 3D (valence, arousal and dominance).

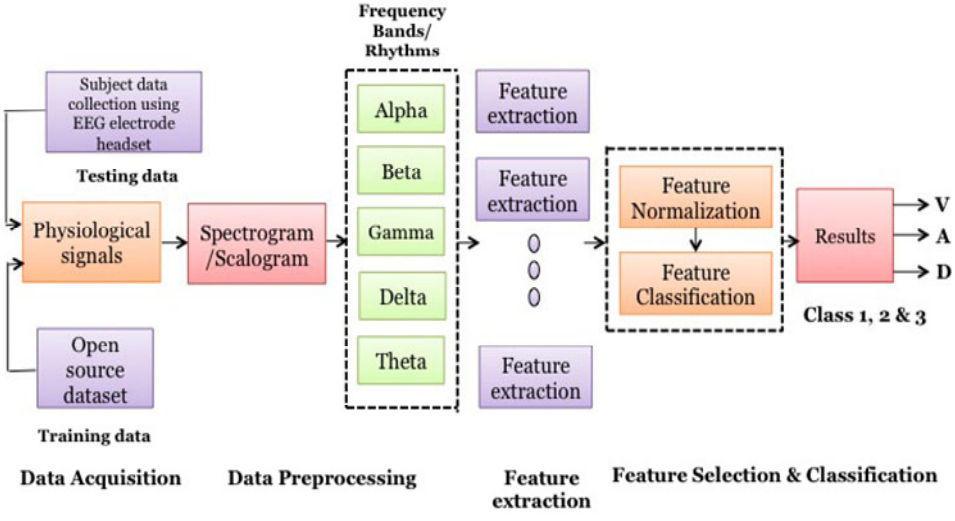
EEG influential features drawn by the authors are (Fdez et al., 2021):

1 high values of alpha and beta powers in the frontal and right parietal lobe respectively result in high valence (positive emotion) and vice versa

2 higher beta power and lower alpha value leads to excitation state i.e. arousal

3 increase in the values of beta/alpha ratio in the frontal lobe and beta in parietal lobe becomes the strength (dominance) of a signal or emotion.

Figure 8 Standard workflow pipeline for AER using EEG signals (see online version for colours)



4 Multimodal AER using facial expressions and EEG signals

Practitioners have put efforts to increase the accuracy of AER by fusing facial expressions and EEG signals. Rapid advancement in the field of ML and information fusion methods has made it possible using multimodality to recognise human emotions.

Network architecture for feature learning and classification: Corneanu et al. (2016) proposed a multimodal emotion recognition system using EEG signals and facial expressions. They adopted the concept of transfer learning to reduce the training and processing time. As CNN weights had been taken from the previously trained model which helped in reducing the processing time. Feature extraction was done using 3D-CNN and Mask RCNN along with ensemble encoding learning for EEG and facial expressions respectively. SVM classifier was used to attain the results on a trained dataset. Further, Huang et al. (2019) worked on fusion of EEG and facial expression data to analyse valence and arousal in binary form for spontaneous expressions. Online experimentation using the proposed system, 13 subjects for 40 trials data had been collected through camera and emotiv mobile devices. Self assessment manikins (SAM) approach was followed as after watching each movie, the subject was asked to submit his level of valence arousal for emotion ranging from 0–9. Pretrained CNN was used to recognise spontaneous facial expressions as transfer learning where testing was done using two datasets: DEAP, MAHNOB-HCI (a multimodal dataset, collected from 27 subjects by showing them emotional videos. Visual, acoustic and physiological signals were recorded in response to affective stimuli. Subjects assessed themselves on emotions: valence, arousal and dominance. For EEG valence arousal detection, raw electrode data was filtered using wavelet transform and further power spectral density (PSD) features were extracted for 14 electrode signal data. Recursive feature eliminations were made for

feature selection. SVM classifies the testing data into low high valence and arousal respectively. Then those SAMs were used against the results obtained from the proposed model for checking the accuracy of the model. Since continuous emotion recognition is a demand of certain HCI applications, So, Li et al. (2019) have made contributions by fusing electroencephalography and facial expressions to recognise the emotions continuously. For EEG signal preprocessing

1 EEGLab in MATLAB was used to preprocess the original signal by applying bandpass filters ranging from 4–47Hz to reduce the artifacts such as eye blinking, face movements etc. Further, independent component analysis (ICA) was applied to get 14 components against 14 electrode channels. EEG data can also be seen in spatial domain by using ICA, then for feature extraction it has to be reconstructed. Since EEG data is non-stationary, STFT is applied to get time and frequency domain segments (stationary) or bands. Hamming window is selected to get short time windows for PSD feature extraction. PSD feature extraction is done separately for all alpha, beta, theta and gamma bands.

