Neural Computing and Applications
https://doi.org/10.1007/s00521-023-09086-8

S.I.: KNOWLEDGE SERVICES: CHALLENGES AND FUTURE PROSPECTS

A deep perceptual framework for affective video tagging through multiband EEG signals modeling

Shanu Sharma[1,2] · Ashwani Kumar Dubey[1] · Priya Ranjan[3] · Alvaro Rocha[4]

Received: 16 August 2022 / Accepted: 20 September 2023

Abstract

Nowadays, multimedia content, such as photographs and movies, is ingrained in every aspect of human lives and has become a vital component of their entertainment. Multimedia content, such as videos or movie clips, is typically created with the intent to evoke certain feelings or emotions in viewers. Thus, by examining the viewer's cognitive state while watching such content, its affectiveness can be evaluated. Considering the emotional aspect of videos, in this paper, a deep learning-based paradigm for affective tagging of video clips is proposed, in which participants' irrational EEG responses are used to examine how people perceive videos. The information behind different brain regions, frequency waves, and connections among them play an important role in understanding a human's cognitive state. Thus, here a contribution is made toward the effective modeling of EEG signals through two different representations, i.e., spatial feature matrix and combined power spectral density maps. The proposed feature representations highlight the spatial features of EEG signals and are therefore used to train a convolution neural network model for implicit tagging of two categories of videos in the Arousal domain, i.e., "Low Arousal" and "High Arousal." The arousal emotional space represents the excitement level of the viewer; thus, this domain is selected to analyze the viewer's engagement while watching video clips. The proposed model is developed using the EEG data taken from publicly available datasets "AMIGOS" and "DREAMER." The model is tested using two different approaches, i.e., single-subject classification and multi-subject classification, and an average accuracy of 90%-95% and 90%-93% is achieved, respectively. The simulations presented in this paper show the pioneering applicability of the proposed framework for the development of brain–computer interface (BCI) devices for affective tagging of videos.

Keywords Affect · Video · Implicit tagging · EEG · Arousal · Excitement · CNN

✉ Ashwani Kumar Dubey
akdubey@amity.edu; dubey1ak@gmail.com

Shanu Sharma
shanu.sharma3@student.amity.edu;
shanu.sharma@abes.ac.in

Priya Ranjan
priya.ranjan@ddn.upes.ac.in

Alvaro Rocha
amr@iseg.ulisboa.pt

[1] Amity School of Engineering and Technology, Amity University Uttar Pradesh, Noida 201313, UP, India

[2] Department of Computer Science and Engineering, ABES Engineering College, Ghaziabad 201009, India

[3] School of Computer Science (SoCS), Internet of Things - Center of Excellence, University of Petroleum and Energy Studies (UPES), Dehradun 248007, India

[4] ISEG, University of Lisbon, Rua Do Quelhas, N86, 1200-781 Lisbon, Portugal

Published online: 17 October 2023

1 Introduction

Recently, due to the wide range of availability of smart handheld and display devices, a massive creation, usage, and circulation of digital images and videos have been observed [1]. With this increasing size of image and video repositories, the need for their automatic evaluation, tagging, and indexing is also in high demand, especially for highly tailored requirements and criteria for their efficient management [2, 3]. Since most of these images and videos are intended for humans, the inclusion of subjective aspects in their analysis methods plays an important role [4]. The multimedia data, i.e., images and videos, are usually developed with certain planned emotions which the makers want to induce in its viewers [5, 6]. As emotions represent affective experiences of one's feelings, they can play a significant role in analyzing affectiveness and interestingness of the multimedia content. Thus, emotion is always considered one of the important factors in analyzing videos. The emotional aspect of video content provides a subjective as well as high-level analysis of it which is considered as a suitable criterion for their indexing and categorization [7, 8].

To date, many research attempts have been conducted to analyze the videos by extracting various content-related data [9, 10]. Numerous theoretical and computational models were also put forth for their efficient analysis. These systems currently in use are mostly based on low-level elements like visual and audio characteristics [4, 10, 11], whereas the role of subjective and perceptual characteristics for their assessment is little explored. In short, an affective analysis and tagging of videos promise a new direction toward the most popular problem areas in the multimedia community, such as automatic video summarization, highlight extraction, indexing, etc. [10, 11].

Furthermore, for the past few years, the multimedia community has been working on simulating the cognitive capacities of humans into machines to improve their efficiency [12]. Today, with the advancement in neuroscience technologies, the field of cognitive psychology has become an interesting way for modern researchers to comprehend human behavior through the study of interdependent mental processes [13, 14]. Neural signal processing and brain–computer interfaces can be used to decipher various brain functions [12, 15]. They can assist complex information dispensation of the human brain for the detection of a range of cognitive states which can further help in analyzing the multimedia content. Neuroimaging studies can be very useful in predicting the unconscious responses of users to different kinds of multimedia information, such as movies, cricket videos, online video commercials, etc. [16, 17]. To date, "Functional Magnetic Resonance Imaging (fMRI)", "Positron emission tomography (PET)", "Computed Tomography (CT)", "Electroencephalograph (EEG)", and other technologies have been successfully applied to record and analyze brain activity [14]. Researchers have also been attempting to investigate the application of portable and affordable EEG devices in a variety of potential domains [13, 15, 16]. The evaluation of multimedia information using EEG waves is another area of active research [17–19]. Here, the affectiveness of video content is modeled by analyzing the affective state of viewers. In the literature, most of the research is done for classifying human emotion using EEG signals [20–23], whereas the field of analyzing the affectiveness of multimedia content using neurophysiological signals is largely unexplored.

In this paper, a framework for facilitating the implicit affective classification of video clips is presented by modeling the human's cognitive state while watching such content. This modeling is inspired by the utilization of EEG signals as well as the role of affect in evaluating video content. Furthermore, it is a well-established fact that the information behind different brain regions, frequency waves, and connection among them plays an important role in the effective analysis of a human's cognitive state. Motivated by this, two effective modeling of EEG signals are proposed here, i.e., spatial feature matrix (SFM) and combined power spectral density maps (PSDM) representation of EEG signals. These feature modeling representations highlight the spatial features of EEG signals and thus they are used to explore the power of deep learning in the current field and a convolution neural network (CNN)-based framework for affective tagging of videos is proposed. The following elements form the foundation of the work presented in this paper:

- As most multimedia content like images and videos, etc., are made to induce strong neural responses in users, it is believed that the incorporation of EEG signals will be beneficial for mapping human perception in existing video analysis frameworks.
- It is a well-established fact in neuroscience that different brain regions and frequency ranges represent different cognitive states of humans. Furthermore, these brain locations do not work in isolation, their connectivity plays an important role in analyzing the human cognitive state. Thus, the spatial characteristics of EEG signals and the connection between brain regions and frequency ranges are considered, and effective modeling of EEG signals is proposed here.
- Different brain regions and frequency ranges represent different cognitive states of humans. Furthermore, these brain locations do not work in isolation, their connectivity plays an important. Thus, the spatial

characteristics of EEG signals and the connection between brain regions and frequency ranges are considered to model the affective state of humans.
- To date, many machine learning (ML) and deep learning (DL) methods have been explored for automatic EEG signals classification. Thus, the power of deep learning is explored along with the higher and lower-level representations of EEG signals to accurately simulate the automated emotive tagging of videos.

To develop a model for affective video tagging using their EEG responses, in this paper videos under two extreme categories are considered, i.e., videos having highly exciting content and boring content. The "Arousal" dimension is taken into consideration for selecting the EEG data from publicly available datasets. The primary contributions the work presented in this paper are:

- First, to analyze the EEG signals corresponding to different video content, PSD-based features are extracted and explored at various brain regions and frequency ranges.
- Second for effective modeling of spatial characteristics of EEG signals and connections among different frequency ranges, two types of representations are used to present the input data, one is a multichannel and multi-frequency feature matrix representation of PSD values at all frequency ranges and channel positions, and another one is an image-based representation of these values through PSD maps.
- Third, to test the applicability of two different ways of EEG signal modeling, a CNN-based model is trained on both SFM and PSDM. The training and testing of the model are done under two categories, i.e., single subject classification and multi-subject classification to analyze the performance of the model on subject-specific and generalized data.

The proposed work toward the affective video tagging framework is further structured in remaining of this paper as; the background and related work section provides detailed information about the topic of study to readers. Here a discussion on the role of affect on video content analysis and the use of EEG signals in diverse fields including multimedia content analysis is discussed. Furthermore, the importance of different brain areas and frequency ranges for analyzing the cognitive state of humans is presented followed by limitations in the current field of work. In the materials and methodology section, the description of the dataset used for the experiment along with the different steps in developing the proposed framework is explained in detail. Various intermediate results obtained during the development of the model and the performance of the proposed framework are analyzed in the results section. At last, the proposed work along with its limitations and future directions are discussed in the conclusion section.

2 Background and related work

"Life is listless and colorless without emotion [5]." Multimedia content, such as movie trailers, music videos, advertisements, etc., are also made to make the viewer feel a certain way [24]. The effective analysis and tagging of videos usually depend on their effective characteristics, such as how interesting that video is or what is the capability of content to get the viewer engaged [6]. For instance, while designing a movie, to keep viewers interested action sequences that evoke surprise are frequently cut between scenes, while horror films that evoke fear are frequently shot in low light to make the setting seem bleak [24]. These cues have an implicit impact on how the audience reacts to the content and can offer a strong link between the content and subjective evaluations of human observers [25, 26]. To date, manual user ratings have been a major driving force in the design of multimedia material to elicit specific emotions, but this requires a sluggish and attention-demanding procedure by a human observer.

"Affective computing" is a popular direction of study that deals with the analysis of affective state of humans [5, 27]. To date, most of the research conducted in this field has been devoted to identifying and categorizing human emotions by analyzing facial expressions or a variety of physiological characteristics such as "EEG", "Electromyogram (EMG)", "Electrocardiogram (ECG)", "Galvanic Skin Response (GSR)", etc. [16, 20–23]. Since, emotion is an important aspect of everyone's life, assigning emotional tags to digital videos has been a popular topic of study in recent years [4, 6]. The emotional tagging of videos can be broadly classified into two categories: direct(explicit) and implicit. The direct method is a way of tagging the videos with emotions depending on their content, whereas implicit techniques derive the emotional tags of videos from a user's irrational reaction while they are watching them [10, 24, 25]. Developing an automated implicit video tagging system considering the emotion of the viewer can result in fascinating applications that can improve already-existing applications such as classifying movie genres or analyzing advertisement's impact, etc. [23, 24].

