

Received August 27, 2020, accepted September 7, 2020, date of publication September 14, 2020, date of current version September 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3023871

Cross-Subject Multimodal Emotion Recognition Based on Hybrid Fusion

YUCEL CIMTAY¹, ERHAN EKMEKCIOGLU¹, AND SEYMA CAGLAR-OZHAN²

¹Institute for Digital Technologies, Loughborough University London, London E20 3BS, U.K.

²Department of Computer Education and Instructional Technology, Hacettepe University, 06800 Ankara, Turkey

Corresponding author: Erhan Ekmekcioglu (e.ekmekcioglu@lboro.ac.uk)

This work was supported by an Institutional Links grant, ID 352175665, under the Newton–Katip Celebi partnership between the U.K. and Turkey. The grant is funded by the U.K. Department of Business, Energy and Industrial Strategy (BEIS) and The Scientific and Technological Research Council of Turkey (TUBITAK) and delivered by the British Council. For further information, please visit www.newtonfund.ac.uk.


ABSTRACT Multimodal emotion recognition has gained traction in affective computing research community to overcome the limitations posed by the processing a single form of data and to increase recognition robustness. In this study, a novel emotion recognition system is introduced, which is based on multiple modalities including facial expressions, galvanic skin response (GSR) and electroencephalogram (EEG). This method follows a hybrid fusion strategy and yields a maximum one-subject-out accuracy of 81.2% and a mean accuracy of 74.2% on our bespoke multimodal emotion dataset (LUMED-2) for 3 emotion classes: sad, neutral and happy. Similarly, our approach yields a maximum one-subject-out accuracy of 91.5% and a mean accuracy of 53.8% on the Database for Emotion Analysis using Physiological Signals (DEAP) for varying numbers of emotion classes, 4 in average, including angry, disgust, afraid, happy, neutral, sad and surprised. The presented model is particularly useful in determining the correct emotional state in the case of natural deceptive facial expressions. In terms of emotion recognition accuracy, this study is superior to, or on par with, the reference subject-independent multimodal emotion recognition studies introduced in the literature.

INDEX TERMS Emotion recognition, multimodal emotion recognition, multimodal data fusion, convolutional neural network, electroencephalogram, galvanic skin response.

I. INTRODUCTION

The study of emotion has a long history. Emotions have been a subject of philosophy long before it was covered in other disciplines [1]. Recognising emotions can help understand and interpret human behaviours. The field of affective computing emerged for emotion recognition using various data sources and leveraging computer-based environments [2]. Affect recognition typically requires tracking and measuring data sources and processing them for estimating emotions. There are various emotion recognition studies published that use multiple sources of data (modalities) [3]–[6].

The literature features different emotion models. Authors in [7] outline a review of them considering nine different studies and in total there are 65 different emotion categories. Emotion can be represented by a continuous 4-D space of valence, arousal, dominance and liking [8].

The associate editor coordinating the review of this manuscript and approving it for publication was Sung Chan Jun .

In many studies the 4-D space is reduced to a 2-D space as valence and arousal [9] to represent emotions. The study by Ekman *et al.* [10] yielded 6 basic emotions: anger, disgust, fear, happiness, sadness, and surprise. These emotional states are widely agreed to be universal across cultures and races. Other emotions map on to different points on the valence-arousal scale. A 2-D normalised emotional circumplex model introduced in [11] is shown in Figure 1. In [11], the valence and arousal values were determined by analysing a large number of blog posts. The basic emotion states are highlighted on Figure 1.

Emotion recognition can be performed by analysing the face, body language or speech, which are the common modalities, although they are susceptible to deliberate deception. On the other hand, physiological signals originating from internal bodily reactions are less affected by deceptions. Electroencephalogram (EEG) and Functional Magnetic Resonance Imaging (fMRI), which measure brain electrical or blood flow activity, are two examples. Other examples

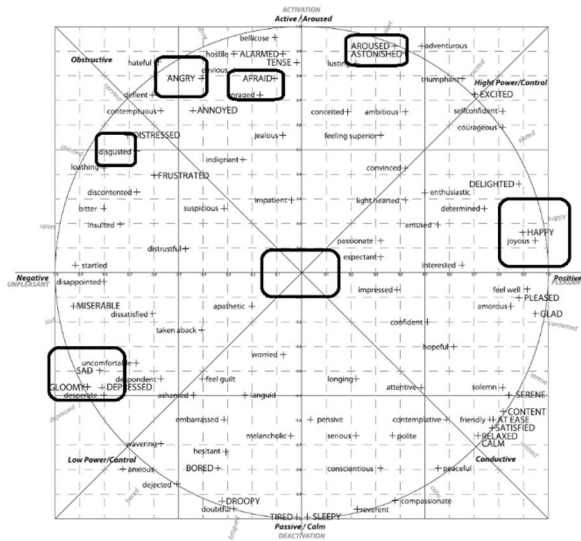


FIGURE 1. 2-D valence-arousal circumplex model [11].

that can be measured using modern consumer-grade sensors include heartbeat, galvanic skin response, blood volume pulse, respiration and temperature.

Emotion classification techniques typically leverage machine learning methods, such as Support Vector Machine (SVM), linear and nonlinear regression, decision trees and K-Nearest neighbour (KNN), which result in a range of classification accuracies. More recently, deep learning models like Convolutional Neural Networks (CNN), Long-Short Term Memory (LSTM) and Convolutional Long-Short Term Memory (CLSTM) have been studied in emotion classification [3], [6]. In deep learning models, unprocessed sensor data can be used directly by the network without pre-extracting features. Deep learning-based models can achieve high accuracy recognition rates especially if there is a high volume of accurately labelled data.

Multimodal emotion recognition aims to combine the predictive capabilities of individual behavioural and biometric traits for accurate classification. This bears a greater complexity than unimodal emotion recognition systems due to the need for jointly processing of multiple data source. Even within multimodal approaches, there is a high degree of variation in terms of prediction accuracy, which necessitates the design of robust approaches. With this aim, in this paper we propose a novel hybrid multimodal emotion recognition model, utilising both soft- and hard-biometric data and test its performance using existing datasets and also on a newly created multimodal dataset.

The remainder of this paper is structured as follows: Section II provides an overview of the relevant literature in the area of multimodal emotion recognition. Section III gives a detailed account of the proposed approach. Datasets used in the process of training and testing are described in Section IV and in Section V we provide the test results with the discussions. Finally, Section VI concludes the paper.

II. BACKGROUND AND LITERATURE REVIEW ON MULTIMODAL EMOTION RECOGNITION

Several unimodal and multimodal emotion recognition studies have been reported in the literature [12]–[16]. Majority of studies use non-physiological data, such as audio, video and text [17]. One of the difficulties with non-physiological modalities is that someone may disguise their actual emotion and not reveal enough cues about their actual state. More recently, researchers are showing an increasing interest in brain signals and peripheral physiological signals due to their nature of irrevocability [18]. However, these are prone to measurement artefacts and noises. A multimodal emotion recognition system reinforces the overall recognition robustness, where temporary defects and problems in some modalities are compensated for by other modalities.

The study in [19] proposes a fusion model based on verbal and nonverbal information to increase the emotion recognition capability of children companion robots. The study in [20] created a multimodal emotion database that includes the visual, audio, physiological, depth and pose data of 60 participants. Researchers in [21] collected participants' facial expressions, heart rate, pupil diameter and EEG data by showing them specific visual stimuli. They reported that the emotion recognition accuracy of models developed based on multimodal features is superior to those based on single modality.

In an application context, the study in [22] focused on two sentiment types: semantic and contraption erudition. Authors developed algorithms to contribute to the business perception and consequences in product improvement. A sentiment analysis technique based on Convolutional Neural Networks for imagery was introduced in [23]. This study leverages transfer leaning and the developed model performs reasonably well under scarce training data condition. The study in [24] reviewed the sentiment analysis approaches like dynamic dictionary handling, lexical variations, and sarcasm sentiment analysis. In [25] researchers applied an end-to-end deep neural network for learning with both text and image demonstrations.

The study in [26] focuses on estimating the depression levels. Authors use multitask modality encoders that get individual modalities' features as inputs and give modality embeddings as outputs. An attention-based fusion network is applied for the fusion of individual modalities. At the final step, a deep network is applied. In [27], a multimodal sentiment analysis by using textual, visual and audio features was proposed. Authors used a CNN model to extract textual features, and a 3D-CNN model to extract audio and visual features. The features coming from each modality are concatenated and for final classification, Support Vector Machine (SVM) is applied. The study in [28] describes a generative Unigram Mixture Model (UMM). This model provides learning a word-emotion society lexicon by using input from a document corpus. They compare their proposed and state-of-the-art baseline methods on word emotion classification and document emotion ranking. In [29], a

Multimodal Dictionary is presented to understand the relationship between the spoken words and facial gestures when introducing sentiment. This approach improves the prediction accuracies in speaker independent sentiment intensity analysis. Multiple other recent studies under the topic of unimodal and multimodal sentiment or emotion analysis are reported in [30]–[41].