2 To remove irrelevant features, LDA and t-stochastic neighbour embedding (SNE) were used for dimensionality reduction. For facial expression, geometric features for mouth corners, eyes, eyebrows etc. have been detected as in 29 landmark coordinates.

3 Individual (EEG and face) valence arousal prediction from SVR (support vector regression) were fed into LSTM for final valence result every time period. Experimentation was conducted on ten subjects by showing movie clips as stimulus from the SEED dataset using emotiv 14 channel headcap and mobile phone camera. It's learnt that T7, T8 electrode channels are most relevant to gamma and beta correlation resp., O1 O2 is for theta correlation coefficient and P8 for alpha band correlation.

Multimodal fusion of facial and EEG signals was implemented at two different levels of fusion (Huang et al., 2017). Facial expressions had been taken as in the form of frames from video dataset. Experimentation was carried out using MAHNOB-HCI dataset for both facial expressions' videos and EEG signals (only two channels were used for emotion recognition as F3 C4 out of 32 channel signals) for emotion recognition. Neural aggregation network (NAN) with CNN had been applied for feature extraction. For EEG signals, frequency features were extracted using linear frequency cepstral coefficients (LFCCs) and complexity using sample entropy features (SampleEn). Before LFCC feature extraction, FFT was applied to get spectrum as preprocessing of a signal. Probabilities were calculated using IMFs for independent components to get the probable values further for fusion to get the final results.

To get the discrete emotion classes with different intensity levels, Li et al. (2019) focussed on two types of fusion methods for multimodal. Face expression feature extraction was performed with a neural network by deploying one hidden layer to classify the emotions as happy, neutral, sad and fear. Other hand, EEG signal feature extraction was carried out using the SVM classifier. Four emotion states of facial expressions along with three intensity levels i.e. strong, medium and low were recognised by conducting two experiments on 20 healthy subjects under movie stimuli. As a future perspective, practitioners suggest improving data training by feeding more input data.

Table 6 Summarisation of analysis of AER using facial expressions

Author year	*Visual features used*	*Preprocessing approach*	*ML domain*	*DL domain*	*Additional classifier*	*Classes*	*Performance*	*Inferences*
Kim et al. (2017)	Learning-based features	Normalised cross correlations for onset, apex & offset frames	--	CNN spatial features followed by LSTM temporal features	--	Angry, disgust, fear, happy, sad, surprise	Recognition accuracy for MMI dataset – 78.61%, CASME II dataset – 60.98%	It can be trained on real time data with no annotations and better fine tuned LSTM to form the applicability.
Nguyen et al. (2019)	Learning based features (high level & mid level)	Tiny face detector, Face clustering & selecting algorithm	--	Multi level CNN	Softmax classifier	Angry, disgust, fear, happy, sad, surprise, neutral	Accuracy on FER2013 – 74.09%, Accuracy on AFEW 7.0 – 49.3%	Preprocessing may be required to improve the accuracy. Also, multimodality can be used to cope with imbalanced data as stated by Akhand et al. (2021).
Akhand et al. (2021)	Learning based features (low level, mid level and complex features)	Fine tuning using emotion data	--	Pretrained VGG-16 model with replaced dense layer and fine tuning of each convolutional block forming a very deep CNN	--	Afraid, angry, disgusted, happy, sad, surprise, neutral	accuracy of 98.78% on KDEF dataset and 100% on JAFFE dataset	Optimising the features and multimodality can help working on real time applications.
Minaee et al. (2021)	Learning based features	--	--	Attentional CNN with grid generator to produce warped data	Softmax classifier	Anger, disgust, fear, happy, sad, surprise, neutral	accuracy of 70.02% on FER 2013, 99.3% on FERG, 92.8% on JAFFE and 98% on CK+	Region of interest based CNN may increase the accuracy on the FER2013 dataset as stated in Melnte et al. (2020). Decreasing the number of CNN layers can help reduce the inference speed too.