In recent years, the utilization of biosensing signals such as "EEG", "ECG", "EMG", "GSR" etc., has sparked interest in the field of affective computing [13, 14]. Furthermore, non-invasive BCI devices are in high demand as a result of the widespread availability and decreasing cost

of EEG systems. As brain signal conveys detailed information about a subject's cognitive state, there have been numerous attempts in the literature to use EEG signals for a diversity of applications, such as emotion recognition [16, 20], behavioral modeling [28], detection of neurological diseases, etc. [29–31].

Electroencephalography is a method wherein electrical signal strength from the brain is measured using a device "electroencephalogram" [14]. It is a non-invasive method, in which several electrodes are placed at different scalp areas to capture signals from different parts of the brain. The human brain is usually classified into four broad sections: the "frontal lobe", "parietal lobe", "temporal lobe", and "occipital lobe" as shown in Fig. 1. Each part of the brain handles the task of processing information differently; for example, when performing a visual activity, the frontal lobe typically deals with decision-related activities, whereas the parietal section deals with action-related activities [12, 14]. The temporal lobe is often active for tasks involving object recognition, whereas the occipital lobe is engaged for attention-oriented tasks [12]. Thus, proper modeling of captured signals at different brain regions is very important to understand the cognitive state concerning the tasks being performed.

The EEG signals contain rich neural activity information about the human brain especially in the frequency range of 2–64 Hz [13]. Furthermore, when the brain is in a certain state, its electrical pattern generates variable frequency patterns in different frequency ranges depending on the cognitive state of the brain. High-frequency ranges such as "Gamma (> 32 Hz)", "Beta (16–32 Hz)" and low-frequency ranges such as "Alpha (8–16 Hz)", "Theta (4–8 Hz)", "Delta (< 4 Hz)" are some of the popular frequency ranges that are effectively employed in the literature for cognition study [12, 13]. These frequency ranges have different roles in assessing a person's cognitive state. It is clear, for instance, that low-frequency ranges are typically linked to unconscious cognition, that's why the alpha waves are very active during a relaxed state of the brain in the occipital and parietal brain areas [20]. Additionally, high-frequency waves in the frontal brain section and other sections of the brain are frequently associated with mental acuity [32].

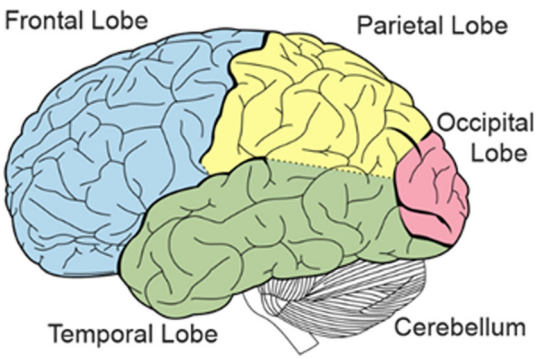

Fig. 1 Human brain structure [12]

In the literature, very few studies have been found for multimedia content using EEG signals. Among earlier studies in the related domain, one was published in [33], where authors performed a "Rapid Serial Visual Presentation (RSVP)" experiment to examine human attention through their EEG responses. Here a series of images along with the target image has been presented to viewers to analyze their attention process and to track their viewing pattern. Various similar studies have been found in [34–36]. EEG signals were also investigated to address a range of computer vision tasks, such as object categorization [37], object segmentation [38], object identification [39], and searching of images [40], etc. A portion of the research also investigated combining different modalities such as EEG, eye-tracking, and user ratings to examine artifacts in images and videos [41]. One of the studies utilizing EEG for visual data categorization is presented by authors in [42]. Here, the authors presented a strategy for merging EEG-based extracted features to increase the object categorization accuracy for six different categories. Some similar works were reported in [25, 26, 43], where authors utilized the EEG signals for analyzing the matched and unmatched tagging of multimedia content.

The proposed work presented in this paper forms a base on the survey and analysis done by authors in [5–8, 27], which presented a thorough analysis of related fields such as affective computing, affective multimedia content analysis, and the need for affective video analysis systems. Some similar attempts to the proposed work can be found in [44–46], where authors presented a hybrid emotional tagging framework using the combination of EEG and various video content features, such as audio and visual [44]. Furthermore, an advanced image-based representation of EEG signals to train the CNN network for emotion classification is provided in [45]. A combination of different deep learning architectures for the emotional tagging of videos is explored in [46]. In [47], the authors explored the use of the graph convolution neural network (GCNN) for identifying the videos using EEG signals. Here authors tried to model the EEG signals as signals on a graph to further train the GCCN.

For understanding human elicitation, the past decade has primarily witnessed two types of research. The emotion of an audience is mapped using facial expressions or other physiological data. Diverse efforts are also made to identify ideal features for efficiently mapping the viewer's emotions [12, 14, 22]. These approaches do not consider visual stimuli. And in other research, emotive video clips are utilized to gauge the viewer's emotions. But most of the

research utilizes subjective input for classifying emotions [20, 21, 23]. Furthermore, it has been noted in the previously published works that most techniques for classifying EEG signals entail the extraction of different kinds of features, including "time-domain", "frequency-domain", and "time–frequency domain features" [13–16]. These feature combinations have also been tested to improve the accuracy of several classification models [21–23, 32]. According to the present literature, the following major limitations have been found, which are considered while developing the proposed model.

- Most of the previous research on affective computing has been devoted to the classification of emotions evoked by emotional visual stimuli. The researchers did not consider the fact which type of content is affecting human emotion mostly.
- EEG signals were explored blindly with different combinations of classification methods for the automatic classification of human affective states.
- The major insight that which brain region and frequency bands contain the most effective information is ignored while designing a model.
- Furthermore, the relationship between multimedia content and physiological responses is not addressed thoroughly with the aim of affective content analysis.

To address the above-mentioned gaps, in this paper, two different representations of EEG signals are used to explore the relationship between various important frequency ranges and brain regions. A deep learning-based model is proposed for affective implicit tagging of videos through EEG signals. For the better representation of EEG signals two different methods are presented here to train the CNN-based model for automatic tagging of videos. Instead of using EEG signals in isolation, these two feature modeling representations are proposed here to explore the power of spatial characteristics of EEG. Furthermore, these two representations are used to train the deep learning model to facilitate the automatic tagging of videos. A detailed explanation of the proposed methodology and experimental results is presented in further sections.

3 Materials and methodology

Video clips are usually created with the intent to produce certain feelings in viewers. Thus, considering the emotional aspects of videos here an affective video classification framework is proposed, which utilizes the irrational EEG responses of the viewers corresponding to video clips for assigning affective tags to them. In this section, a step-by-step explanation of the development of the proposed affective video tagging framework is described. The workflow of the proposed approach is presented in Fig. 2.

Here, EEG signals corresponding to two different types of videos are used to develop the model. The EEG signals captured at different brain locations are first analyzed at well-known frequency ranges by extracting power spectral density-based features. As discussed in Sect. 2, specific brain areas, frequency ranges, and connections among them serve distinct functions in the study of the human cognitive state, thus the extracted features are then encoded to model the human cognitive state to measure the power of engagement of video content. The feature encoding is done to generate a better representation of EEG features, which can maintain the spatial characteristics at different brain locations and connections between frequency ranges which is the most important thing in EEG signal modeling. These two types of encoded features are then used to train the CNN model to facilitate affective tagging of videos.

3.1 Dataset description

Suitable modeling for emotions must be done to examine a given visual piece at an affective level. According to research in cognition and neuroscience, human emotions can commonly be categorized as continuous and discrete [5, 6]. In contrast to categorical categories like surprise, happiness, sadness, disgust, fear, or rage, continuous emotions are those that are defined using dimensions like positive–negative or calm-aroused [7, 8]. To generate different types of emotions among viewers different types of emotional clips are usually used. Various publicly available datasets are also present in the literature, where different neuro-physiological signals are recorded while presenting the emotional clips or images to the viewers. These emotional videos are developed on the dimension of "Valence-Arousal" (V-A) emotional dimensional space [43]. This dimensional model is used to provide a quantitative analysis of the emotion. Valence is usually measured on the scale of positive and negative, in the range of pleasantness or unpleasantness respectively, whereas arousal represents the strength of evoked or induced emotion in the range of low to high as presented in Fig. 3.

Various publicly available datasets such as MANHOB-HCI [43], AMIGOS [48], DREAMER [49], DEAP [50], and DECAF [51], etc., exist, which contain the EEG recordings of the participants corresponding to the affective video clips. These datasets differ in many aspects, such as modalities to collect physiological recordings, number of participants, type of video, self-assessment, etc. Due to variations in the quantity and kind of retrieved features, many classical techniques cannot be generalized across datasets. In addition to the forms of audio-visual stimulation (music videos vs. movie clips), the datasets differ in