Data fusion is a critical step involved in multimodal emotion recognition for producing the estimation. The literature about emotional data fusion involves three data fusion techniques, which are early fusion (feature fusion) [42], [43], late fusion (decision fusion) [44]–[46] and hybrid approaches [17], [47], [48].

In feature level fusion, features are extracted from each modality and one feature set is created by combining all modalities. This combined feature set is used for the training of emotion recognition model. Some of the studies which use feature fusion are [42], [43], [49]. In [42], EEG and audio-visual feature sets are fused. It mentions that using the fusion of full-band EEG power spectrum and video audio-visual features achieves the best recognition accuracies. It reports 96.8% classification accuracy for positive/negative valence and 97.8% for High/Low arousal. The study in [43] integrates low-level audio-visual features extracted from videos and brain functional activity measured by magnetic resonance imaging (fMRI). For classification they use multimodal deep Boltzmann machine. In [49], the researchers use Fisher Criterion Score and Davies-Bouldin index feature selection methods in order to select significant multimodal features from physiological data. They use HMM (Hidden Markov Model) for the classification of valence and arousal.

Feature level fusion is an imitation of human mechanism of emotion recognition, which is based on collecting all sensory information from different modalities and combining them. There is no isolated emotion classification done using any individual modality, and the classification is based on a combined multimodal feature set. The limitation for feature level fusion is that once a model is trained using a specific combination of feature sets, all future test data should have exact same feature structure with no tolerance to data loss. Any missing modality or feature will result in a failed classification. One possible solution is maintaining multiple models trained using different combinations of features (or modalities) to create backups and prevent failures resulting from missing data. Or, missing data should be recovered before creating test samples, as done in [6].

In [6], different missing data techniques were applied by using early fusion with deep recurrent networks. The study in [3] uses Deep Belief Networks (DBN) for doing a feature level fusion of audio, face video, body video and physiological data. Authors in [50] also apply feature level fusion of face video and audio modalities and classify the emotional state using a 2-Layer LSTM.

The study in [4] uses EEG and eye movement modalities to implement a feature level fusion by using SVM and Deep

neural networks. It reports that the mean recognition accuracy is increased combining feature fusion and deep neural network compared to combining feature fusion with SVM. In [18], EEG and peripheral physiological signals were used as modalities. It applies feature fusion with SVM and bimodal LSTM classifiers. Comparing to feature fusion with SVM, it reports that bimodal LSTM has achieved the maximum recognition accuracy at 93.97% on Shanghai Jiao Tong University Emotion EEG Dataset (SEED) [46].

Decision level fusion combines the resulted emotion labels coming from each classifier that use different modalities. In this type of fusion, each modality leads to an independent emotional output. These outputs can be used separately or jointly through machine learning methods. Some studies have implemented decision level fusion as reported in [44], [45], [49], [52], [53].

The study in [44] a decision level fusion is applied for audio and visual data to identify emotions. The proposed method is applied on eNTERFACE'05 database [54]. The study in [49] uses face, voice and head movement modalities for emotion recognition. It conducts a late classification by using a Bayesian framework. It reports that the decision level fusion is successful especially for detecting happiness. The fusion strategy increases the average accuracy, from 55% to about 62% comparing to unimodal application.

In [52], the scientists propose a decision level multimodal emotion recognition based on EEG and face modalities. They use a stimulus which is based on a subset of clips that correspond to four specific emotions: happiness, neutral, sadness, and fear on the valence-arousal emotional space. For facial expression recognition, these emotion states are detected by a neural network classifier. For EEG data recognition, four basic emotion states and three emotion intensity levels (strong, ordinary, and weak) are detected by support vector machines (SVM). They apply a sum and production rule for fusion of the unimodal results. They report that the mean recognition accuracy is 82% for multimodal method which is higher than the accuracies of 74.38% and 66.88% of facial expression and EEG modalities respectively.

The advantage of decision level fusion is that when any of the modalities is missing, the decision can be made by using the other modalities. However, this requires an intelligent system, which can detect the missing modalities. One of the earliest decision level fusion strategies uses a voting technique [55]. In this strategy, the classification state reached by most of the modalities was ultimately agreed on and adopted. The drawback of this strategy is the likely tie situations. Another approach, which avoids possible ties and increases the accuracy, is to use the prior knowledge about each modality. The prior knowledge includes the classification confusion matrix for each modality and for each method used. According to the known prior classification accuracies, for each emotional state, a weighted voting strategy can be used. Voting weights can be calculated from the recognition rate or error rates that each classifier has shown with the training or test data [56].

Beyond the use of voting schemes, the study in [57] uses a lookup table during training to record the classification combinations, classification output, correct labels and the number of occurrences of combinations. The classification confidence of combinations is measured with the number of the occurrences. The outcome with the highest level of confidence in the lookup table is chosen.

The study in [58] computes confidence for each modality and weights the unimodal decisions by the measured quality of raw signals. This operation decreases the classification error rate and increases the quality of recognition [58]. In [59], Relevance Vector Machines (RVM), which correspond to embedded SVMs are used. RVMs resemble SVM but run in an embedded Bayesian frame. It calculates the affiliation probabilities of each emotion category across the classifiers. These probabilities are used as classification weights for decision level fusion.

The study in [60] uses a decision fusion strategy, which combines the emotion outputs of visual and audio paths using KNN or Artificial Neural Networks (ANN). The study in [12] is one of the first examples of approaches combining audio, face and physiological data by applying decision level fusion. It reports the Concordance Correlation Coefficient (CCC) metric result for arousal and valence and concludes that multimodality increases the emotion recognition accuracy.

In [52], a decision level fusion was applied to outcomes from EEG and face modalities. It applied SVM classification for EEG data and Neural Network for face data. The emotional states are grouped according to the intensity levels. According to the labels of training data (happiness, neutral, sadness, fear) it tried to find the optimal weights of each modality in obtaining the final decision.

Another fusion technique is called hybrid fusion, which typically combines both the early- and late-fusion techniques. For instance, one classifier may deploy a feature-level fusion for the face and body gestures modalities, while another classifier may do so for physiological signals. Another decision level classifier above these two classifiers can process the results of two feature level classifiers to come up with the final emotion label. Hence, such a system is called a hybrid fusion system.

In [61], a simple hybrid fusion was employed where the output of an early fusion classifier is feeding input to a decision-level fusion system. A recent study in [48] uses a latent space map for the fusion of audio and video modalities; and then, by using a Dempster-Shafer (DS) theory-based evidential fusion method, the projected features on the cross-modal space are fused with the textual modality.

Multimodal techniques have also been applied in cross-subject and subject-independent emotion recognition studies. The study in [62] uses physiological signals with feature level fusion. Authors trained a deep CNN model and achieved 94% subject-independent average accuracy on BP4D+ dataset [63] for 10 emotion classes. The study in [64] uses speech and body motions modalities with two stages Gaussian Mixture Model (GMM) mapping framework.

It achieves a maximum accuracy of 63%, 51%, 50% for valence, arousal and dominance, respectively, on USC CreativeIT multimodal emotion database [65].

The study in [66] uses a decision fusion on physiological data including photoplethysmography (PPG), a respiratory belt (RB) and fingertip temperature (FTT) and conducts one-subject-out test on the DEAP dataset. It achieves 72.18% and 73.08% mean accuracies on high-low valence and high-low arousal, respectively. The study in [67] uses a three-stage decision method for classifying four emotions on the DEAP dataset. It achieves 77.57% average accuracy when classifying the labels as high- and low- arousal, and 43.57% when classifying as high- and low- valence.

A. LIMITATIONS OF THE REVIEWED MULTIMODAL EMOTION RECOGNITION STUDIES

While the published studies report varying recognition and classification accuracies under varying contexts, we noted some common limitations applying to a multiple of these studies. They are summarised below.

- Studies that exploit facial expression analysis for emotion detection broadly assume that the participants do not show deceptive expressions, while this is not always the case. DEAP dataset is a good example for this. Visual-based multimodal emotion recognition studies tend to yield inferior accuracy results with this dataset.
- Multimodal emotion recognition studies generally do not discriminate between the predictive capabilities of individual modalities. For example, GSR is effective in predicting the arousal, but not valence. Incorporating all modalities equally in a model could reduce the recognition accuracy.
- The effects of emotion transitions are not reflected on physiologic outputs with the same delay. Yet, many reviewed studies jointly process multiple modalities measured at co-located time windows that can limit the achievable recognition accuracy.