Table 6 Summarisation of analysis of AER using facial expressions (continued)

Author year	*Visual features used*	*Preprocessing approach*	*ML domain*	*DL domain*	*Additional classifier*	*Classes*	*Performance*	*Inferences*
Melinte and Vladareanu (2020)	Learning -based features	Pretrained on VGG, Resnet, inception V3	--	Faster R-CNN, single shot detector CNN	--	Anger, disgust, fear, happy, sad, surprise, neutral	SSD inception model accuracy of 97.42% and faster R-CNN accuracy of 97.8%	Transfer learning can be practiced on different variants of pre-trained CNNs for real time applicability. Multimodality is another important parameter to be considered for it.
Noroozi et al. (2019)	Geometric and visual features	--	Euclidean distances for geometric features	Googlenet CNN for visual features	Multiclass SVM and random forest	Angry, disgust, fear, happy, sad, surprise	98.10%, 98.33% and 98.89% using SVM with PCA for SAVEE, eNTERCAE'05 and RML	Performance of multimodality may get improved for real time data by practicing different fusion schemes as used by Huang et al. (2017)
Ackermann et al. (2019)	Low level empirical, high level self-learning and multimodal features	--	--	CNN	SVM		Accuracy – 94.41%	In place of combining whole face features, key area features can be explored as shown in Figure 2.

Table 7 Summarisation of analysis of AER using EEG signals

Author paper	*EEG channels used*	*Preprocessing approach*	*Proposed network architecture ML domain*	*DL domain*	*Additional classifier*	*Discrete*	*Dimensional*	*Performance*	*Inferences*
Tao et al. (2020)	F3, FC5, P3, P7, F8,Cz, C4, P8,T8, F8	Removing baseline signals and 3-s sliding window	--	Attention based convolutional recurrent neural network (ACRNN)	--	--	Valence, arousal, dominance	Accuracy on DEAP Database, for V and A: 93.72% and 93.38% DREAMER Database, for VAD: 97.93%, 97.78%, 98.23%	It can have trials on more lobe channels with different subjects to check the performance of the system.
Ackermann et al. (2016)	AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8 and AF4	Butterworth filter and hamming window of size 1s	STFT features with PSD, HOC, HHS and mRMR feature selection	--	Random forest and support vector machine	--	Positive, negative	Recognition is improved using RF than SVM.	Gamma features, prefrontal and left temporal lobe EEG channel locations play an important role in recognising emotions. Different frequency bands of EEG signals' contribution can be better checked on continuous data.
Li et al. (2016)	32 channel data	Continuous wavelet transform DB-4 wavelet followed by STFT features	--	C-RNN LSTM	--	--	Valence and arousal	Accuracy for valence – 72.06% arousal – 74.12%	Automated features with 12 hand engineered extracted features can work better for real time data.
Zhang et al. (2017)	64 channel data	Sliding window	--	Cascaded and parallel RCNN-LSTM	Softmax classifier, SVM	Multiclass	--	Accuracy on both models, i.e., parallel and cascaded RCNN – 98.1% and 96.6% resp.	Proposed preprocessing approaches can be enhanced to work on missing values of input data.

Table 7 Summarisation of analysis of AER using EEG signals (continued)

Author paper	*EEG channels used*	*Preprocessing approach*	*Proposed network architecture ML domain*	*DL domain*	*Additional classifier*	*Discrete*	*Dimensional*	*Performance*	*Inferences*
Huang et al. (2017)	F3,F4	Empirical mode decomposition for Intrinsic mode functions (IMFs), sliding window	Empirical wavelet transform followed by autoregressive	--	SVM using Gaussian kernel function	--	Valence and arousal	Arousal – 64.3% and valence – 67.3%.	It has used the highest energy beta band for feature extraction. But more channels can be involved to get more inferences.
Lin et al. (2017)	40 channel data	Bandpass filters	--	Pretrained Alexnet model	K-means clustering	--	Valence and arousal	Accuracy and F1 score for arousal as 87.30% and 78.24% and for valence 85.50 &anda 80.06%	It describes handcrafted features for various peripheral signals such as eye blink rate, minima value, blood volume pressure features etc. which may be useful for preprocessing raw data in future too.
Yang et al. (2018)	8 channels out of 32 channel headset	Frequency pattern decomposition	--	Continuous-CNN	SVM	--	Valence and arousal	Arousal and valence as 90.24% and 89.45%	Combination of different channel features may improve the accuracy with respect to no reaction brain signals.
Bhatti et al. (2019)	14 and 118 channel	Bandpass filters ranging from 8–32 Hz and 6–30 Hz	Common spatial patterns followed by their linear discriminant analysis score and further optimised by sequential backward floating selection	--	Radial basis functional neural network	--	Motor imagery signals	Real time environment accuracy – 85% and with emotiV dataset – 93%	There may be chances of important features getting removed while noise removal from real time data. So, optimal feature selection approaches can be used here.