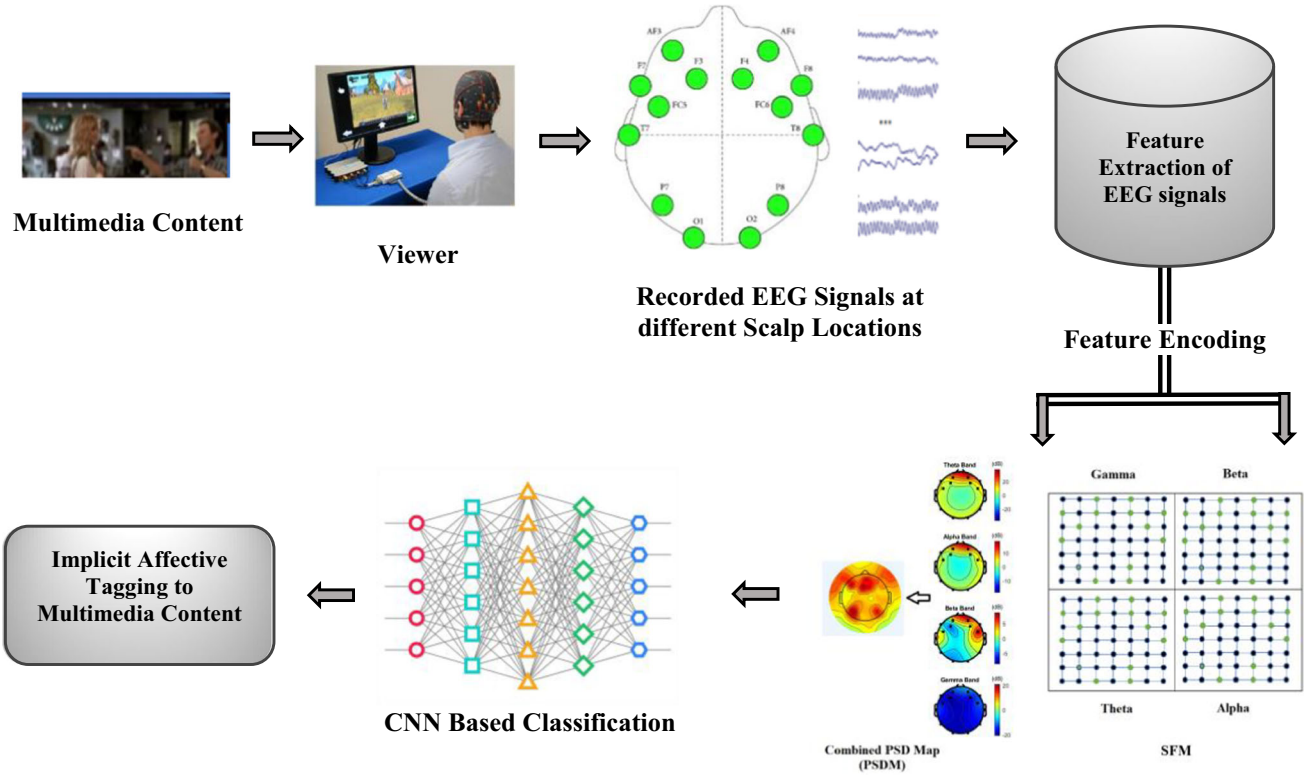

Fig. 2 The workflow of the proposed framework

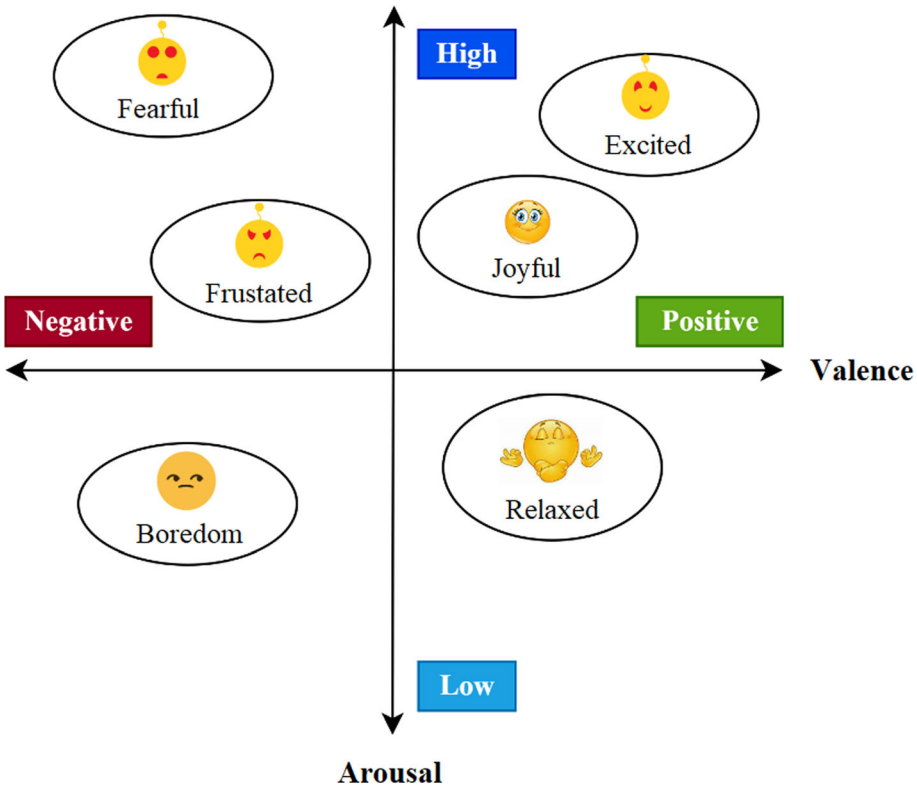

Fig. 3 Valence-Arousal (V-A) space for affective video selection

the length of the trial and the availability of baseline data also.

In this paper, AMIGOS [40] and DREAMER [41] datasets are used to develop and test the model. These two datasets are chosen, as they contain a uniform number of EEG recordings, i.e., at 14 channel locations using the same signal-capturing device as well as the same sampling rate of 128 Hz, which facilitates uniform modeling of the proposed approach. The detailed information on these two datasets is presented in Table 1. Furthermore, various steps involved in the development of the video tagging model in further subsections are explained with the help of EEG data from the AMIGOS dataset.

3.1.1 AMIGOS dataset [48]

It is a dataset for exploring the emotion, behavior, and mood of the viewers while watching short and long video clips. In this dataset physiological recordings (EEG, ECG, and GSR) of forty (thirteen female) participants are present concerning different affective video clips for two different experiments, i.e., short experiment and long experiment. The clips used in the experiment were chosen depending on where they fell on the V-A dimensional space [Fig. 3]. In the short experiment, participants watched 16 video clips (four from each quadrant of V-A space) of a duration of approximately 250 s, whereas, in the long experiment, they watched four videos of about 20 min duration. The affective reactions of the participants were observed in two separate situations: one, when the participant watched videos alone, and the other one is when they watched videos in a group. The dataset also includes participants' high-definition facial and full-body depth recordings.

3.1.2 Video stimuli

In this paper, for developing the proposed model, the EEG dataset of forty participants corresponding to affective video stimuli from the arousal quadrant, i.e., videos having content capable of generating high arousal (HA) and low arousal (LA) is selected. The arousal emotional space represents the excitement level of the viewer. Thus, it is used to analyze the videos based on the fact that how much the viewer is attentive while watching it. The details of these videos corresponding to the two quadrants are presented in Table 2.

3.1.3 Trial Structure

In the short experiment, the recording of EEG data of participants for 16 videos in the arousal category is done in 16 different trials. The structure of the trial is presented in Fig. 4. Here self-assessment of participants is taken before and after each trial to track the participant's emotional state, for which participants were asked to give responses on the dimensions of valence, arousal, dominance, liking, and familiarity. Then, participants were asked to list at least one basic emotion, such as Neutral, Happiness, Sadness, Surprise, Fear, Anger, or Disgust. After self-assessment, a fixation cross is presented for five seconds followed by a presentation of the video clip. The required sensory data such as EEG, ECG, and GSR were collected during both baseline and visual stimuli periods.

3.2 EEG Data analysis and feature extraction

3.2.1 Pre-Processing of EEG Data

The pre-processing of EEG signals is a crucial step in their analysis since due to modest voltage variations they frequently have very low signal-to-noise ratio. They are also impacted by other artifacts, such as signal line noise, noise from muscle actions, eye blinks, etc. In the AMIGOS [48] and DREAMER [49] datasets, EEG data of participants at various brain regions is recorded using 14 channel electrodes as shown in Fig. 5. Here to remove noise from the EEG data, first, ocular artifacts from EEG data at all channel locations are removed using the independent component analysis (ICA)-based blind source separation technique [18]. In addition, the re-referencing of EEG data is performed followed by band-pass filtering between 4 and 65 Hz [18, 19].

Table 1 Dataset Description Used in the Experiment

Dataset	No. of participants	Stimuli and duration	EEG Data recording specification
AMIGOS (2018) [48]	40 (13 female)	16 movie clips (51–150 s)	Sampling rate: 128 Hz No. of Channels: 14
DREAMER (2018) [49]	23 (16 female)	18 movie clips (65 to 393 s)	Channel positions: AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4, M1 and M2" Recording device: "Emotiv EPOC Neuroheadset"

Table 2 Description of Video Stimuli Used in the Experiment

Dataset	Video category	Video description [48]	Total No. of movie clips
AMIGOS (2018) [48]	HA (High Arousal)	"Airplane, When Harry Met Sally, Hot Shots, Love, Silent Hill, Prestige, Pink Flamingos, Black Swan"	8
	LA (Low Arousal)	"Exorcist, My Girl, My Body Guard, The Thin Red Line Fatigue level detection, August Rush, Love Actually, House of Flying Daggers, Mr Beans' Holiday"	8

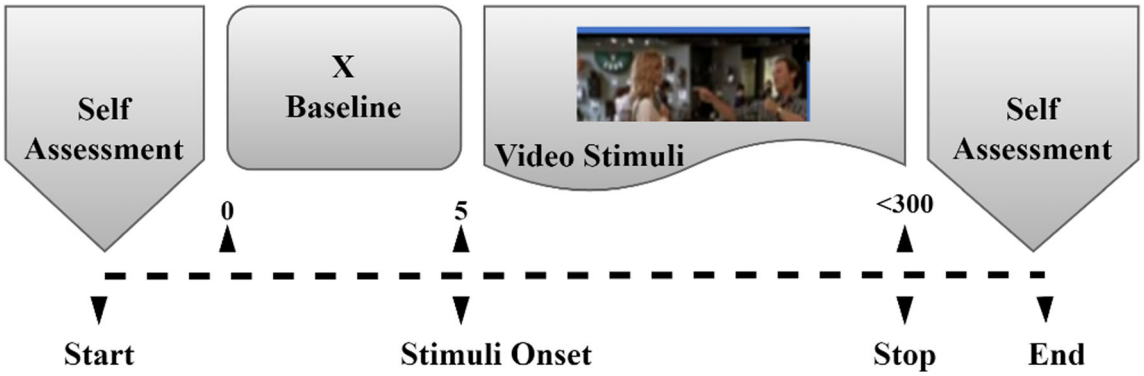

Fig. 4 Stimuli trial structure for EEG recording (time in seconds)

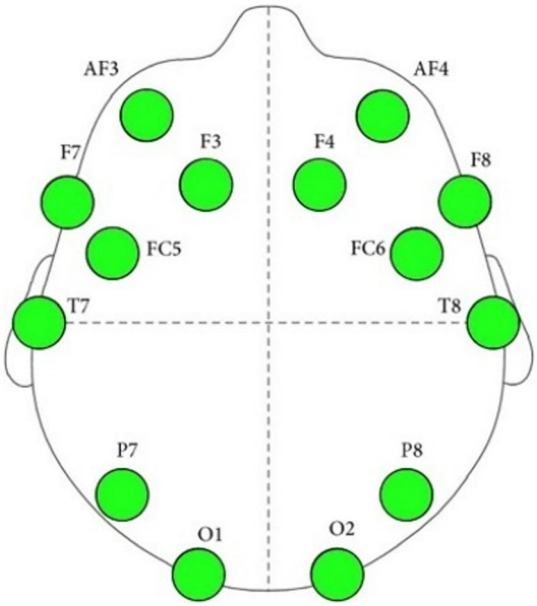

Fig. 5 Channel positions for EEG data recording in AMGOS [48] and DREAMER [49] dataset

3.2.2 Frequency ranges extraction

The importance of analyzing EEG signals at different brain regions and frequency ranges is already discussed in Sect. 2. Thus, here EEG signals are decomposed to extract various frequency ranges required to examine the human cognitive state. EEG signals are typically non-stationary in nature, therefore converting signals from the time domain to the frequency domain is always required to determine a meaningful spectral component from them. A popular technique for exploring non-stationary data in the frequency domain is the discrete wavelet transform (DWT). It gives a multiresolution description for a non-stationary signal [52]. The multiresolution analysis of any signal involves the process of analyzing the signal at different frequencies with different resolutions.