B. CONTRIBUTIONS OF THIS WORK

- We employ subsets of modalities and features derived from them, in association with the targeted dimension, i.e., valence or arousal. Our decision to select the subsets of modalities is informed by the former studies.
- The presented emotion recognition model is made robust against natural deceptive facial expressions by comparing facial-expression based dominant emotions against emotions inferred via other modalities.
- We adopt a hybrid approach composed of feature- and decision-level fusion. Physiologic modalities measured at varying time instances are jointly processed for improved accuracy.
- Our method is tested on multiple datasets, which include instances of deceptive facial expressions.
- The feature extraction capability of a pretrained CNN model (InceptionResnetV2 [61]) is utilised to eliminate the need for manual feature extraction.

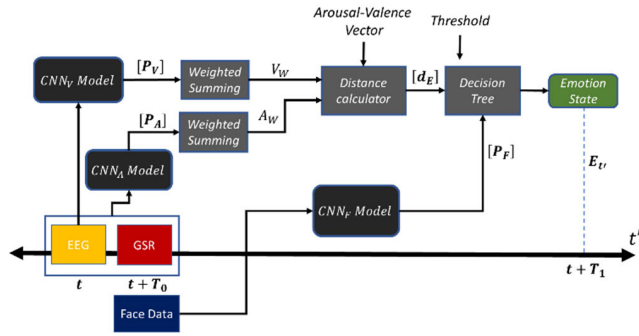


FIGURE 2. Proposed hybrid multimodal emotion recognition method.

III. PROPOSED METHOD

In this study a new multimodal emotion recognition method is introduced, which uses face, EEG and GSR data modalities. The overall system diagram is shown in Figure 2, which will be described later in detail. In this study, to detect the emotional state, the valence probability array coming from EEG modality at time t , the arousal probability array coming from the combined data of EEG and GSR at time instant $t + T_0$ and, the discrete emotion probability array coming from face modality (P_F) at time $t + T_0$ are used.

The reason we introduce a time delay T_0 between EEG sample and other modalities' samples is due to the fact that physically the emotion signal first emerges in the brain and the brain sends the exciter signal to trigger other modalities with some delay. T_1 depends on the processing power of used hardware. Empirically we have chosen T_0 as 0.35 sec. For prediction of arousal we use EEG and GSR raw data together, and for valence we only use EEG data. There are three separate InceptionResnetV2 CNN models for arousal, valence and face-based discrete emotion, respectively.

A. FACIAL EXPRESSION ANALYSIS

The face data used in this study for training our multimodal emotion recognition model is composed of well-known public face imagery datasets including Cohn-Kanade+ face dataset [69], Radboud Faces Database [70], FacesDB [71] and AffectNet [72]. Since we get use of the spatial information contained in the images, for the purpose of training, rather than video, static images are preferred. For each of the datasets used, all the facial images were manually double-checked to identify and leave out the samples where we don't have a consensus with the associated emotion label. This filtering step is applied in our work to minimise the influence of training samples with potentially wrong or arguable labels on the classifier performance. Table 1 shows the percentage of eliminated labelled facial image samples from each dataset following the application of the filtering step.

Following the filtering step, a final dataset is obtained by combining all labelled samples from the four datasets. The number of face images belonging to each emotion category is given in Table 2.

TABLE 1. Percentage of eliminated images from each dataset.

Name of Dataset	Percentage of Eliminated Images
Cohn Kanade [69]	17
FacesDB [71]	12
Radboud [70]	14
AffectNet [72]	22

TABLE 2. Number of labelled face images used in training.

Emotion Category	#Images
Angry	1117
Disgust	1173
Afraid	784
Happy	2417
Neutral	1401
Sad	874
Surprised	1126
Total	8898

It is important to note that this combined dataset features labelled face image samples taken using different camera settings (angle of shooting, camera zoom level, resolution, average brightness) in order to handle different conditions. We don't use the face data of LUMED-2 [73] and DEAP [74] dataset, since they are the test datasets for this study. Face modality is used in the form of still images of faces. A pre-trained state-of-the-art InceptionResnetV2 CNN model is trained again using the previously discussed and refined face datasets as input, and their associated discrete emotion labels as output. The reason for choosing InceptionResnetV2 is due to its proven capability of yielding one of the highest recognition accuracies in the image classification context [75]. The initial network weights were adopted from ImageNet training. Hence, it is called pre-trained. The parameters of training are given in Table 3.

This CNN model is trained with 8898 face images (see Table 2) in the emotion categories of angry, disgust, afraid, happy, neutral, sad and surprised. Image data augmentation and normalisation is applied on the training set prior to training in order to handle different shifting, zooming and lighting conditions.

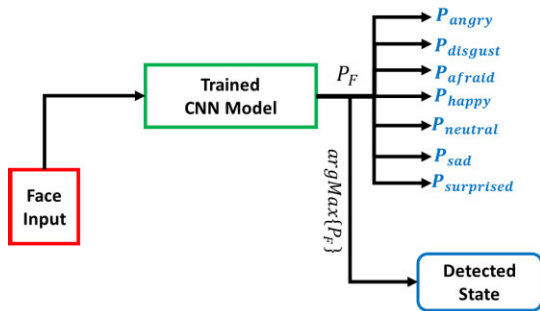
The outputs of trained model CNN_F in Figure 3 are the probability vectors of seven discrete emotion categories (based on the Ekman's model, and additionally the neutral emotion state) and following the application of the argmax function, the final emotion state is yielded. The diagram of emotion classification based on the face data modality is shown in Figure 3.

B. GALVANIC SKIN RESPONSE (GSR) ANALYSIS

Second modality used in the proposed hybrid multimodal architecture is the GSR data. When humans are exposed to an

TABLE 3. Training parameters of face network.

Property	Value
Base Model	InceptionResnetV2
#Images	8898
#Extra Layers	3 Dense layers with depth of 1024
Rotation range	10
Width-Shift Range	0.1
Height-Shift Range	0.1
Zoom Range	0.2
Horizontal Flip	True
Regularization	L2
Optimizer	Adam
Loss	Categorical cross entropy
#Epochs	30
Shuffle	True
Batch Size	64
#Classes	7
Environment	Win 10, Parallel GPU (2 GPUs), TensorFlow

**FIGURE 3.** Face emotion recognition model.

event or stimuli, the emotion is triggered first in the brain [76]. The brain signals start to change and send exciter signals to the other parts of the body.

For instance, face expression and voice tone change. Another important change occurs in physiological signals. One of them is GSR, which is a measure of change in the electrical resistance of the skin. When one is emotionally aroused, the electrical conductivity of the skin changes. GSR, which is also known as EDA (Electrodermal activity) or SC (Skin Conductance), is one of the most important measures of emotional arousal [77].

GSR measures the skin secretion that is an unconscious process under the control of body's sympathetic nervous system and reflects the changes in arousal. When people are aroused due to various effects like fear, joy or stress, skin starts to sweat. GSR has been investigated in many studies and it is still regarded as one of the most powerful methods for understanding and measuring physiological arousal.

GSR has been used in a wide range of research activities, which include but not limited to physiological research, clinical research and psychotherapy, consumer neuroscience and

marketing, media and ad testing, usability testing and User Experience (UX) design [78]. Under normal conditions, for a healthy individual GSR is near stable. Once an individual is exposed to some arousing stimuli, more frequent changes in the GSR data start to be observed. Following to the transition to a calmer state, GSR change activity decreases. Since GSR is highly linked to one's level of arousal, we use this modality together with the EEG modality for predicting the arousal level of individuals.

C. ELECTROENCEPHALOGRAM (EEG) ANALYSIS

One way of measuring the brain activities is using the EEG technology. The change of the electric potential formed on the skull is measured by using actual and reference electrodes. Some commercially available EEG devices in the consumer market include Emotiv, Neurosky, and Neuroelectronics [79]. These devices have various spatial and temporal resolutions. Spatial resolution is related to number of electrodes placed on the head and the temporal resolution is related to the number of electric potential changes recorded in a second. Generally, EEG has a low spatial, but a high temporal resolution compared to other brain monitoring technologies such as fMRI and functional near-infrared spectroscopy (fNIRS). There are different electrode positioning standards for EEG, such as 10-20, 10-10 and 10-5 that result in different spatial resolutions [80].

EEG is preferred in this work due to its reliability and portability compared to other brain monitoring technologies. Clinical applications have traditionally been one of the application areas of EEG. EEG has been used to investigate the signal patterns related to epilepsy [81] and sleep [82]. Detection of the hyperactivity and consciousness related disorders [83], [84], measurement of mental workload [85], level of attention [86]–[88], mood and emotions [89]–[91] have been the other application areas of EEG.

Although EEG has gained popularity in the context of emotion recognition studies due to its resistance to deceptive actions of humans, it exhibits varying distributions for different people as well as for the same person at different time instances [92], [93]. This is a problem, which decreases the accuracy of emotion recognition for subject-independent applications. One way of mitigating this kind of a problem is decreasing the number of emotion classes by grouping them and using complex techniques on feature extraction to increase the accuracy [94]. In this study we leveraged the research outputs of our previous study in [94]. Although the subject-independent emotion recognition accuracy of EEG is relatively low, when the number of emotion categories is reduced to two (e.g., high and low valence/arousal), the accuracy increases accordingly.