Table 7 Summarisation of analysis of AER using EEG signals (continued)

Author paper	*EEG channels used*	*Preprocessing approach*	*Proposed network architecture ML domain*	*DL domain*	*Additional classifier*	*Discrete*	*Dimensional*	*Performance*	*Inferences*
Wang et al. (2018)	--	Fast Fourier transform with range from 8 Hz–35 Hz followed by 1d-aggregate approximation by inputting segment data	--	LSTM-RNN	Softmax regression	--	Motor imagery signals	Accuracy – 76.47%	Risk of overfitting can be reduced by training the model with offline EEG data and testing on real time online data.
Fdez et al. (2021)	62 channel data	--	--	Differential entropy using Gaussian distribution, CNN using stratified normalisation	--	Positive, negative, neutral	--	Accuracy for binary (positive and negative) – 91.6% ternary (positive, negative and neutral) – 79.6%	It is observed that stratified normalisation helps well in EEG AER with inter subject data comparatively. That can also be deployed in other AER systems.
Maheshwari et al. (2021)	AF3, F7, F15, T7, P7, O1, O2, P8, FC6, F4, F8, AF4, FC2, FP2, T8	Butterworth forth order bandpass filter	--	Deep CNN with optimisation property		--	Valence, arousal and dominance	Valence – 98.56%, arousal – 98.82% and dominance – 98.99%	Optimisation of features by deploying a number of filters at CNN layers can be used for real time AER systems.
Bin et al. (2019)	Fp1, Fp2, F7, F3, Fz, F4, F8, T3, T4, T5, T6, C3, Cz, C4, P3, Pz, P4, O1, O2	Butterworth bandpass filter	Power spectral density	--	K-nearest neighbour, linear discriminant analysis	--	--	Highest accurate channels: P3, T6 – 99.44%, parietal region – 98.33% (alpha band), 95.56% (beta band), temporal lobe – 98.20% (alpha frequency band), 92.92% (beta frequency band)	Observed accurate channels can be an application specific based on EEG signal AER systems as a future inference.

Table 8 Summarisation of analysis of AER using facial expressions and EEG signals

Author year	*Network architecture for EEG signals*	*Network architecture for facial expressions*	*Training dataset*	*Transfer learning*	*Additional classifier*	*Data level fusion*	*Feature level fusion*	*Decision/score level fusion*	*Testing data*	*Performance with classes*
Abdullah et al. (2021)	3D CNN	Mask RCNN with ensemble encoding	Pretrained model	TRUE	SVM	Stacking and bagging	--	--	DEAP	Valence (96.13%) and arousal (96.79%)
Huang et al. (2019)	PSD features	CNN	Pretrained model	TRUE	SVM	--	--	Enumerated and AdaBoost fusion	Equipment data (online experimentation)	Valence (69.75%) and arousal (70%)
Zhang et al. (2021)	CNN-LSTM	1D CNN	DEAP and MAHNOB-HCI	FALSE	--	--	--	--	DEAP and MAHNOB-HCI	Valence and arousal (99.22%)
Fang et al. (2021)	LFCC and SampleEN	NAN-CNN	DEAP and MAHNOB-HCI	FALSE	SVM	--		Bayesian function	DEAP and MAHNOB-HCI	Valence (68.30/69.21) and arousal (66.73/69.38)
Huang et al. (2017)	PSD features	feed forward CNN	Equipment data	FALSE	SVM	--	--	Sum and production rule	Equipment data	Four classes with three intensity levels (81.25% and 82.75%)

- *Fusion methods:* To take advantage of complimentary information of facial expressions and EEG signal data to enhance AER performance, multimodalities are combined using different types of fusion schemes as mentioned below:
  1 Corneanu et al. (2016) worked on Data level and score level fusion strategies using two approaches stacking and bagging. For data level fusion, EEG chunks and facial expression video chunks with equal time distribution have been fed to the system.
  2 Researchers worked on score level fusion, where scores as individual results are fused together to attain the final score for classes of emotions (Corneanu et al., 2016). Further, Huang et al. (2019) used enumerated fusion and Adaboost (adaptive boosting) methods to attain the scores for valence and arousal binary classification.
  3 Decision level fusion by Zhang et al. (2021) using LSTM yielded the results as (0.625+–0.029) with concordance correlation coefficients (CCC) for facial and EEG modalities. Also, decision level fusion has been performed using sum and product rule. EEG signals and face expressions were fused together to form a multimodal system (Li et al., 2019).