By breaking a signal down into several scales, DWT can analyze it at multiple resolutions. This can be achieved by convolving the signal with various translations and scalings of a tiny oscillatory function, known as the mother wavelet. Here, translation denotes the location of the window, which, in the transformed domain, represents the passage of time. The global and localized information of a signal are both represented by the term scale. While analyzing the frequencies, the globalized information is usually contained in high scales, i.e., low frequencies whereas the narrow scales, i.e., high frequencies represent the localized and detailed information of the signal [52]. The DWT performs the signal decomposition by applying the filters of different cut-off frequencies [53], where high pass and

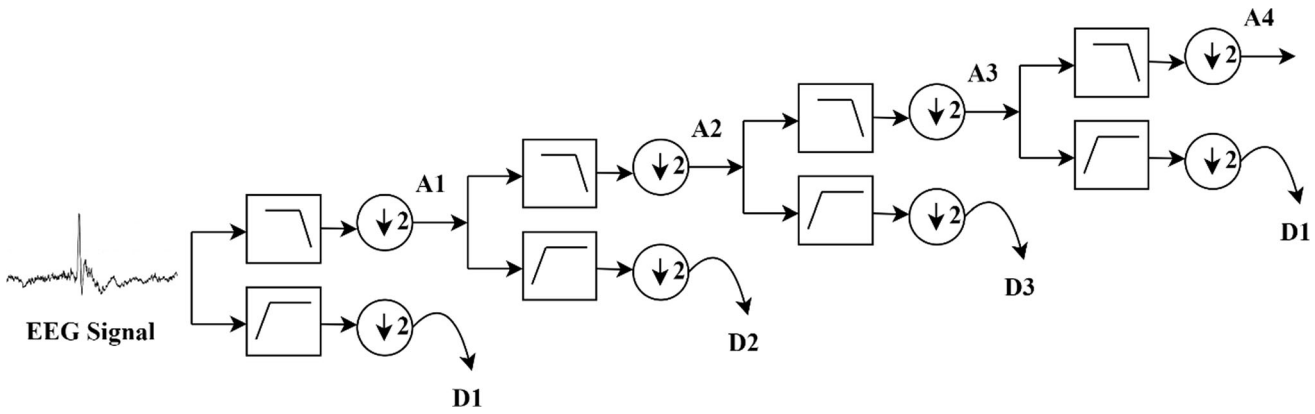

Fig. 6 EEG signal decomposition using DWT for affective video analysis

Table 3 Details of Extracted Frequency Ranges

Coefficients	Frequency range	Band name
D1	3264 Hz	Gamma
D2	1632 Hz	Beta
D3	816 Hz	Alpha
D4	48 Hz	Theta
A4	04 Hz	Delta

low pass filters are used in a recursive convolution to separate the signal into its high and low frequency components [54, 55]. Additionally, scaling of the signal is carried out through downsampling, to lowers the signal's sampling rate [52].

In the current study, the DWT-based decomposition of EEG signals is done using Daubechies-four (db4) mother wavelet. Db4 wavelet is chosen after testing many wavelets from the Daubechies wavelet family because of its good performance and lower computing complexity than other wavelets. Furthermore, to find the number of decomposition levels for DWT-based decomposition, the dominant frequency component of the signal is considered. As the EEG data used in the current study is recorded at the sampling rate of 128 Hz, only four decomposition levels could be employed to split the signal into the necessary frequency ranges, namely "Gamma", "Beta", "Alpha", "Theta", and "Delta".

Equations (1) and (2) represent the wavelet and scaling functions, respectively [55].

$$\emptyset_{j,k}(m) = 2^{-\frac{k}{2}} h\left(2^{-k}m - p\right) \tag{1}$$

$$\psi_{j,k}(m) = 2^{-\frac{k}{2}} g\left(2^{-k}m - p\right) \tag{2}$$

Here m ∈ 0,1,2....N-1, k ∈ 0,1,2... L-1, p ∈ 0,1,2... 2^{k} -1. The length of the signal is denoted by N, and no. of levels is represented by L which is 4 here.

By applying the high pass and low pass filtering recursively along with the downsampling ratio of two, the approximate (A_i) and detailed coefficients (D_i) are generated at each level of decomposition. Equations (3) and (4) are used to calculate A_i and Di at the i_{th} level, respectively [54].

$$A_i = \frac{1}{\sqrt{N}} \sum_m x(m).\emptyset_{k,p}(m) \tag{3}$$

$$D_i = \frac{1}{\sqrt{N}} \sum_m x(m).\psi_{k,p}(m) \tag{4}$$

The required frequency ranges are then extracted by further decomposition of approximation coefficients to extract the detailed information stored in the signal as depicted in Fig. 6.

The extracted frequency ranges as a result of DWT-based signal decomposition are shown in Table 3, which are further used to explore the human responses corresponding to emotional video stimuli.

EEG Feature Extraction: Power spectral density (PSD) is the most popular method in the literature for EEG signal modeling [56]. The "frequency content" of any signal, or how the signal power is distributed over frequency, can be seen in its power spectrum. In this study, PSD value is estimated for "Gamma", "Beta", "Alpha" and "Theta" frequency ranges. Furthermore, as shown in Fig. 4, during each trial EEG data is recorded for five second baseline period and for the video stimuli period, where each trial's length varies depending on how long the video is. Thus, to analyze the effect of any particular video stimuli on the human cognitive state the relative analysis of EEG signals in the visual stimuli period is done with respect to the baseline period. The EEG data corresponding to 4–4 video clips under the arousal quadrant, i.e., HA and LA [Table 2] is used for this study. PSD values are extracted for both baseline and video stimuli periods. Then these values are normalized using Eq. 5 to normalize power levels in the

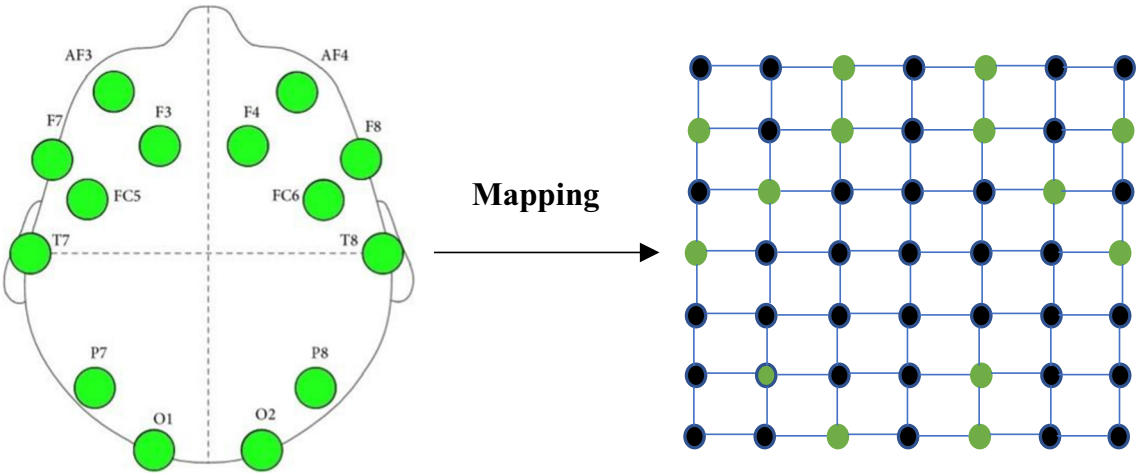

Fig. 7 Electrode positions and the corresponding feature matrix

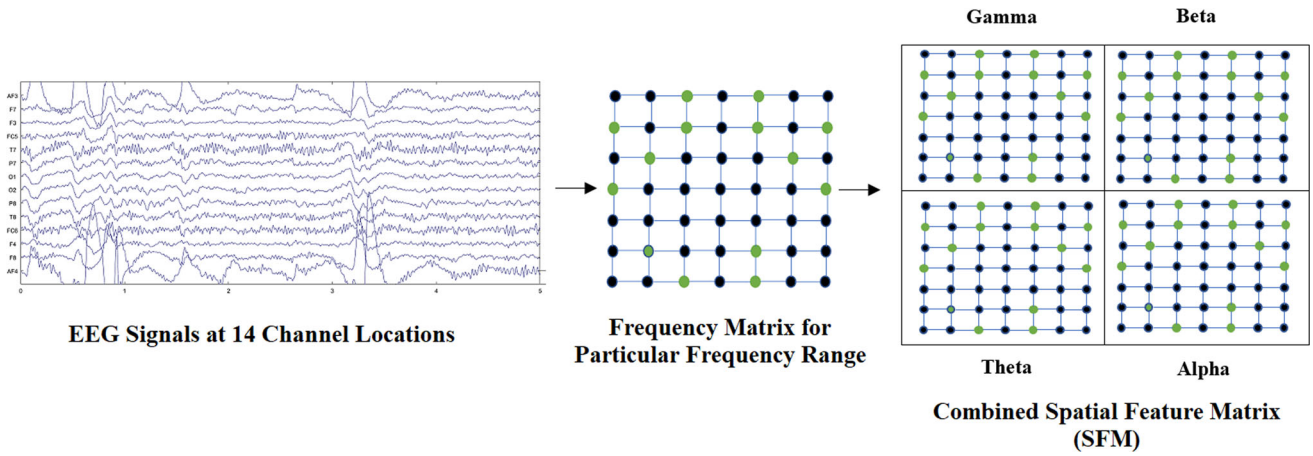

Fig. 8 Spatial feature matrix generation

stimulus period with respect to the baseline (fixation) period, which would eliminate any constant and unrelated activities from the EEG signals during the recording period. This normalized power is calculated for mentioned frequency ranges corresponding to all eight trials under mentioned categories of videos of forty participants at all fourteen scalp locations (Fig. 5).

$$P^{N}_{c,b} = \left(P^{\text{stimuli}} - P^{\text{fixation}}\right)/P^{\text{fixation}} \quad (5)$$

Here, P^{stimuli} and P^{fixation} represents the power values for channel c and frequency band b in stimuli and the fixation time respectively, where c ε [1 to 14], b ε [1 to 4], and P^{N} denotes normalized power. In this way, the EEG signals corresponding to each video trial of a particular participant is characterized by a feature matrix of 14*4 dimension.