Our previous study on EEG-based emotion recognition [94] was a leave-one-subject-out cross-validated study. This is depicted in Figure 4a, where the outputs of EEG-based emotion recognition system are the probabilities of high- and low- valence (P_{V-H} and P_{V-L}) considering two classes.

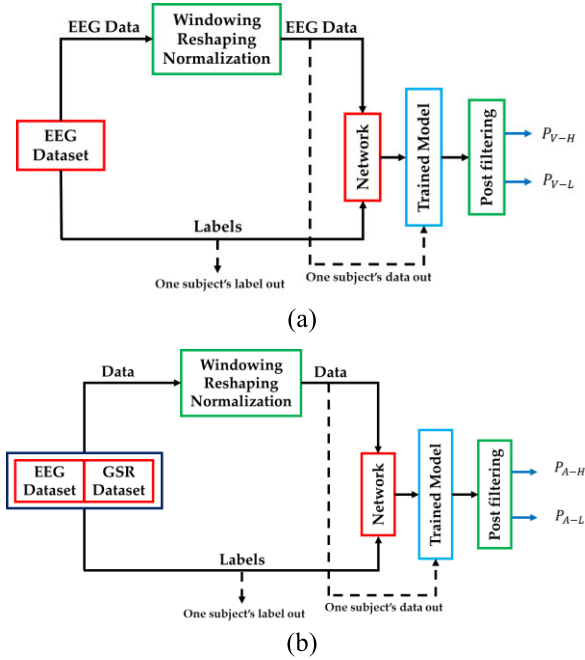


FIGURE 4. (a) One-subject-out valence recognition model, (b) One-subject-out arousal recognition model.

TABLE 4. Training parameters of EEG and GSR+EEG network.

Property	Parameters
Base Model	InceptionResnetV2
Additional Layers	Global Average Pooling, 5 Dense Layers
Regularization	L2
Optimizer	Adam
Loss	Categorical cross entropy
Max. #Epochs	100
Batch Size	64
#Classes	2 (H-L Valence/Arousal)
Environment	Win 10, Parallel GPU (2 GPUs), TensorFlow

In order to increase the overall accuracy, in the proposed method, we do a pre-grouping on the valence and arousal labels in order to reduce the total number of output classes. We group valence as high-/low- valence and arousal as high-/low- arousal. In this study, for estimation of arousal, we have combined the prediction capabilities of both the GSR and EEG modalities.

The outputs of the network that relies on the early fusion of raw GSR and EEG are the probabilities of high- and low-arousal (P_{A-H} and P_{A-L}), as shown in Figure 4b. The training parameters of EEG network in Figure 4a, and the network combining both EEG and GSR modalities in Figure 4b are given in Table 4. The parameters used are the same. Note that, like in the case of the face modality, for the physiological signals we also used the InceptionResnetV2 Convolutional Neural Network as the underlying network architecture.

The outputs of CNN_A in Figure 3, are the probability vectors of high- and low- arousal. Similarly, the outputs

of CNN_V are the probability vectors of high- and low-valence. The outputs of CNN_V and CNN_A are fed into a weighting unit. This unit calculates the weighted sums of valence and arousal, respectively. Weighted valence V_W and weighted arousal A_W are then sent to the distance calculator. Distance calculator calculates the emotional distance from the weighted valence-arousal to actual valence-arousal pairs, which correspond to the seven discrete emotion states. Emotional distances are fed to the final decision tree system. The output of CNN_F is fed to decision tree system at this stage too.

The decision tree system, which is explained in more detail below, outputs the final computed emotional state at time $t + T_1$. Distance calculator calculates the emotional distance from the weighted valence-arousal to actual valence-arousal pairs, which correspond to the seven discrete emotion states. Emotional distances are fed to the final decision tree system. The output of CNN_F is fed to decision tree system at this stage too. The decision tree system, which is explained in more detail below, outputs the final computed emotional state at time $t + T_1$. The reason we put a time delay T_0 between EEG sample and other modalities' samples is due to the fact that physically the emotion signal first emerges in the brain and the brain sends the exciter signal to trigger other modalities with a of delay.

D. DECISION TREE

At the core of the proposed method lies the function of a decision tree. Here we put the emphasis first on the probability vector of the face modality-based emotion detection system. We consider not only the dominant emotion (i.e., the one with the maximum probability), but also the emotion with the second highest probability. This is because individuals may hide their actual emotion and the likelihood of the second most probable emotion state being the actual emotion state increases. Also, face modality-based emotion systems may sometimes lead to false detections in terms of the emotional states in the case of confusing gestures. At this point, the other physiological modalities (i.e., GSR and EEG) help detect the emotional state with better accuracy.

EEG is a nonstationary signal and shows different distributions from subject to subject. Therefore, it is not very successful in terms of predicting the discrete emotional state or the valence and arousal when subject-independency is sought [92], and especially when the number of output classes is relatively high. GSR on the other hand is mostly a measurement of arousal and doesn't give useful information about the valence dimension.

We set the normalised centre values of high and low as -0.75 and 0.75, respectively. These centre values for low and high can change according to the choice of low and high intervals. In our case, the resulted valence value is the average of valence vector $[P_V]$ and arousal is the average of arousal vector $[P_A]$. We also use the associated (mapped) valence and arousal values of the used seven discrete emotion states, shown in Figure 1.

These are empirical values previously introduced in [11]. These values ([valence, arousal]) include: [0,0], [0.89,0.17], [-0.81,-0.40], [-0.68,0.49], [-0.40,0.79], [-0.12,0.79], [0.42,0.88] for emotion states 'Neutral', 'Happy', 'Sad', 'Disgust', 'Angry', 'Afraid', and 'Surprised', respectively.

In the decision tree, if the maximum probability in $[P_F]$ vector coming from face modality based network is over a pre-defined threshold that is empirically set to 0.9, we take the final emotion state as the one discrete emotion with maximum probability value in $[P_F]$. If the maximum probability is below the threshold, then we use the distances from the weighted valence-arousal to each valence-arousal pair of discrete emotion states (i.e., output of distance calculator). We compare the emotional distances of the emotion states with the highest and second highest probabilities in the $[P_F]$ vector. We check whether the emotional distance of the emotion state with second highest probability in $[P_F]$ is higher than the one with highest probability. If it is higher, then we take the one with second highest probability in $[P_F]$ as the final emotional state. Otherwise, we assign the discrete emotional state with the highest probability in $[P_F]$ as the final emotional state output.

E. THE OVERALL HYBRID FUSION WORKFLOW

The proposed method is suitable for real-time operation. To summarise the overall hybrid fusion based multimodal emotion recognition technique, which uses face data, GSR and EEG inputs, a pseudocode is provided below. In the pseudocode, GT_V and GT_A represent the ground truth valence and arousal pairs taken from valence-arousal circumplex model shown in [11], and MSE stands for Mean-Square Error. Other symbols in the pseudocode are as defined in the previous sub-sections of Section III.

```
def DiffCalculator ( $V_W, A_W, [GT_V, GT_A]$ )
    [ $d_E$ ] = MSE ( $V_W, A_W, [GT_V, GT_A]$ )
    return  $d_E$ 

def DecisionTree ( $[P_F], [d_E], Th$ )
    if max ( $[P_F]$ ) >  $Th$ 
        emotionstate = argmax ( $[P_F]$ )
    else
        if  $d_E$  (argsecondmax ( $[P_F]$ )) <  $d_E$  (argmax ( $[P_F]$ ))
            emotionstate = argsecondmax ( $[P_F]$ )
        else
            emotionstate = argmax ( $[P_F]$ )
    return emotionstate
```

//Main hybrid fusion-based emotion recognition cycle

```
load  $CNN_V$ , load  $CNN_A$ , load  $CNN_F$ , load [ $GT_V, GT_A$ ]
 $Th$  = some value between [0.7,1]
 $[W_V]$  = first element is between [-1, -0.5], second element is between [0.5,1]→the center value of high and low valence
```

$[W_A]$ = first element is between [-1, -0.5], second element is between [0.5,1] →the center value of high and low arousal

while True

EEG = readEEGData ()→ at time T

sleep for $T_0 \rightarrow T_0$ is chosen as empirically, optimally set to 0.35 sec.

GSR = readGSRData ()

Face = readFaceData ()

EEG_GSR = concatenate (EEG,GSR)

$P_V = CNN_V$ (EEG)

$P_A = CNN_A$ (EEG_GSR)

$[P_F] = CNN_F$ (Face)

$V_W = \text{sum} (P_V * [W_V]), A_W = \text{sum} (P_A * [W_A])$

$d_E = \text{DiffCalculator} (V_W, A_W, [GT_V, GT_A])$

FinalState = DecisionTree ($[P_F], [d_E], Th$)

IV. DATASETS

The multimodal datasets used in the proposed emotion recognition system include LUMED-2 [73] and DEAP [74] multimodal emotional databases. These two datasets are open to public access.