4.1 Summarisation

We have summarised a large number of papers to provide insights into AER using facial expressions and EEG signals. The overall analysis has been encapsulated in Table 6, Table 7 and Table 8 for AER using facial expressions, EEG signals and both respectively.

5 Discussion

During the literature survey, the standard pipeline to recognise the emotions using facial expressions is observed as data acquisition, preprocessing, feature learning and result classification. As shown in Table 6, most of the researchers have used learning based automated visual features over handcrafted features because the automated features are capable of extracting semantic or detailed information too along with other low level and mid level features. For extracting semantic information, deep learning network architectures are best suited over machine learning methods. Deep learning neural networks, its variants that may be RNN-LSTM or pretrained CNN models such as Alexnet, Googlenet give best results among all existing static methods. Also, additional classifiers are not required in case of deploying deep learning models for AER using facial expressions as fully connected layers solve the purpose of data classification using SVM and softmax classifiers mainly. In spite of exponential growth in AER using facial expressions, still some challenges need to be addressed as (Corneanu et al., 2016): improving the algorithms to deal with naturalistic environments robustly. Also, it's vital to deal with the concerns existing in the input data such as large rotations, occlusions and multiple persons. On the other hand, AER using EEG signals is also a booming area and is preferred for cognitive research over other physiological signals because they have better resolution than other physiological signals (Maheshwari et al., 2021). It's observed that the preprocessing phase is significantly important because EEG signals contain irrelevant information that may hamper the final results. So, using the sliding window to

remove baseline signals, bandpass filters are good ways to remove unnecessary information from input data. As summarised in Table 7, the majourly used network model is RNN-LSTM, a time variant CNN because the EEG signal is a both spatial and temporal domain associated time series data. Also, its effectiveness can be seen to work on lengthy and variable physiological signal data. Table 7 investigates the commonly and rarely used channels to recognise the emotion dimensions as in binary form. Still some major concerns in EEG emotion recognition are insufficient training data available to work in a real time environment. Though some papers have worked using a number of filters at each layer to reduce the feature complexity or optimising features, still optimisation of features needs to be focussed more by researchers to increase the efficiency of the system (Maheshwari et al., 2021). Apart from this, for clear cognitive understanding, only hand engineered features isn't sufficient, in depth features of EEG signals should be extracted too and finding the most contributive channels of EEG signals is also a future concern (Li et al., 2016). Further, multiple physiological and non-physiological modalities can be fused together to build strong AER systems. Since, only discrete emotion classes are not sufficient to describe one's complete state of mind. Dimensional values of emotions in the form of valence, arousal and dominance also describe the state of elicited emotion. Based on the multimodal AER using face and EEG signal research, fusing visual features and EEG signals lead to concrete and objective outputs in terms of discrete and dimensional emotions both. Where fusion approaches to combine different modalities is still a wide area of research and researchers can foresee it as one of the future directions in multimodal AER.

6 Conclusions and future work

This paper presented recent work in AER using facial expressions and EEG signals, also latest developments in this area using machine learning and deep learning techniques. We have presented experimental workflow in terms of preprocessing approaches, network architectures, unconstrained or controlled datasets, classifiers, fusion scheme, results achieved in tabular form for facial emotion recognition, EEG emotion recognition and multimodal emotion recognition. Also, it's explained that emotions are not only limited to seven basic classes but dimensions of emotions can also be equally important part of it to be recognised. Henceforth, practical adaptability of AER needs real time high performance systems. It will push researchers to build large uncontrolled or unconstrained datasets, experimentation on exploring deep learning approaches, combination of multiple modalities (physiological or non-physiological), real time data for testing of systems for ideal recognition of human emotions.

References

Abdullah, S.M.S., Ameen, S.Y.A., Sadeeq, M.A. and Zeebaree, S. (2021) 'Multimodal emotion recognition using deep learning', *Journal of Applied Science and Technology Trends*, Vol. 2, No. 2, pp.52–58.