3.3 EEG feature encoding for CNN-based classification

In the literature, very few work exists on using deep learning frameworks utilizing EEG signals. The reason behind this is, that capturing images and videos is much easier than capturing any kind of bio-sensing signals. Furthermore, different bio-sensing signals are usually recorded with different types of devices and have different profiles. Thus, to generalize the deep learning frameworks for these types of signal processing is not an easy task. In existing deep learning frameworks, CNN-based architectures are the most popular models for analyzing spatial information. These models have shown their significance in various computer vision-related tasks also. However, the use of these frameworks for EEG signal processing needs more attention during pre-processing and modeling of input data.

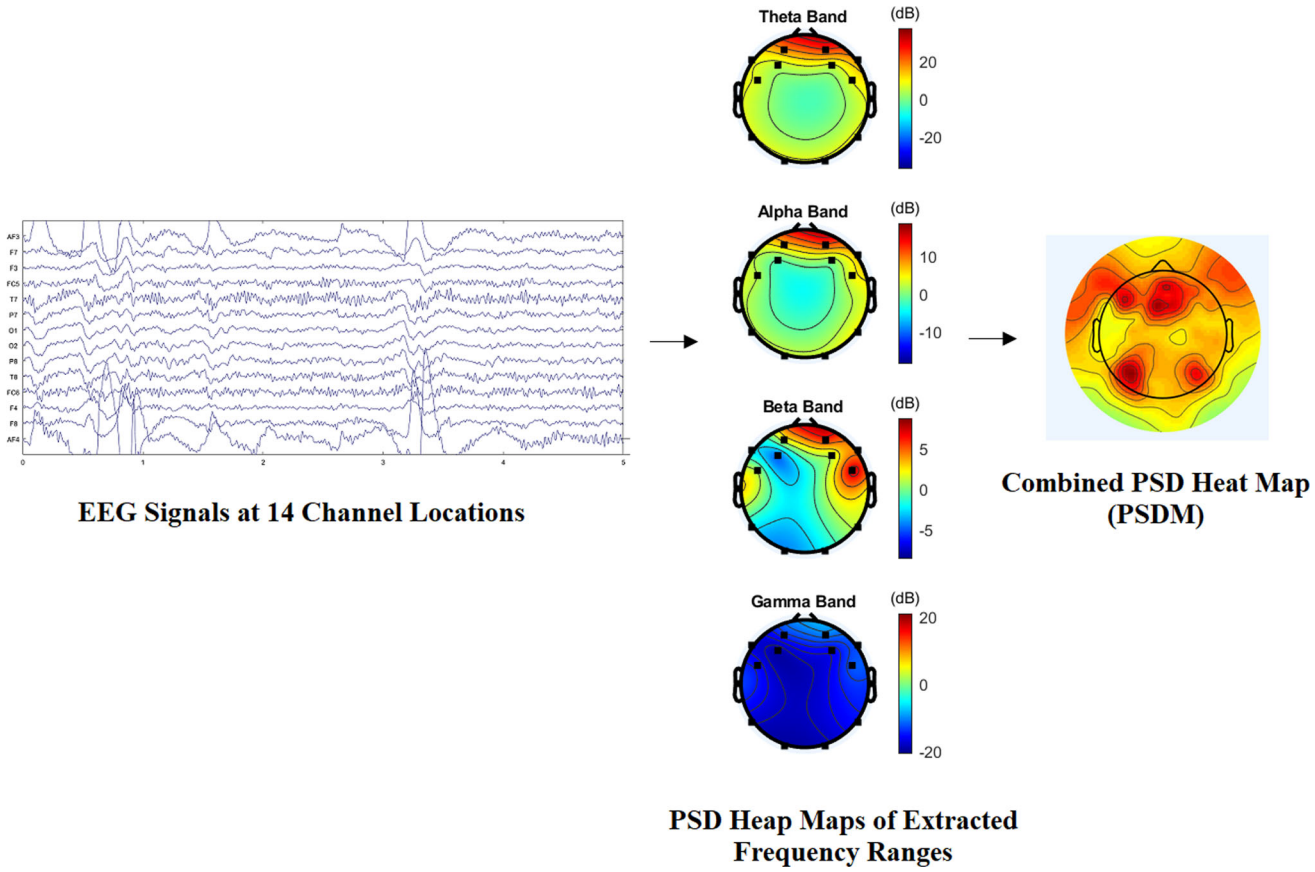

Fig. 9 Generation of combined power spectral density maps (PSDM)

Furthermore, most of the existing work using deep learning frameworks for analyzing EEG signals uses two-dimensional EEG spectrograms to train the networks [45, 46]. These spectrograms represent the information of a single brain location at a time. Thus, while training the model the spatial characteristics of brain regions are not captured by the models. Whereas as discussed in Sect. 2, the interaction among different brain regions as well as frequency ranges are very important in understanding the cognitive state of humans. As EEG signals contain significant information at different brain locations and frequency ranges, they should be affectively modeled while using deep learning frameworks. Thus here, two different EEG feature encoding approaches are proposed to generate input representations for CNN.

- Spatial Feature Matrix (SFM)
- Combined PSD Map (PSDM)

These feature encoding approaches are designed to generate the better representation of EEG signals, so that the responses of different brain regions and frequency ranges can be used in a combined manner to model the affective state of humans. Furthermore, these representations are used in the form of images which can further facilitate the utilization of CNN-based networks for further classification task. The two proposed approaches are discussed below in detail:

3.3.1 SFM-based feature encoding

In SFM-based feature encoding approach the extracted normalized PSD values are arranged to generate a matrix representation. The positions of the channel locations [Fig. 5] are used to assign the corresponding locations in the matrix. These positions can be converted into a corresponding matrix of size 7×7 as shown in Fig. 7. Here, green-colored points represent the corresponding channel locations on the scalp. These matrix points are then filled using extracted normalized PSD values of the signal at that particular channel location. The dark points, which do not belong to any channel locations are filled with a default value.

Similarly, the feature matrix for signal responses in four extracted frequency ranges [Table 3] i.e., Gamma, Beta, Alpha and Theta is prepared. These frequency-wise matrices are then again arranged and combined to generate SFM of size 14×14 as shown in Fig. 8. SFM facilitates the representation of frequency domain features of EEG

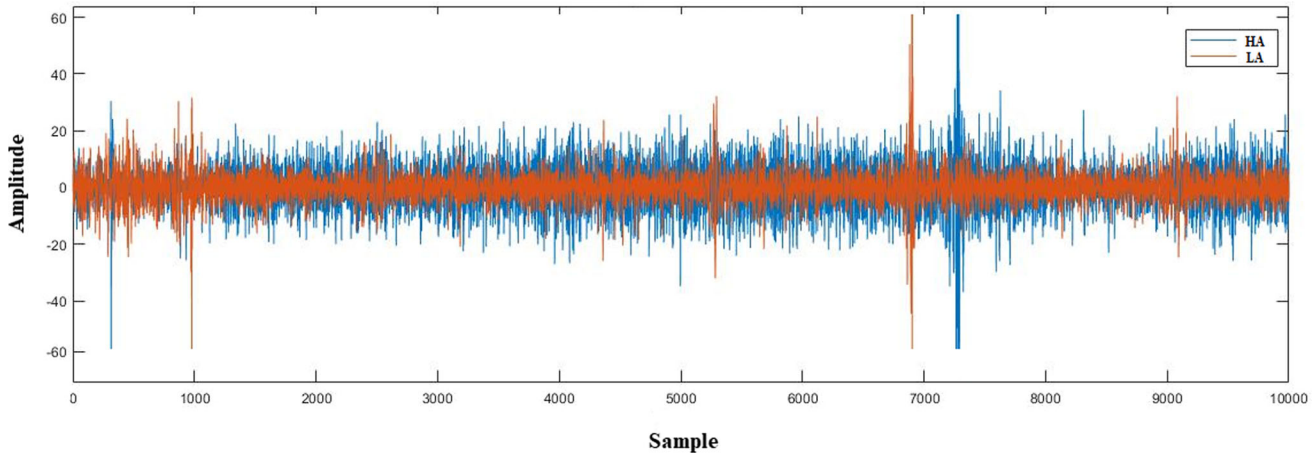

Fig. 10 The EEG signal of two categories of videos HA and LA in the time domain at channel AF3

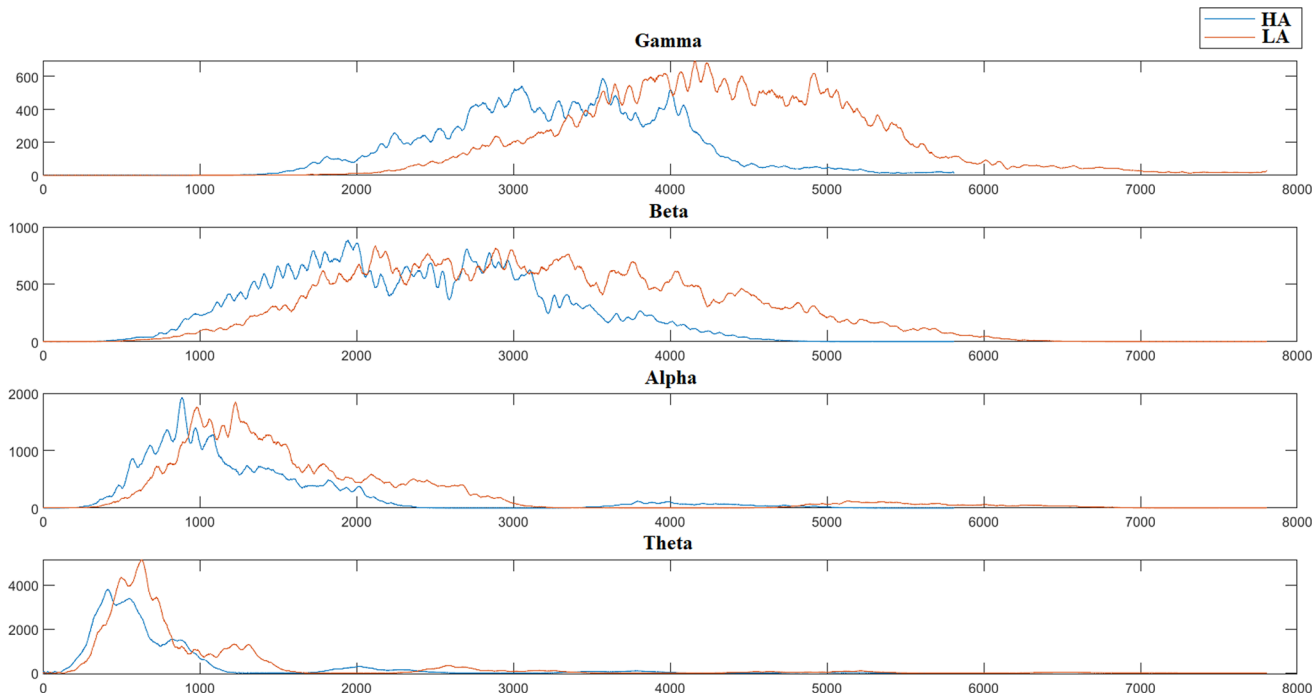

Fig. 11 EEG signal representation in different frequency ranges for two categories of videos HA and LA at channel

signals through PSD at different brain locations and channel locations at a single place. Thus, it preserves the spatial and frequency band characteristics of EEG signals which is somewhere missing in most of the literature. Furthermore, this kind of representation can be generalized to any kind of EEG data having varying no. of channel positions.