A. OVERVIEW OF DEAP DATASET

As explained in [8], DEAP is a multimodal emotion dataset which includes peripheral physiological signals and the EEG of 32 participants. Frontal face video of 22 participants was also additionally recorded. To elicit emotions in the participants, one-minute long video clips (music) were shown, and the data was recorded in real-time. Each participant watched 40 separate videos. Participants rated the videos in terms of levels of arousal, valence, like or dislike, dominance, and familiarity along a number scale between 1 and 9. A 32-channel EEG device was used. The raw EEG data was down sampled to 128 Hz. The EOG artefacts were removed and a bandpass filter was applied whose cut-off frequencies are 4.0 Hz and 45.0 Hz. The data was segmented into 60-second intervals and a 3 second baseline data was removed. Face videos are also set to 60-second long segments at 50 fps and 720 × 576 resolution.

B. OVERVIEW OF LUMED-2 DATASET

Loughborough University Multimodal Emotion Database-2 (LUMED-2) is a new multimodal emotion dataset that was created by the researchers of Loughborough University, UK, and Hacettepe University, Turkey, by collecting simultaneous multimodal data from 13 participants (6 females and 7 males) by showing audio-visual stimuli [73]. The total duration of all stimuli is 8 minutes and 50 seconds, which consist of short video clips selected from the web to elicit specific emotions. Between each video clip, in order to let participants, have a rest, a 20-second grey screen was showed.



FIGURE 5. LUMED-2 data collection setup.

Although it is anticipated that each video clip elicits a distinct emotional state and thus can determine the label of the resulting emotion, in reality the same content might trigger differing emotions for different participants. Therefore, after each session, the participants were additionally asked to label the clips with the felt emotional state while watching them. Three different emotions were resulted from labelling: sad, neutral and happy. The facial expressions of the participants were captured using a webcam at a resolution of 640×480 and at 30 fps.

Participants' EEG data was captured using an ENOBIO 8-channels wireless EEG device, which has a temporal resolution of 500 Hz [70]. We filtered EEG data for the frequency range $[0, 75\text{Hz}]$ and applied baseline subtraction for each window. As for the peripheral physiological data, an EMPATICA E4 Wristband [95], powered by Bluetooth, was used to record participants' GSR. A screenshot of the data capturing system is shown in Figure 5. We prepared a fully wireless setup, which facilitates a more comfortable experience for the participants while watching the stimuli, reducing the inherent distress induced by wired devices surrounding their body. This is one of the important advantages of our multimodal data collection setup.

V. RESULTS AND DISCUSSION

In this work, leave-one-subject-out cross-validated classification tests were conducted on two different multimodal emotion datasets: DEAP and LUMED-2. For DEAP dataset, we first generated the discrete emotion labels using the participants' rating of valence and arousal (already existing in the dataset). We extracted the central valence and arousal values of seven discrete emotion states from the 2-D emotional circumplex model shown in Figure 1. Then the Least Mean Squares (LMS) distance between each emotional state and the participant ratings were calculated. The emotion state, which gives the minimum distance and has a maximum normalised distance of 0.2, was assigned as the discrete emotional label

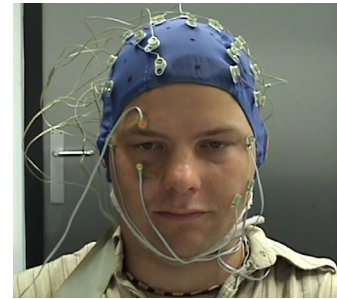


FIGURE 6. An example face imagery of emotion "Angry" from DEAP dataset.

for the corresponding samples in the dataset. For LUMED-2, the discrete emotion labels are given as "sad", "neutral" and "happy", i.e., three classes of emotions.

We classified the emotions under three test conditions: using (a) face modality only, (b) physiological modalities only, and (c) face and physiological modalities together. For DEAP dataset's physiological (EEG+GSR) data, we have achieved an average cross-subject high-low valence classification accuracy of 86.6% and an average high-low arousal classification accuracy of 84.7%. In these two-class classifications, we defined the interval between 7 and 9 on the valence/arousal scale as "High", and the interval between 1 and 3 as "Low". Note that these are not the classification accuracies obtained for discrete emotion states.

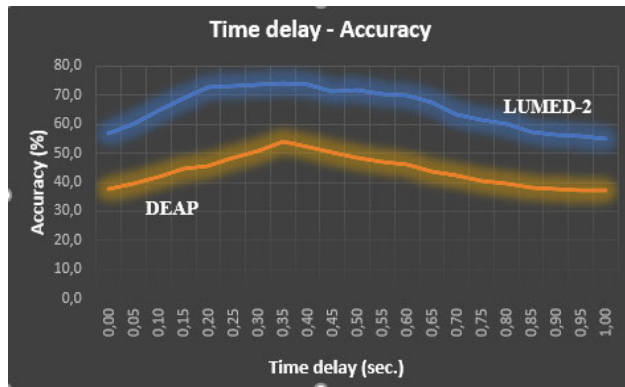
Table 5 shows the classification accuracy results for the three test conditions using the DEAP dataset and for 10 randomly selected participants who had their face video also taken. The number of unique emotion states are also given in the table. It can be seen from the table that using EEG and GSR modalities with face modality increases the recognition accuracy in comparison to using only the face modality and/or only the physiological modalities, for all selected participants. The mean accuracy is increased by approximately 10% compared to using only the face modality.

The proposed hybrid multimodal emotion recognition method also increases the accuracy compared to using physiological modalities only. It increases the mean accuracy by approximately 13%. An important point to note is that the average emotion recognition accuracies are relatively low for the DEAP dataset. This is mainly because the participants' faces are covered by multiple Electromyogram (EMG) electrodes and connectors. EMG setup acts as a hindrance that prevents participants from showing facial expressions freely. This is regarded as deceptive facial actions in terms of our study, which may lead to a higher percentage of false classifications for the face modality.

An example face imagery from the DEAP dataset is shown in Figure 6. A face-based emotion classifier yields for this specific face image a 93% probability of "happy" state and a 4% probability of "neutral" state. On the other hand, the generated emotion label based on the valence and arousal values is "Angry". It is also difficult to recognise the actual emotion state with the proposed multimodal method when

TABLE 5. One-subject-out” prediction accuracies (%) For DEAP dataset.

User	Face	Physiological (GSR+EEG)	Face + (GSR+EEG)	#Emotion states
User 1	0.03	4.51	11.45	5
User 2	31.35	24.72	47.12	4
User 3	85.50	56.89	91.51	3
User 4	29.34	21.33	39.56	4
User 5	46.31	36.42	57.88	5
User 6	45.39	43.01	58.74	4
User 7	25.04	27.49	44.78	4
User 8	25.44	28.30	40.66	5
User 9	11.88	16.48	33.91	3
User 10	18.01	20.21	37.33	4
Average (22 users)	37.11	32.45	53.87	-
Std.Dev. (22 users)	20.35	15.56	16.44	-

**FIGURE 7.** Emotion recognition accuracy with respect to the Time delay between the considered modalities.

the dominant emotion probability is over a high threshold. We have set the threshold probability as 0.9. When the participant starts to show the actual emotion naturally, meaning that the maximum value of probability vector produced by face modality drops under 0.9 and even if it is in the second place in the face recognition probability order, our proposed method mostly catches the actual emotion by use of EEG and GSR. These results can be observed from Table 5 and 6.

Table 6 shows the accuracy results for the LUMED-2 dataset for three discrete emotion states: sad, neutral and happy. As shown in the table, face and physiological modalities establishes superiority to each other for different users, however using all modalities together with proposed method improves the accuracy for all users. The average accuracy is 52% better than using face only case and 42.3% better than using physiological modality. Standard deviation is reduced by 75% and 50% compared to face only and physiological only cases respectively.

The inter time-delay has been chosen empirically. It is one of the core parameters that impacts the recognition accuracy. We employ different time delay values from 0 to 1 sec. and

TABLE 6. One-subject-out” prediction accuracies (%) For LUMED-2 dataset.

User	Face	Physiological (GSR+EEG)	Face + (GSR+EEG)
User 1	70.2	63.0	74.2
User 2	73.3	58.4	79.3
User 3	72.9	70.5	81.4
User 4	77.9	61.1	83.6
User 5	34.2	56.2	71.3
User 6	43.1	40.1	73.3
User 7	28.6	45.7	66.0
User 8	38.3	53.3	71.9
User 9	45.5	54.2	75.5
User 10	39.3	36.6	70.0
User 11	6.0	39.0	67.4
User 12	40.5	52.2	73.1
User 13	44.3	49.1	78.2
Average	47.2	52.2	74.2
Std. Dev.	20.8	10.0	5.2

observe the changing in the accuracies. Figure 7 shows the accuracy change for the DEAP and LUMED-2 datasets as the introduced time delay changes between 0 to 1s. The optimal value for both datasets is around 0.35. So, we set the time delay as 0.35 sec. Figure 7 shows how the value of time delay affects the mean recognition accuracy. This proves that the bio-signals do not react simultaneously and there is a time lag between brain response and other bio-signals. This is due to the time it takes to transmit biochemicals to other parts of the body, after being initiated inside the brain.