Ackermann, P., Kohlschein, C., Bitsch, J.A., Wehrle, K. and Jeschke, S. (2016) 'EEG-based automatic emotion recognition: Feature extraction, selection and classification methods', *2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom)*, pp.1–6.

Akhand, M., Roy, S., Siddique, N., Kamal, M.A.S. and Shimamura, T. (2021) 'Facial emotion recognition using transfer learning in the deep CNN', *Electronics*, Vol. 10, No. 9, p.1036.

Baghdadi, A., Aribi, Y., Fourati, R., Halouani, N., Siarry, P. and Alimi, A.M. (2021) *Dasps Database*, https://doi.org/10.21227/barx-we60.

Bhattacharyya, A., Tripathy, R.K., Garg, L. and Pachori, R.B. (2020) 'A novel multivariate-multiscale approach for computing eeg spectral and temporal complexity for human emotion recognition', *IEEE Sensors Journal*, Vol. 21, No. 3, pp.3579–3591.

Bhatti, M.H., Khan, J., Khan, M.U.G., Iqbal, R., Aloqaily, M., Jararweh, Y. and Gupta, B. (2019) 'Soft computing-based eeg classification by optimal feature selection and neural networks', *IEEE Transactions on Industrial Informatics*, Vol. 15, No. 10, pp.5747–5754.

Bin, N.W., Awang, S.A., Fook, C.Y., Chin, L.C. and Ying, O.Z. (2019) 'A study of informative EEG channel and brain region for typing activity', *Journal of Physics: Conference Series*, Vol. 1372, No. 1, p.012008, https://doi.org/10.1088/1742-6596/1372/1/012008.

Chaudhari, P., Kulkarni, N. and More, P. (2021) 'Antispoofing for facial recognition through separable convolution', *2021 IEEE Pune Section International Conference (PuneCon)*, pp.1–4, https://doi.org/10.1109/PuneCon52575.2021.9686544.

Corneanu, C.A., Sim´on, M.O., Cohn, J.F. and Guerrero, S.E. (2016) 'Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: history, trends, and affect-related applications', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 38, No. 8, pp.1548–1568, https://doi.org/10.1109/TPAMI.2016.2515606.

Dhall, A., Goecke, R., Lucey, S. and Gedeon, T. (2011) 'Static facial expression analysis in tough conditions: data, evaluation protocol and benchmark', *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pp.2106–2112, https://doi.org/10.1109/ICCVW.2011.6130508.

Fang, Y., Rong, R. and Huang, J. (2021) 'Hierarchical fusion of visual and physiological signals for emotion recognition', *Multidim. Syst. Sign. Process*., Vol. 32, pp.1103–1121, https://doi.org/10.1007/s11045-021-00774-z.

Fdez, J., Guttenberg, N., Witkowski, O. and Pasquali, A. (2021) 'Cross-subject EEG-based emotion recognition through neural networks with stratified normalization', *Frontiers in Neuroscience*, Vol. 15, p.626277.

Goeleven, E., De Raedt, R., Leyman, L. and Verschuere, B. (2008) 'The Karolinska directed emotional faces: a validation study', *Cognition & Emotion*, Vol. 22, No. 6, pp.1094–1118.

Huang, D., Zhang, S. and Zhang, Y. (2017) 'EEG-based emotion recognition using empirical wavelet transform', *2017 4th International Conference on Systems and Informatics (ICSAI)*, pp.1444–1449.

Huang, Y., Yang, J., Liu, S. and Pan, J. (2019) 'Combining facial expressions and electroencephalography to enhance emotion recognition', *Future Internet*, Vol. 11, No. 5, p.105.

Jackson, P. and Haq, S. (n.d.) *Surrey Audio-Visual Expressed Emotion (Savee) Database*.

Katsigiannis, S. and Ramzan, N. (2018) 'Dreamer: a database for emotion recognition through EEG and ecg signals from wireless low-cost off-the-shelf devices', *IEEE Journal of Biomedical and Health Informatics*, Vol. 22, No. 1, pp.98–107, https://doi.org/10.1109/JBHI.2017.2688239.

Kaur, S. and Kulkarni, N. (2021a) 'A deep learning technique for emotion recognition using face and voice features', *2021 IEEE Pune Section International Conference (PuneCon)*, pp.1–6, https: //doi.org/10.1109/PuneCon52575.2021.9686510.