3.3.2 PSDM-based feature encoding

Power spectral density heat maps are one of the most popular methods among neuroscientists to visualize the function of various brain regions during a cognitive task. PSD features of EEG signals are used to generate corresponding heat maps. These heat maps are used to visualize the effective brain regions through their topographical representations. Here, the idea is to use the image-based representations of PSD values through heat maps to generate a better representation for CNN-based networks. In this study, heat maps are generated which contain the topographical information of 14 brain locations. Furthermore, to model the information at different frequency ranges, first calculated PSD values of EEG signals are used to generate heat maps corresponding to different extracted frequency ranges, i.e., Gamma, Beta, Alpha and Theta are generated as shown in Fig. 9.

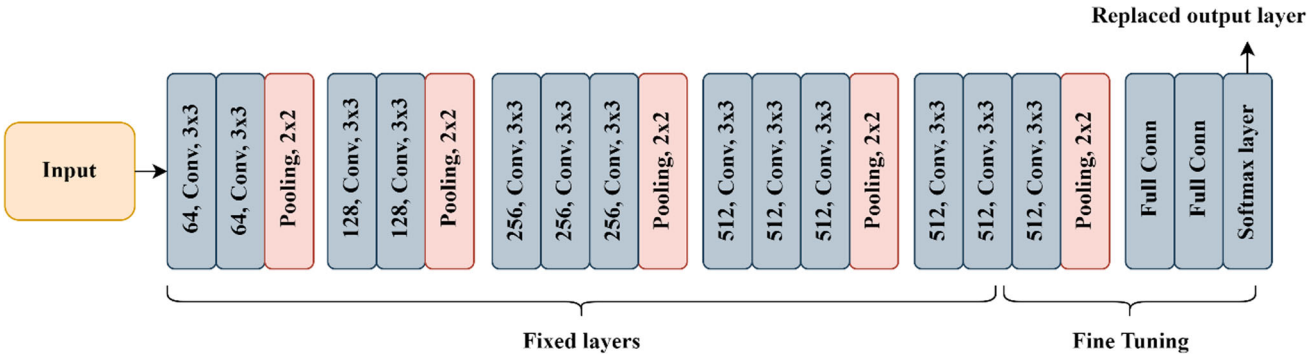

Fig. 12 CNN network training based on VGG-16

Table 4 Performance of Proposed Approach on Single Subject's Data

Model type	HA		LA	
	Precision	Recall	Precision	Recall
SFM-based CNN Model	**0.95**	**0.85**	**0.90**	**0.83**
PSDM-based CNN Model	**0.97**	**0.94**	**0.92**	**0.90**

The color in the heat map represents the active region of the brain for a particular cognitive task, e.g., here dark blue color is representing less activity and the orange color is representing high activity as presented in the bar in Fig. 9. Thus, to create an image-based combined representation of PSD maps, the three most significant maps are selected to combine than like an RGB image. After rigorous analysis, it has been found that Gamma frequency range can be dropped, as it's showing less activity at most of the channels. Thus, Alpha, Beta and Theta PSD maps are further used to generate combined PSDM. Then, the CNN model for the emotional labeling of videos is trained using these aggregated heat maps as described in the methodology section.

Table 5 Performance of Proposed Approach on Multiple Subjects Data

Model type	HA	LA		
	Precision	Recall	Precision	Recall
SFM-based CNN Model	**0.92**	**0.83**	**0.87**	**0.81**
PSDM-based CNN Model	**0.95**	**0.91**	**0.91**	**0.89**

Table 6 Comparative Analysis with Similar Works

Ref	Task	Data used	Methodology	Performance and analysis
[42]	Video tagging	Self-collected data 13 participants	Different classifiers were tested considering each channel location using several combinations of feature extraction techniques, such as DWT, FFT, STFT, etc	80% Accuracy Single Channel position AF8 is used No detailed analysis to support the choice of specific features, classifiers, or channel location
[47]	Video tagging	DEAP database	Graph Convolution Neural Network (GCNN)-based classification model is proposed using power and entropy-based features	48.25%—62.94% Accuracy
[19]	Video tagging	AMIGOS dataset	Various frequency ranges are extracted using DWT-based method. Different frequency domain features are extracted and modeled to model human attentiveness followed by SVM-based classification for video tagging	93.2% Accuracy Two categories of videos H_AH_V and L_AL_V
Proposed work	Vido tagging	AMIGOS and DREAMER dataset	CNN based classification model is proposed with two featured modeling methods considering the importance of brain regions and frequency ranges	Accuracy: DREAMER dataset- 90%-95% AMIGOS Data Set- 90%-93% Two categories of video HA and LA

4 Results

In this paper, an attempt has been made for developing a deep learning model utilizing EEG signals for providing an automatic affective tag to video clips. For this, the relationship between affective video content and their related EEG responses are encoded at different frequency ranges and brain areas, and two encoded EEG feature representations are developed, i.e., SFM and PSDM as described above. These two representations are used to train two different CNN models for affective tagging of videos. In this section, the two mentioned EEG feature encoding approaches are evaluated for CNN model development. The results are evaluated on AMIGOS [48] and DREAMER [49] datasets. Various intermediate results obtained during the different phases of the development of the models are further presented and discussed in detail with respect to video clips from the AMIGOS dataset [48]. As discussed in the methodology section, EEG data of 40 participants corresponding to 16 short movie clips from HA and LA quadrants are used to extract and modeling of features.

4.1 Data preparation

Here, the EEG signal of any participant for a particular video clip is represented as $E_{\text{channels x timepoints}}$, where channels represent the total number of electrode positions used to capture the brain signals and timepoints represent the total sampling points, which depend on the sampling rate and duration of the video clip. In AMIGOS dataset the signals are recorded at 14 channel locations using 128 Hz sampling rate. Here, all video clips slightly vary in duration, thus accordingly timepoints vary for each EEG recording also. In Fig. 10, EEG response of participant at channel position AF3 while seeing video clip of HA and LA video category is presented.

$E_{\text{channels x timepoints}}$, this time domain representation of EEG signals is highly dimensional. Various deep learning models have been tested which directly takes this time domain representation of the EEG signals for the development of the model [45, 46]. Whereas, as discussed previously the EEG signals are usually very complex in nature, and the information behind different brain regions and frequency ranges should be properly modeled for effective model generation. Thus, for effective modeling of EEG signals, first frequency domain processing is performed here to convert the time domain representation of EEG signals into frequency domain representation and different required frequency ranges are then extracted. A representation of EEG signals in different frequency ranges for the two mentioned categories of videos is presented in Fig. 11.

After frequency band extraction, normalized PSD values are extracted, which estimated the amount of power content in different frequency ranges. The power content in four frequency ranges is extracted, which resulted in 4 different features from the EEG signals, and now the sub-feature space is represented by $F_{\text{channels x features}}$, where channels are electrode positions, i.e., 14 and features are 4 corresponding to four frequency ranges. $F_{\text{channels x features}}$ represent a reduced frequency domain representation of EEG signals and have been used by various researchers for the development of different types of classification models [42, 47].

The power content in different frequency ranges and brain locations contains important information about the cognitive state of the human. However, their direct use, i.e., array representation of these features may not generate an effective model. The reason behind this is, the array representation of extracted features does not explain the spatial characteristics of the brain scalp, as well as the connection and communication between different brain regions during a cognitive task. Furthermore, the relative performance of different frequency ranges also plays an important role. Thus, two different ways of arranging these features are explored here, i.e., SFM and PSDM. A detailed discussion on how to encode EEG signals in these types of representations is already explained in Sect. 3.3. During SFM-based feature modeling a 14-channel EEG signal $E_{\text{channels x features}}$ is converted into a 7×7 sub-matrix using the method presented in SFM-based feature encoding section, which is further converted into a 14×14 SFM. Each cell of the SFM contains the average normalized PSD at a specific brain location and frequency range. For PSD image-based encoding, PSD maps of the three most significant frequency ranges are used to develop a combined PSDM as discussed in Sect. 3.3.2. These two different representations are then used to train and test the CNN-based classification model.

As AMIGOS dataset contains EEG recordings of participants for very few affective video clips, i.e., 16, which is a small dataset to train the model. Thus, data augmentation is done here using a windowing approach. The EEG signals are windowed every ten seconds and further advanced by one second. The dataset was expanded in this manner to provide more than 200 trials for a single participant. SFM and PSDM representations for the augmented trials are then used to train the CNN model.

4.2 CNN-based training models development

In the realm of computer vision, CNN is the most used deep learning framework. It is an extended version of a

neural network that has the capability of extracting features from a grid-like matrix having some pattern. It delivers more versatile and deep feature extraction, ranging from low-level to high-level feature extraction from large datasets [27]. The resultant SFM and PSDM representations of EEG signals also contain a pattern of EEG spectral responses across various brain regions and frequency ranges for a particular video stimulus. This pattern contains the information behind certain cognitive states of the viewer. Thus, to capture the salient features from these representations, the power of CNN is explored here.