The proposed method has also been compared to other cross-subject unimodal and multimodal emotion recognition studies in the literature. There are not many multimodal emotion recognition studies that do classification on discrete emotion states basis. Most of the multimodal emotion recognition studies implement emotion recognition on the scales of valence and arousal.

Therefore, to be able to make a comparison, we will regard the high/low valence and high/low arousal classification methods as they implement 2 emotion states classification. Table 7 shows the accuracy comparison between the proposed method and 15 other multimodal and unimodal emotion recognition studies. The number of output emotion classes is an important aspect, which determines the mean classification accuracy. Note that the left-hand side of the ‘/’ symbol stands for the high-low arousal accuracy and the right-hand side of it stands for the high-low valence accuracy. If there is no ‘/’ symbol, then the comparison is done based on either valence or arousal with maximum accuracy.

Proposed method provides the accuracy results based on discrete emotional states. The proposed method performs the best in terms of 2-states, although it implements a 4-state classification. For a fair comparison, we have calculated the mean

TABLE 7. Benchmark for “One-subject-out” Mean prediction accuracies (%) on the DEAP Dataset.

Study	Modality	Accuracy	Number of emotion states
<i>Proposed</i>	Multimodal	53.8	4
<i>Proposed</i>	Multimodal	79.2	2 or 3
<i>Proposed</i>	Multimodal	82.7	2
<i>Study-1 [67]</i>	Multimodal	43.5 / 77.5	2 / 2
<i>Study-2 [66]</i>	Multimodal	72.1 / 73.0	2 / 2
<i>FAWT [96]</i>	EEG	79.9	2
<i>T-RFE [97]</i>	EEG	78.7	2
<i>Inc.ResnetV2[94]</i>	EEG	72.8	2
<i>ST-SBSSVM [98]</i>	EEG	72.0	2
<i>VMD-DNN [99]</i>	EEG	62.5	2
<i>MIDA [100]</i>	EEG	48.9	2
<i>TCA [101]</i>	EEG	47.2	2
<i>SA [102]</i>	EEG	38.7	2
<i>ITL [103]</i>	EEG	40.5	2
<i>GFK [104]</i>	EEG	46.5	2
<i>KPCA [105]</i>	EEG	39.8	2
<i>Study-3 [106]</i>	Face	71.1/72.3	2
<i>Study-3 [106]</i>	EEG	68.8/75.3	2
<i>Study-3 [106]</i>	Multimodal	74.2/80.3	2
<i>Study-4 [107]</i>	EEG	68.4/78.9	2

accuracy of the classification results of users who have 2 and 3 emotion states. We put these results in Table 7 too. Based on that, when the number of states is equal to 2, proposed study yields better mean accuracy compared to others on the DEAP dataset.

We could find only one study which reports the accuracy based on face modality, on DEAP. However, this study updates the recognition model parameters according to dataset and, they use some of the users' face data to train their model. However, we report the face-based subject independent accuracy without using any users' data from DEAP dataset. Therefore, we foresee that if we use some users' data to train our model in terms of face modality, the face-based recognition accuracy would improve.

Since Lumed-2 is a new multimodal dataset, we have been unable to reproduce the respective methods in other studies. It needs to be emphasized that the accuracy figures we report here are one-subject-out figures (subject-independent results). This should not be confused with the subject-dependent accuracy results, which are reported in most studies in the literature and tend to be much higher than one-subject-out results.

VI. CONCLUSION

The objective of this paper is to classify emotional states by using a multimodal approach. We use a hybrid fusion of face, EEG and GSR modalities. We use feature fusion on EEG and

GSR modalities for estimating the level of arousal. In the final step we use late fusion of EEG, GSR and face modalities. Our proposed model has the capability of detecting the actual emotional state when it is dominant, or it is hidden due to natural deceptive face actions.

We present a subject independent emotion recognition system that is suitable for real time operations, since it will not require feature extraction. Hence, this brings flexibility and reduces the overall processing load. Although subject-independent recognition accuracy of EEG modality is relatively low due to its nonstationary properties, it is effective if the number of output classes is limited. Therefore, we limited the number of classes to two (high and low states for valence and arousal) to make use of its success when detecting the actual emotional state.

Future studies should focus on subject-independent emotion recognition for person-independent applicability and concentrate on reducing the number of modalities for practicality. Reported user-based emotion recognition accuracies are higher compared to cross-subject and subject-independent emotion recognition, but these are less practical. In addition, since the comfort of use and non-intrusiveness is important for end users, significant effort should be made towards reducing the weight and form factor of sensor devices.

ACKNOWLEDGMENT

The authors would like to thank the creators of DEAP dataset for openly sharing the dataset and the wider research community. The authors would also like to thank all volunteering staff and students with Loughborough University London and Hacettepe University for participating the recording sessions to generate the LUMED-2 dataset and Perihan TEKELI, Hacettepe University, for the excellent coordination activities.

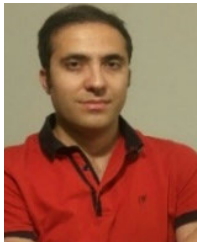
REFERENCES

- [1] A. Dąbrowski, “Emotions in philosophy. A short introduction,” *Studia Humana*, vol. 5, no. 3, pp. 8–20, 2016.
- [2] R. W. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 1995.
- [3] H. Ranganathan, S. Chakraborty, and S. Panchanathan, “Multimodal emotion recognition using deep learning architectures,” in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–9.
- [4] W.-L. Zheng, W. Liu, Y. Lu, B.-L. Lu, and A. Cichocki, “EmotionMeter: A multimodal framework for recognizing human emotions,” *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 1110–1122, Mar. 2019.
- [5] K. Bahreini, R. Nadolski, and W. Westera, “Data fusion for real-time multimodal emotion recognition through webcams and microphones in E-Learning,” *Int. J. Hum.-Comput. Interact.*, vol. 32, no. 5, pp. 415–430, May 2016.
- [6] B. Bucur, I. Șomfelean, A. Ghiurțan, C. Lemnaru, and M. Dinșoreanu, “An early fusion approach for multimodal emotion recognition using deep recurrent networks,” in *Proc. IEEE 14th Int. Conf. Intell. Comput. Commun. Process. (ICCP)*, Sep. 2018, pp. 71–78.
- [7] Z. Wang, S.-B. Ho, and E. Cambria, “A review of emotion sensing: Categorization models and algorithms,” *Multimedia Tools Appl.*, pp. 1–30, Jan. 2020, doi: 10.1007/s11042-019-08328-z.
- [8] S. Koelstra, C. Mueh, M. Soleymani, J. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, “DEAP: A database for emotion analysis using physiological signals,” *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan./Mar. 2012.