Kaur, S. and Kulkarni, N. (2021b) 'Emotion recognition – a review', *International Journal of Applied Engineering Research*, Vol. 16, No. 2, pp.103–110, ISSN 0973-4562.

Kim, D.H., Baddar, W.J., Jang, J. and Ro, Y.M. (2017) 'Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition', *IEEE Transactions on Affective Computing*, Vol. 10, No. 2, pp.223–236.

Koelstra, S., Muehl, C., Soleymani, M., Lee, J-S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A. and Patras, I. (2012) 'DEAP: a database for emotion analysis using physiological signals (PDF)', *IEEE Transactions on Affective Computing*, Vol. 3, No. 1, pp.18–31.

Koole, S.L. (2009) 'The psychology of emotion regulation: an integrative review', *Cognition and Emotion*, Vol. 23, No. 1, pp.4–41, https://doi.org/10.1080/02699930802619031.

Li, D., Wang, Z., Wang, C., Liu, S., Chi, W., Dong, E., Song, X., Gao, Q. and Song, Y. (2019) 'The fusion of electroencephalography and facial expression for continuous emotion recognition', *IEEE Access*, Vol. 7, pp.155724–155736, iNSPEC Accession Number 19087757.

Li, S. and Deng, W. (2020) 'Deep facial expression recognition: a survey', *IEEE Transactions on Affective Computing*, arXiv: 1804.08348

Li, X., Song, D., Zhang, P., Yu, G., Hou, Y. and Hu, B. (2016) 'Emotion recognition from multi-channel eeg data through convolutional recurrent neural network', *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp.352–359.

Lin, W., Li, C. and Sun, S. (2017) 'Deep convolutional neural network for emotion recognition using eeg and peripheral physiological signal', *International Conference on Image and Graphics*, pp.385–394.

Liu, W., Qiu, J-L., Zheng, W-L. and Lu, B-L. (2021) 'Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition', *IEEE Transactions on Cognitive and Developmental Systems*, Vol. 14, No. 2, pp.715–729.

Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z. and Matthews, I. (2010) 'The extended cohn-kanade dataset (ck+): a complete dataset for action unit and emotion-specific expression', *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition – Workshops*, pp.94–101, https://doi.org/10.1109/CVPRW.2010.5543262.

Lyons, M.J., Kamachi, M. and Jiro, G. (1998) *The Japanese Female Facial Expression (Jaffe) Dataset* [online] https: //zenodo.org/record/3451524#.Yh5iQC8RpQI (accessed 6 May 2022).

Machajdik, J. and Hanbury, A. (2010) 'Affective image classification using features inspired by psychology and art theory', in *Proceedings of the 18th ACM International Conference on Multimedia*, pp.83–92.

Maheshwari, D., Ghosh, S., Tripathy, R., Sharma, M. and Acharya, U.R. (2021) 'Automated accurate emotion recognition system using rhythm-specific deep convolutional neural network technique with multi-channel EEG signals', *Computers in Biology and Medicine*, Vol. 134, p.104428, ISSN: 0010-4825, https://doi.org/10.1016/j.compbiomed.2021.104428.

Mauss, I.B. and Robinson, M.D. (2009) 'Measures of emotion: a review', *Cognition and Emotion*, Vol. 23, No. 2, pp.209–237.

Melinte, D.O. and Vladareanu, L. (2020) 'Facial expressions recognition for human–robot interaction using deep convolutional neural networks with rectified Adam optimizer', *Sensors*, Vol. 20, No. 8, p.2393.

Minaee, S., Minaei, M. and Abdolrashidi, A. (2021) 'Deep-emotion: facial expression recognition using attentional convolutional networks', *Sensors*, Vol. 21, No. 9, p.3046.

Nguyen, D.H., Kim, S., Lee, G-S., Yang, H-J., Na, I-S. and Kim, S.H. (2019) 'Facial expression recognition using a temporal ensemble of multi-level convolutional neural networks', *IEEE Transactions on Affective Computing*, Vol. 10, No. 1, pp.4–6.

Noroozi, F., Marjanovic, M., Njegus, A., Escalera, S. and Anbarjafari, G. (2019) 'Audio-visual emotion recognition in video clips', *IEEE Transactions on Affective Computing*, Vol. 10, No. 1, pp.60–75, https: //doi.org/10.1109/TAFFC.2017.2713783.