However, building a CNN from scratch to solve a particular problem takes a lot of time and requires a lot of computer resources like memory, processing capacity, and expensive CPUs and GPUs. In addition, training the network requires access to a relevant sizable dataset. Transfer learning has evolved as an effective strategy to overcome these issues. In order to perform a different but similar task with a smaller dataset, a pre-trained CNN model can use the existing knowledge learned from a larger dataset. Also, fine-tuning any pre-trained model on the target dataset is comparatively simple and takes less time and resources.

Visual Geometry Group (VGG)-16, is a standard 16-layer CNN architecture pre-trained on the ImageNet dataset [57]. In the literature, the researchers have explored the power of a pre-trained VGG-16 network for the classification of other types of datasets also [27, 58]. Thus, in this paper, to classify video clips in high arousal and low arousal categories, VGG-16 pre-trained network is used. VGG-16 is trained on millions of images to classify 1000 different output labels, whereas here the task is to classify the input into two categories of arousal dimension. So, some of the starting layers, i.e., convolution and pooling layers of the VGG-16 are transferred as it is to utilize its power of analyzing the pattern in images, and the last layers, i.e., fully connected and output layer are changed according to the required classification as presented in Fig. 12.

In this setup, the fixed layers supply the standard features, while the fully connected layers and output layers are fine-tuned with the help of the target dataset. The pre-trained VGG-16 architecture accepts inputs with dimensions of 224 $\times$ 224x3. Thus, the SFM and PSDM encoded feature representations are scaled to a dimension of 224 $\times$ 224x3 to prepare the input layer. Each model was developed using the Pytorch framework and trained using a single 24 GB Nvidia Titan RTX GPU. The performance of the models is analyzed for different batch sizes.

4.3 Performance evaluation

The VGG-16-based CNN models using SFM and PSDM feature representations are trained and tested using two different classification approaches:

- Single subject classification
- Multi-subject classification

In a single-subject classification model, the model is trained and tested using the EEG data of one subject to create a subject-specific model. The fivefold cross-validation technique is adopted here. For the multi-subject classification model, data from all 40 subjects is used. LOSO (Leave one Subject Out) cross-validation method is adopted here, to test the generalizing ability of the model. With this assessment process, the model is trained on all the subjects except one, and the evaluation is done on the other one. Table 4 and 5 present the performance evaluation of the developed models using SFM and PSDM feature representations where precision and recall are calculated using Eq. 6 and 7, respectively.

Precision represents the measure of correctness of any machine learning model with respect to all predictions for positive class. It defines how much portion of the positive prediction is actually correct with respect to all positive predictions.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6}$$

Recall represents the sensitivity of the model. It defines that out of all actual positive predictions, how many samples are correctly predicted.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{7}$$

Here, during the evaluation of the proposed model, equal importance is given to classification in both HA and LA affective categories. Thus, the precision and recall are calculated for both categories as presented in Tables 4 and 5 to evaluate the performance of model for classifying both categories.

The same procedure is adopted to evaluate the model performance on the DREAMER dataset. An average accuracy of 90%-95% is reported during the single-subject classification model, and 90%-93% during the multi-subject classification model.

4.4 Findings and comparative analysis

EEG signals have been studied for a variety of purposes, including psychological state analysis, motor rehabilitation, stress analysis, etc. A lot of work has been done on reading EEG signals to infer a person's emotional state

also. It has been observed during the literature study, that most of the previous work is centered on analyzing the human emotions evoked by a particular emotional visual stimulus. The researchers have explored various optimal feature extraction approaches with the purpose of mapping the viewers' emotions with their physiological recordings or facial expressions. Whereas very few studies have been done on the emotive labeling of multimedia information using EEG data for understanding the fact that which type of content affects viewers the most. Furthermore, during processing, EEG signals were explored blindly with different combinations of classification methods for the automatic classification of human affective states. The major insight regarding the relationship between various important frequency ranges and brain regions is ignored while designing a model.

The major findings of the work presented in this paper are, first the relationship between multimedia content and physiological responses is addressed here with the aim of analyzing the two categories of video clips under the arousal category. The arousal emotional space represents the excitement level of the viewer. Thus, it is used to analyze the video on the basis of the fact that how much the viewer is attentive while watching video clips. Second, for effective modeling of EEG signals the relationship between various important frequency ranges and brain regions is explored and two different feature representation approaches are presented to further facilitate the CNN-based model preparation. The significance of the proposed effort is illustrated by a comparison to some of the most relevant publications in Table 6.

5 Conclusion

The categorization and labelling of visual content are activities that can be performed extremely easily and without much effort by humans. In addition, this visual content is typically designed with specific feelings in mind, which the creators hope to evoke in their viewers. Considering the emotional aspects of videos, in this paper, an intriguing application of EEG signals is presented to facilitate implicit affective tagging of video clips by assessing the human's cognitive state through EEG responses of viewers while watching that content. Furthermore, due to the importance of various brain regions, their communication and connection during cognitive tasks, and the information content in different frequency ranges, a contribution is made toward the effective modeling of EEG signals to explore the power of their spatial characteristics and the importance of different frequency ranges. In this paper two different effective representations of EEG signals are presented, i.e., spatial feature matrix and power spectral density maps. The proposed feature representations highlight the spatial features of EEG signals and are further used to train a CNN model for the implicit tagging of two categories of videos. The result shows that the two different representations provide better classification accuracy in comparison to the isolated representation of EEG signals at different brain locations. Here the simulations presented in this paper show the pioneering applicability of the proposed system for affective tagging of video based on human excitement level.

The current work presents the usage of EEG signals for analyzing the cognitive state of the participant with respect to visual stimuli, whereas the publicly available dataset contains other physiological recordings also such as ECG, GSR, etc., thus a combination of EEG signals with other modalities can also be explored for modeling of affective state. Also, in the future, the work can be extended by taking the physical features of video content also in account, where various audio-visual features of video content can be combined with EEG features to train a deep learning-based network. Furthermore, the current study utilized the pre-trained CNN model for the development of affective video tagging framework, which can be further extended by designing a dedicated model for EEG signal classification. Emotions represent affective experiences of one's feelings, thus the emotional aspect of video content can be considered a suitable criterion for their indexing and categorization. It can provide a subjective as well as high-level analysis of the video. In short, an affective analysis and tagging of video content promise a new direction toward the most popular problem areas in the multimedia community such as automatic video summarization, highlight extraction, indexing, etc., and motivating researchers for developing affective models.

Acknowledgements The authors are grateful to the database creators for granting us access to the dataset.

Author's contribution SS: Conceptualization, Methodology, Software, Data Curation, Validation, Writing- Original Draft Preparation. AKD: Conceptualization, Methodology, Supervision, Reviewing, and Editing. PR: Conceptualization, Supervision, Reviewing, and Editing. AR: Reviewing and Editing.

Funding This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data and material availability AMIGOS, DREAMER.

Code availability Custom code.

Declarations

Conflicts of interest Authors declare that they have no conflict of interest.

Ethical approval The article includes approaches used on datasets AMIGOS and DREAMER that are available to the public. According to the dataset description, participants gave the developers their written approval before participating.

Animal Rights There are no animal studies conducted by any of the authors in this article.