- [9] J. A. Russell, "A circumplex model of affect," *J. Personality Social Psychol.*, vol. 39, no. 6, pp. 1161–1178, 1980.
- [10] P. Ekman, W. V. Friesen, M. O'Sullivan, A. I. Chan, T. Diacyanni, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, and P. E. Ricci-Bitti, "Universals and cultural differences in the judgments of facial expressions of emotion," *J. Personality Social Psychol.*, vol. 53, no. 4, pp. 712–717, 1987.
- [11] G. Paltoglou and M. Thelwall, "Seeing stars of valence and arousal in blog posts," *IEEE Trans. Affect. Comput.*, vol. 4, no. 1, pp. 116–123, Jan. 2013.
- [12] R. Fabien, S. Björn, V. Michel, J. Shashank, M. Erik, L. Denis, C. Roddy, and P. Maja, "AV+EC 2015: The first affect recognition challenge bridging across audio, video, and physiological data," in *Proc. 5th Int. Workshop Audio/Vis. Emotion Challenge*, 2015, pp. 3–8.
- [13] A. Kapoor, W. Burleson, and R. W. Picard, "Automatic prediction of frustration," *Int. J. Hum.-Comput. Stud.*, vol. 65, no. 8, pp. 724–736, Aug. 2007.
- [14] H. Zhang, A. Jolfaei, and M. Alazab, "A face emotion recognition method using convolutional neural network and image edge computing," *IEEE Access*, vol. 7, pp. 159081–159089, 2019.
- [15] J. Deng, S. Fröhholz, Z. Zhang, and B. Schuller, "Recognizing emotions from whispered speech based on acoustic feature transfer learning," *IEEE Access*, vol. 5, pp. 5235–5246, 2017.
- [16] C. Qing, R. Qiao, X. Xu, and Y. Cheng, "Interpretable emotion recognition using EEG signals," *IEEE Access*, vol. 7, pp. 94160–94170, 2019.
- [17] H. A. Osman and T. H. Falk, *Multimodal Affect Recognition: Current Approaches and Challenges*. Rijeka, Croatia: IntechOpen, 2016.
- [18] T. Hao, L. Wei, Z. Wei-Long, L. Bao-Liang, L. Derong, X. Shengli, L. Yuanqing, Z. Dongbin, and E. El-Sayed, "Multimodal emotion recognition using deep neural networks," in *Proc. Int. Conf. Neural Inf. Process.*, 2017, pp. 811–819.
- [19] J. Chen, Y. She, M. Zheng, Y. Shu, Y. Wang, and Y. Xu, "A multimodal affective computing approach for children companion robots," in *Proc. 7th Int. Symp. Chin. CHI*, 2019, pp. 57–64.
- [20] D. Hazer-Rau, S. Meudt, A. Daucher, J. Spohrs, H. Hoffmann, F. Schwenker, and H. Traue, "The uulmMAC database—A multimodal affective corpus for affective computing in human-computer interaction," *Sensors*, vol. 20, no. 8, p. 2308, 2020.
- [21] K. Masui, T. Nagasawa, H. Doi, N. Tsumura, "Continuous estimation of emotional change using multimodal affective responses," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jun. 2020, pp. 290–291.
- [22] B. Singh, N. Kushwaha, and O. P. Vyas, "An interpretation of sentiment analysis for enrichment of Business Intelligence," in *Proc. IEEE Region 10 Conf. (TENCON)*, Nov. 2016, pp. 18–23.
- [23] J. Islam and Y. Zhang, "Visual sentiment analysis for social images using transfer learning approach," in *Proc. IEEE Int. Conference Big Data Cloud Comput. (BDCloud), Social Comput. Netw. (SocialCom), Sustain. Comput. Commun. (SustainCom) (BDCloud-SocialCom-SustainCom)*, Oct. 2016, pp. 124–130.
- [24] P. Yadav and D. Pandya, "SentiReview: Sentiment analysis based on text and emoticons," in *Proc. Int. Conf. Innov. Mech. Ind. Appl. (ICIMIA)*, Feb. 2017, pp. 467–472.
- [25] N. Xu and W. Mao, "A residual merged neutral network for multimodal sentiment analysis," in *Proc. IEEE 2nd Int. Conf. Big Data Anal. (ICBDA)*, Mar. 2017, pp. 6–10, doi: [10.1109/icbda.2017.8078794](https://doi.org/10.1109/icbda.2017.8078794).
- [26] S. A. Qureshi, S. Saha, M. Hasanuzzaman, G. Dias, and E. Cambria, "Multitask representation learning for multimodal estimation of depression level," *IEEE Intell. Syst.*, vol. 34, no. 5, pp. 45–52, Sep. 2019.
- [27] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh, and A. Hussain, "Multimodal sentiment analysis: Addressing key issues and setting up the baselines," *IEEE Intell. Syst.*, vol. 33, no. 6, pp. 17–25, Nov/Dec. 2018.
- [28] A. Bandhakavi, N. Wiratunga, S. Massie, and D. Padmanabhan, "Lexicon generation for emotion detection from text," *IEEE Intell. Syst.*, vol. 32, no. 1, pp. 102–108, Jan./Feb. 2017.
- [29] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intell. Syst.*, vol. 31, no. 6, pp. 82–88, Nov. 2016.
- [30] E. Cambria, "Affective computing and sentiment analysis," *IEEE Intell. Syst.*, vol. 31, no. 2, pp. 102–107, Mar. 2016.
- [31] A. Esuli, A. Moreo, F. Sebastiani, and E. Cambria, "Cross-lingual sentiment quantification," *IEEE Intell. Syst.*, vol. 35, no. 3, pp. 106–114, May 2020.
- [32] J. Schuurmans, F. Frasincar, and E. Cambria, "Intent classification for dialogue utterances," *IEEE Intell. Syst.*, vol. 35, no. 1, pp. 82–88, Jan. 2020.
- [33] N. Majumder, S. Poria, H. Peng, N. Chhaya, E. Cambria, A. Gelbukh, and E. Cambria, "Sentiment and sarcasm classification with multitask learning," *IEEE Intell. Syst.*, vol. 34, no. 3, pp. 38–43, May 2019.
- [34] Q. Yang, Y. Rao, H. Xie, J. Wang, F. L. Wang, W. H. Chan, and E. Cambria, "Segment-level joint topic-sentiment model for online review analysis," *IEEE Intell. Syst.*, vol. 34, no. 1, pp. 43–50, Jan. 2019.
- [35] M. Dragoni, S. Poria, and E. Cambria, "OntoSentNet: A common-sense ontology for sentiment analysis," *IEEE Intell. Syst.*, vol. 33, no. 3, pp. 77–85, May 2018.
- [36] E. Cambria, S. Poria, A. Gelbukh, and M. Thelwall, "Sentiment analysis is a big suitcase," *IEEE Intell. Syst.*, vol. 32, no. 6, pp. 74–80, Nov. 2017.
- [37] M. Ebrahimi, A. H. Yazdavar, and A. Sheth, "Challenges of sentiment analysis for dynamic events," *IEEE Intell. Syst.*, vol. 32, no. 5, pp. 70–75, Sep. 2017.
- [38] A. Weichselbraun, S. Gindl, F. Fischer, S. Vakulenko, and A. Scharl, "Aspect-based extraction and analysis of affective knowledge from social media streams," *IEEE Intell. Syst.*, vol. 32, no. 3, pp. 80–88, May 2017.
- [39] N. Majumder, "Deep learning-based document modeling for personal-ty detection from text," *IEEE Intell. Syst.*, vol. 32, no. 2, pp. 74–79, Mar./Apr. 2017.
- [40] E. Cambria, N. Howard, J. Hsu, and A. Hussain, "Sentic blending: Scalable multimodal fusion for the continuous interpretation of semantics and sentics," in *Proc. IEEE Symp. Comput. Intell. Hum.-Like Intell. (CIHLI)*, Singapore, Apr. 2013, pp. 108–117, doi: [10.1109/CIHLI.2013.6613272](https://doi.org/10.1109/CIHLI.2013.6613272).
- [41] S. Poria, E. Cambria, N. Howard, G.-B. Huang, and A. Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," *Neurocomputing*, vol. 174, pp. 50–59, Jan. 2016.
- [42] B. Xing, H. Zhang, K. Zhang, L. Zhang, X. Wu, and X. Shi, "Exploiting EEG signals and audiovisual feature fusion for video emotion recognition," *IEEE Access*, vol. 7, pp. 59844–59861, 2019.
- [43] J. Han, X. Ji, X. Hu, L. Guo, and T. Liu, "Arousal recognition using audio-visual features and fMRI-based brain response," *IEEE Trans. Affect. Comput.*, vol. 6, no. 4, pp. 337–347, Oct. 2015.
- [44] S. Sahoo and A. Routray, "Emotion recognition from audio-visual data using rule based decision level fusion," in *Proc. IEEE Students' Technol. Symp. (TechSym)*, Sep. 2016, pp. 7–12.
- [45] K.-S. Song, Y.-H. Nho, J.-H. Seo, and D.-S. Kwon, "Decision-level fusion method for emotion recognition using multimodal emotion recognition information," in *Proc. 15th Int. Conf. Ubiquitous Robots (UR)*, Jun. 2018, pp. 472–476.
- [46] A. Metallinou, S. Lee, and S. Narayanan, "Decision level combination of multiple modalities for detection and analysis of emotional expression," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2010, pp. 2462–2465.
- [47] M. Mansoorzadeh and N. M. Charkari, "Hybrid feature and decision level fusion of face and speech information for bimodal emotion recognition," in *Proc. 14th Int. CSI Comput. Conf.*, Oct. 2009, pp. 652–657.
- [48] S. Nemati, R. Rohani, M. E. Basiri, M. Abdar, N. Y. Yen, and V. Makarenkov, "A hybrid latent space data fusion method for multimodal emotion recognition," *IEEE Access*, vol. 7, pp. 172948–172964, 2019.
- [49] J. Chen, B. Hu, L. Xu, P. Moore, and Y. Su, "Feature-level fusion of multimodal physiological signals for emotion recognition," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2015, pp. 395–399.
- [50] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-End multimodal emotion recognition using deep neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 8, pp. 1301–1309, Dec. 2017.
- [51] *SEED Dataset*. Accessed: Apr. 19, 2020. [Online]. Available: <http://bcmi.sjtu.edu.cn/~seed/>
- [52] Y. Huang, J. Yang, P. Liao, and J. Pan, "Fusion of facial expressions and EEG for multimodal emotion recognition," *Comput. Intell. Neurosci.*, vol. 2017, pp. 1–8, 2017.
- [53] B. Reuderink, M. Poel, K. Truong, R. Poppe, and M. Pantic, "Decision-level fusion for audio-visual laughter detection," in *Machine Learning for Multimodal Interaction (MLMI)* (Lecture Notes in Computer Science), vol. 5237, 2008, pp. 137–148.
- [54] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE'05 audio-visual emotion database," in *Proc. 22nd Int. Conf. Data Eng. Workshops (ICDEW)*, Atlanta, GA, USA, Apr. 2006, p. 8, doi: [10.1109/ICDEW.2006.145](https://doi.org/10.1109/ICDEW.2006.145).