Pampouchidou, A., Simos, P.G., Marias, K., Meriaudeau, F., Yang, F., Pediaditis, M. and Tsiknakis, M. (2019) 'Automatic assessment of depression based on visual cues: a systematic review', *IEEE Transactions on Affective Computing*, Vol. 10, No. 4, pp.445–470, https://doi.org/10.1109/TAFFC.2017. 2724035.

Peng, K-C., Chen, T., Sadovnik, A. and Gallagher, A. (2015) 'A mixed bag of emotions: model, predict, and transfer emotion distributions', *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.860–868, https://doi.org/10.1109/CVPR.2015.7298687.

Poria, S., Cambria, E., Bajpai, R. and Hussain, A. (2017) 'A review of affective computing: from unimodal analysis to multimodal fusion', *Information Fusion*, Vol. 37, pp.98–125, https://doi.org/ https://doi.org/10.1016/j.inffus.2017.02.003.

Salama, E.S., El-Khoreibi, R.A., Shoman, M.E. and Shalaby, M.A.W. (2021) 'A 3d-convolutional neural network framework with ensemble learning techniques for multi-modal emotion recognition', *Egyptian Informatics Journal*, Vol. 22, No. 2, pp.167–176.

Sharmila, A. (2021) 'Hybrid control approaches for hands-free high level human–computer interface-a review [PMID: 33191811]', *Journal of Medical Engineering & Technology*, Vol. 45, No. 1, pp.6–13, https: //doi.org/10.1080/03091902.2020.1838642.

Singh, H.P. and Kumar, P. (2021) 'Developments in the human machine interface technologies and their applications: a review [PMID: 34184601]', *Journal of Medical Engineering & Technology*, Vol. 45, No. 7, pp.552–573, https://doi.org/10.1080/03091902.2021.1936237.

Tao, W., Li, C., Song, R., Cheng, J., Liu, Y., Wan, F. and Chen, X. (2020) 'EEG-based emotion recognition via channel-wise attention and self attention', *IEEE Transactions on Affective Computing*, Early Access, p.1.

Valstar, M.F. and Pantic, M. (2002) *MMI Facial Expression Database* [online] https://mmifacedb.eu (accessed 19 May 2022).

Wang, P., Jiang, A., Liu, X., Shang, J. and Zhang, L. (2018) 'LSTM-based EEG classification in motor imagery tasks', *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, Vol. 26, No. 11, pp.2086–2095.

Wei, W., Jia, Q., Feng, Y., Chen, G. and Chu, M. (2020) 'Multi-modal facial expression feature based on deep-neural networks', *Journal on Multimodal User Interfaces*, Vol. 14, No. 1, pp.17–23.

Yang, Y., Wu, Q., Fu, Y. and Chen, X. (2018) 'Continuous convolutional neural network with 3d input for EEG-based emotion recognition', *International Conference on Neural Information Processing*, pp.433–443.

Zhang, D., Yao, L., Zhang, X., Wang, S., Chen, W. and Boots, R. (2017) 'EEG-based intention recognition from spatio-temporal representations via cascade and parallel convolutional recurrent neural networks', pp.1–8, arXiv preprint arXiv:1708.06578.

Zhang, Y., Hossain, M.Z. and Rahman, S. (2021) 'DeepVANet: a deep end-to-end network for multi-modal emotion recognition', *Human-Computer Interaction – INTERACT 2021, INTERACT 2021. Lecture Notes in Computer Science*, Vol. 12934, Springer, Cham, https://doi.org/10.1007/978-3-030-85613-7_16.

Zhao, S., Ding, G., Gao, Y., Zhao, X., Tang, Y., Han, J., Yao, H. and Huang, Q. (2020) 'Discrete probabil ity distribution prediction of image emotions with shared sparse learning', *IEEE Transactions on Affective Computing*, Vol. 11, No. 4, pp.574–587, https://doi.org/10.1109 /TAFFC.2018.2818685.

Zhao, S., Yao, H., Gao, Y., Ji, R., Xie, W., Jiang, X. and Chua, T-S. (2016) 'Predicting personalized emotion perceptions of social images', *Proceedings of the 24th ACM International Conference on Multimedia*, pp.1385–1394, https://doi.org/10.1145 /2964284.2964289.

Websites

https://mahnob-db.eu/hci-tagging/ (accessed 22 April 2022).

https://openbci.com/community/publicly-available-eeg-datasets/ (accessed 13 May 2022).