References

1. Caviedes JE (2012) The evolution of video processing technology and its main drivers. Proc IEEE 100(4):872–877. https://doi.org/10.1109/JPROC.2011.2182072
2. Pouyanfar S, Yang Y, Chen SC, Shyu ML, Iyengar SS (2018) Multimedia Big Data Analytics. ACM Comput Surv 51(1):1–34. https://doi.org/10.1145/3150226
3. Pereira F, Ascenso J, Brites C, Fonseca P, Pinho P, Baltazar J (2007) Evolution and Challenges in Multimedia Representation Technologies. In: M.S. Pereira (Ed) A Portrait of State-of-the-Art Research at the Technical University of Lisbon. Springer, Dordrecht, 2007. 275–294. https://doi.org/10.1007/978-1-4020-5690-1
4. Scherp A, Mezaris V (2014) Survey on modeling and indexing events. Multimedia Tools Appl 70:7–23. https://doi.org/10.1007/s11042-013-1427-7
5. Baveye Y, Chamaret C, Dellandréa E, Chen L (2018) Affective video content analysis: a multidisciplinary insight. IEEE Trans Affect Comput 9(4):396–409. https://doi.org/10.1109/TAFFC.2017.2661284
6. Hanjalic A, Xu L (2005) Affective video content representation and modeling. IEEE Trans Multimed 7(1):143–154. https://doi.org/10.1109/TMM.2004.840618
7. Assabumrungrat R, Sangnark S, Charoenpattarawut T, Polpakdee W, Sudhawiyangkul T, Boonchieng E, Wilaiprasitporn T (2022) ubiquitous affective computing: a review. IEEE Sens J 22:1867–1881. https://doi.org/10.1109/jsen.2021.3138269
8. Wang D, Zhao X (2022) Affective video recommender systems: a survey. Front Neurosci 16:984404. https://doi.org/10.3389/fnins.2022.984404
9. Slaney M (2011) Web-scale multimedia analysis: does content matter? IEEE Multimedia 18(2):12–15. https://doi.org/10.1109/mmul.2011.34
10. Dimitrova N, Zhang HJ, Shahraray B, Sezan I, Huang T, Zakhor A (2002) Applications of video-content analysis and retrieval. IEEE Multimed 9(3):42–55. https://doi.org/10.1109/MMUL.2002.1022858
11. Smith MA, Chen T (2005) 9.1: image and video indexing and retrieval. In: Bovik AL (ed) In: communications, networking and multimedia, handbook of image and video processing, 2nd edn. Academic Press, New York. https://doi.org/10.1016/B978-012119792-6/50121-2
12. Müller V, Boden MA (2008) Mind as machine: a history of cognitive science 2 vols. Mind Mach 18:121–125. https://doi.org/10.1007/s11023-008-9091-9
13. Hassanien AE, Azar A (2014) Brain computer interfaces: current trends and applications, intelligent systems reference library, vol 74. Springer, Cham
14. Ghaemmaghami P (2017) Information retrieval from neuro-physiological signals. Ph.D. Thesis. University of Trento. Canada
15. Zabcikova M, Koudelkova Z, Jasek R, Lorenzo Navarro JJ (2022) Recent advances and current trends in brain-computer interface research and their applications. Int J Dev Neurosci 82:107–123. https://doi.org/10.1002/jdn.10166
16. Alarcao SM, Fonseca MJ (2018) Emotions recognition using EEG signals: a survey. IEEE Trans Affect Comput. https://doi.org/10.1109/TAFFC.2017.2714671
17. Yang X, Yan J, Wang W, Li S, Hu B (2022) Lin J (2022) Brain-inspired models for visual object recognition: an overview. Artif Intell Rev 55:5263–5311. https://doi.org/10.1007/s10462-021-10130-z
18. Sharma S, Dubey AK, Ranjan P, Rocha A (2023) Neural correlates of affective content: application to perceptual tagging of video. Neural Comput & Applic 35:7925–7941. https://doi.org/10.1007/s00521-021-06591-6
19. Sharma S, Dubey AK, Ranjan P (2022) Affective video tagging framework using human attention modelling through EEG signals. International Journal of Intelligent Information Technologies (IJIIT) 18(1):1–18. https://doi.org/10.4018/IJIIT.306968
20. Gawali BW, Rao S, Abhang P, Rokade P, Mehrotra SC (2012) Classification of EEG signals for different emotional states. In: Fourth international conference on advances in recent technologies in communication and computing (ARTCom2012), pp 177–181. https://doi.org/10.1049/cp.2012.2521
21. Li J, Zhang Z, He H (2018) Hierarchical convolutional neural networks for EEG-based emotion recognition. Cogn Comput 10:368–380. https://doi.org/10.1007/s12559-017-9533-x
22. Hiyoshi-Taniguchi K, Kawasaki M, Yokota T, Bakardjian H, Fukuyama H, Cichocki A, Vialatte FB (2015) EEG correlates of voice and face emotional judgments in the human brain. Cogn Comput 7:11–19. https://doi.org/10.1007/s12559-013-9225-0
23. Frydenlund A, Rudzicz F (2015) Emotional affect estimation using video and EEG data in deep neural networks. In: Barbosa D, Milios E (eds) Advances in artificial intelligence. Canadian AI 2015. Lecture notes in computer science, vol 9091. Springer, Cham. https://doi.org/10.1007/978-3-319-18356-5_24
24. Hee Lin Wang (2006) Loong-fah cheong: affective understanding in film. IEEE Trans Circuits Syst Video Technol 16:689–704. https://doi.org/10.1109/tcsvt.2006.873781
25. Soleymani M, Pantic M (2013) Multimedia implicit tagging using EEG signals. In: 2013 IEEE international conference on multimedia and expo (ICME), San Jose, CA, USA, 2013, pp 1–6. https://doi.org/10.1109/ICME.2013.6607623
26. Koelstra S, Muhl C, Patras I (2009) EEG analysis for implicit tagging of video data. In: 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops. ACII 2009. IEEE, Amsterdam, Netherlands. pp 1–6 https://doi.org/10.1109/acii.2009.5349482.
27. Garg D, Verma GK, Singh AK (2023) A review of deep learning based methods for affect analysis using physiological signals. Multimed Tools Appl 82:26089–26134. https://doi.org/10.1007/s11042-023-14354-9
28. Vecchiato G, Cherubino P, Maglione AG, Ezquierro MT, Marinozzi F, Bini F, Trettel A, Babiloni F (2014) How to measure cerebral correlates of emotions in marketing relevant tasks. Cogn Comput 6:856–871. https://doi.org/10.1007/s12559-014-9304-x
29. Kumar S, Riddoch MJ, Humphreys G (2013) Mu rhythm desynchronization reveals motoric influences of hand action on object recognition. Front Hum Neurosci 7:66. https://doi.org/10.3389/fnhum.2013.00066
30. Sharma S, Mishra A, Kumar S, Ranjan P, Ujlayan A (2018) Analysis of action oriented effects on perceptual process of object recognition using physiological responses. In: Tiwary, U. (ed) Intelligent Human Computer Interaction. IHCI 2018. Lecture Notes in Computer Science. 46–58. doi: https://doi.org/10.1007/978-3-030-04021-5_5.
31. Padfield N, Zabalza J, Zhao H, Vargas VM, Ren J (2019) EEG-based brain-computer interfaces using motor-imagery: techniques and challenges. Sensors. https://doi.org/10.3390/s19061423

32. Kumar JS, Bhuvaneswari P (2012) Analysis of electroencephalography (EEG) signals and its categorization–a study. Procedia Eng 38:2525–2536. https://doi.org/10.1016/j.proeng.2012.06.298
33. Bigdely-Shamlo N, Vankov A, Ramirez RR, Makeig S (2008) Brain activity-based image classification from rapid serial visual presentation. IEEE Trans Neural Syst Rehabil Eng 16(5):432–441. https://doi.org/10.1109/TNSRE.2008.2003381
34. Wang J, Pohlmeyer E, Hanna B, Jiang YG, Sajda,P, Chang SF (2009) Brain state decoding for rapid image retrieval. In: Proceedings of the 17th ACM international conference on multimedia, pp 945–954. ACM, New York. https://doi.org/10.1145/1631272.1631463
35. Huang Y, Erdogmus D, Pavel M, Mathan S, Hild KE (2011) A framework for rapid visual image search using single-trial brain evoked responses. Neurocomputing 74(12):2041–2051. https://doi.org/10.1016/j.neucom.2010.12.025
36. Lees S, Dayan N, Cecotti H, McCullagh P, Maguire L, Lotte F, Coyle D (2018) A review of rapid serial visual presentation-based brain- computer interfaces. J Neural Eng 15(2):021001. https://doi.org/10.1088/1741-2552/aa9817
37. Kapoor A, Shenoy P (2008) Combining brain computer interfaces with vision for object categorization. In: 2008 IEEE conference on computer vision and pattern recognition, pp 1–8. https://doi.org/10.1109/CVPR.2008.4587618
38. Mohedano E, Healy G, McGuinness K, Giró-i-Nieto X, O'Connor NE, Smeaton AF (2014) Object segmentation in images using EEG signals. In: Proceedings of the 22Nd ACM international conference on multimedia, pp 417–426. ACM, New York. https://doi.org/10.1145/2647868.2654896
39. Mohedano E, McGuinness K, Healy G, O'Connor NE, Smeaton AF, Salvador A, Porta S, Nieto XG (2015) Exploring EEG for object detection and retrieval. In: Proceedings of the 5th ACM on international conference on multimedia retrieval, pp 591–594. ACM, New York. https://doi.org/10.1145/2671188.2749368
40. Healy G, Smeaton AF (2011) Optimising the number of channels in EEG-augmented image search. In: Proceedings of the 25th BCS conference on human–computer interaction, pp 157–162. British Computer Society, Swinton
41. Tauscher JP, Mustafa M, Magnor M (2017) Comparative analysis of three different modalities for perception of artifacts in videos. ACM Trans Appl Percept. https://doi.org/10.1145/3129289
42. Mutasim AK, Tipu RS, Bashar MR, Amin MA (2017) Video category classification using wireless EEG. In: Zeng Y, He Y, Kotaleski JH, Martone M, Xu B, Peng H, Luo Q (eds) Brain informatics. Lecture notes in computer science, vol 10654. Springer, Cham, pp 39–48. https://doi.org/10.1007/978-3-319-70772-3_4
43. Soleymani M, Lichtenauer J, Pun T, Pantic M (2012) A multimodal database for affect recognition and implicit tagging. IEEE Trans Affect Comput 3(1):42–55. https://doi.org/10.1109/T-AFFC.2011.25
44. Wang S, Zhu Y, Wu G, Ji Q (2014) Hybrid video emotional tagging using users' EEG and video content. Multimed Tools Appl 72:1257–1283. https://doi.org/10.1007/s11042-013-1450-8
45. Martínez-Rodrigo A, García-Martínez B, Huerta Á, Alcaraz R (2021) Detection of negative stress through spectral features of electroencephalographic recordings and a convolutional neural network. Sensors 21:3050. https://doi.org/10.3390/s21093050
46. Mishra A, Ranjan P, Ujlayan A (2020) Empirical analysis of deep learning networks for affective video tagging. Multimed Tools Appl 79:18611–18626. https://doi.org/10.1007/s11042-020-08714-y
47. Jang S, Moon S-E, Lee J-S (2018) Eeg-based video identification using graph signal modeling and graph convolutional neural network. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) Calgary, AB, Canada, 2018, pp. 3066–3070. https://doi.org/10.1109/icassp.2018.8462207.
48. Correa JAM, Abadi MK, Sebe N, Patras I (2018) AMIGOS: a dataset for affect, personality and mood research on individuals and groups. IEEE Trans Affect Comput 12(2):479–493. https://doi.org/10.1109/TAFFC.2018.2884461
49. Katsigiannis S, Ramzan N (2018) DREAMER: a database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices. IEEE J Biomed Health Inform 22:98–107. https://doi.org/10.1109/jbhi.2017.2688239
50. Koelstra S, Mühl C, Soleymani M, Jong-Seok L, Yazdani A, Ebrahimi T, Pun T, Nijholt A, Patras I (2012) Deap: a database for emotion analysis; using physiological signals. IEEE Trans Affect Comput 3(1):18–31
51. Abadi MK, Subramanian R, Kia SM, Avesani P, Patras I, Sebe N (2015) DECAF: MEG-based multimodal database for decoding affective physiological responses. IEEE Trans Affect Comput 6(3):209–222. https://doi.org/10.1109/TAFFC.2015.2392932
52. Mallat S (1989) A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans Pattern Anal Mach Intell 11(7):674–693. https://doi.org/10.1109/34.192463
53. Kehtarnavaz N (2008) Chapter 7: frequency domain processing. In: Kehtarnavaz N (ed) Digital signal processing system design, 2nd edn. Academic Press, London, pp 175–196. https://doi.org/10.1016/B978-0-12-374490-6.00007-6
54. Vivas EL, García-González A, Figueroa I, Fuentes RQ (2013) Discrete wavelet transform and ANFIS classifier for brain-machine interface based on EEG. In: 2013 6th international conference on human system interactions (HSI), pp 137–144. https://doi.org/10.1109/HSI.2013.6577814
55. Subasi A (2007) EEG signal classification using wavelet feature extraction and a mixture of expert model. Expert Syst Appl 32:1084–1093. https://doi.org/10.1016/j.eswa.2006.02.005
56. Welch P (1967) The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. IEEE Trans Audio Electroacoust 15(2):70–73. https://doi.org/10.1109/TAU.1967.1161901
57. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.
58. Asghar MA, Fawad Khan MJ, Amin Y, Akram A (2020) EEG-based Emotion Recognition for Multi-Channel Fast Empirical Mode Decomposition using VGG-16. International Conference on Engineering and Emerging Technologies (ICEET). Lahore, Pakistan, 2020, pp. 1–7. https://doi.org/10.1109/iceet48479.2020.9048217