- [55] C. Y. Suen and L. Lam, "Multiple classifier combination methodologies for different output levels," in *Multiple Classifier Systems. MCS* (Lecture Notes in Computer Science), vol. 1857. Berlin, Germany: Springer, 2000, doi: 10.1007/3-540-45014-9_5.
- [56] F. Lingenfelder, J. Wagner, and E. André, "A systematic discussion of fusion techniques for multi-modal affect recognition tasks," in *Proc. 13th Int. Conf. Multimodal Interfaces (ICMI)*, 2011, pp. 19–26.
- [57] Y. S. Huang and C. Y. Suen, "The behavior-knowledge space method for combination of multiple classifiers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 1993, p. 347.
- [58] R. Gupta, M. Khomami Abadi, J. A. Cárdenes Cabré, F. Morreale, T. H. Falk, and N. Sebs, "A quality adaptive multimodal affect recognition system for user-centric multimedia indexing," in *Proc. ACM Int. Conf. Multimedia Retr. (ICMR)*, 2016, pp. 317–320.
- [59] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, Jun. 2001.
- [60] H. Miao, Y. Zhang, W. Li, H. Zhang, D. Wang, and S. Feng, "Chinese multimodal emotion recognition in deep and traditional machine learning approaches," in *Proc. 1st Asian Conf. Affect. Comput. Intell. Interact. (ACII Asia)*, May 2018, pp. 1–6.
- [61] J. Kim, E. André, M. Rehm, T. Vogt, and J. Wagner, "Integrating information from speech and physiological signals to achieve emotional sensitivity" in *Proc. Interspeech*, 2005, pp. 809–812.
- [62] A. Sharma and S. Canavan, "Multimodal physiological-based emotion recognition," Univ. South Florida, Tampa, FL, USA, Tech. Rep., 2019. [Online]. Available: <https://www.semanticscholar.org/paper/Multimodal-Physiological-based-Emotion-Recognition-Sharma-Canavan/4ab5469bf1f0690956aeb1271831c23a46b2bbd4>
- [63] Z. Zhang, J. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, J. Cohn, Q. Ji, and L. Yin, "Multimodal spontaneous emotion corpus for human behavior analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3438–3446.
- [64] F. Syeda and E. Engin, "Cross-subject continuous emotion recognition using speech and body motion in dyadic interactions," in *Proc. Interspeech*, 2017, pp. 1731–1735.
- [65] A. Metallinou, Z. Yang, C.-C. Lee, C. Busso, S. Carnicke, and S. Narayanan, "The USC CreativeIT database of multimodal dyadic interactions: From speech and full body motion capture to continuous emotional annotations," *Lang. Resour. Eval.*, vol. 50, no. 3, pp. 497–521, Sep. 2016.
- [66] D. Ayata, Y. Yaslan, and M. E. Kamasak, "Emotion recognition from multimodal physiological signals for emotion aware healthcare systems," *J. Med. Biol. Eng.*, vol. 40, no. 2, pp. 149–157, Apr. 2020.
- [67] J. Chen, B. Hu, Y. Wang, P. Moore, Y. Dai, L. Feng, and Z. Ding, "Subject-independent emotion recognition based on physiological signals: A three-stage decision method," *BMC Med. Informat. Decis. Making*, vol. 17, no. S3, Dec. 2017.
- [68] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. AAAI*, vol. 4, 2017, p. 1.
- [69] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (Workshops)*, Jun. 2010, pp. 94–101.
- [70] O. Langner, R. Dotsch, G. Bijlstra, D. H. J. Wigboldus, S. T. Hawk, and A. van Knippenberg, "Presentation and validation of the radboud faces database," *Cognition Emotion*, vol. 24, no. 8, pp. 1377–1388, Dec. 2010.
- [71] *FacesDB*. Accessed: Apr. 10, 2020. [Online]. Available: <http://app.visgraf.impa.br/database/faces/>
- [72] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, Jan./Mar. 2019.
- [73] *LUMED-2 Dataset*. Accessed: Jul. 3, 2020. [Online]. Available: https://figshare.com/articles/dataset/Loughborough_University_Multimodal_Emotion_Dataset_-_2/12644033
- [74] *DEAP Dataset*. Accessed: Apr. 21, 2020. [Online]. Available: <https://www.eecs.qmul.ac.uk/mnm/datasets/deap/>
- [75] *Pretrained Deep Neural Networks*. Accessed: Mar. 1, 2020. [Online]. Available: <https://uk.mathworks.com/help/deeplearning/ug/pretrained-convolutional-neural-networks.html>
- [76] K. A. Lindquist, T. D. Wager, H. Kober, E. Bliss-Moreau, and L. F. Barrett, "The brain basis of emotion: A meta-analytic review," *Behav. Brain Sci.*, vol. 35, no. 3, pp. 121–143, 2012.
- [77] M. Benedek and C. Kaernbach, "A continuous measure of phasic electrodermal activity," *J. Neurosci. Methods*, vol. 190, no. 1, pp. 80–91, Jun. 2010.
- [78] *Imotions*. Accessed: Jun. 27, 2020. [Online]. Available: <https://imotions.com/blog/gsr-why-5-application-trends-biometric-research/>
- [79] *Top 14 EEG Hardware Companies*. Accessed: Apr. 5, 2020. [Online]. Available: <https://imotions.com/blog/top-14-eeeg-hardware-companies-ranked/>
- [80] V. Jurcak, D. Tsuzuki, and I. Dan, "10/20, 10/10, and 10/5 systems revisited: Their validity as relative head-surface-based positioning systems," *NeuroImage*, vol. 34, no. 4, pp. 1600–1611, Feb. 2007.
- [81] U. R. Acharya, S. V. Sree, G. Swapna, R. J. Martis, and J. S. Suri, "Automated EEG analysis of epilepsy: A review," *Knowl.-Based Syst.*, vol. 45, pp. 147–165, Jun. 2013.
- [82] K. Aboalayon, M. Faezipour, W. Almuhammadi, and S. Moslehpour, "Sleep stage classification using EEG signal analysis: A comprehensive survey and new investigation," *Entropy*, vol. 18, no. 9, p. 272, Aug. 2016.
- [83] D. A. Engemann, "Robust EEG-based cross-site and cross-protocol classification of states of consciousness," *Brain*, vol. 141, no. 11, pp. 3179–3192, 2018.
- [84] M. Arns, C. K. Conners, and H. C. Kraemer, "A decade of EEG theta/beta ratio research in ADHD: A meta-analysis," *Journal of Attention Disorders*, vol. 17, no. 5, pp. 374–383, 2013.
- [85] W. K. Y. So, S. W. H. Wong, J. N. Mak, and R. H. M. Chan, "An evaluation of mental workload with frontal EEG," *PLoS ONE*, vol. 12, no. 4, Apr. 2017, Art. no. e0174949.
- [86] N.-H. Liu, C.-Y. Chiang, and H.-C. Chu, "Recognizing the degree of human attention using EEG signals from mobile sensors," *Sensors*, vol. 13, no. 8, pp. 10273–10286, Aug. 2013.
- [87] A. Y. Shestiyuk, K. Kasinathan, V. Karapoonindott, R. T. Knight, and R. Gurumoorthy, "Individual EEG measures of attention, memory, and motivation predict population level TV viewership and Twitter engagement," *PLoS ONE*, vol. 14, no. 3, Mar. 2019, Art. no. e0214507.
- [88] M. Mohammadpour and S. Mozaffari, "Classification of EEG-based attention for brain computer interface," in *Proc. 3rd Iranian Conf. Intell. Syst. Signal Process. (ICSPIS)*, Dec. 2017, pp. 34–37.
- [89] S. Thejaswini, K. M. Ravikumar, L. Jhenkar, N. Aditya, and K. K. Abhay, "Analysis of EEG based emotion detection of DEAP and SEED-IV databases using SVM," *Int. J. Recent Technol. Eng. (IJRTE)*, vol. 8, pp. 207–211, May 2019.
- [90] J. Liu, H. Meng, A. Nandi, and M. Li, "Emotion detection from EEG recordings," in *Proc. 12th Int. Conf. Natural Comput., Fuzzy Syst. Knowl. Discovery (ICNC-FSKD)*, Aug. 2016, pp. 1722–1727.
- [91] A. Gómez, L. Quintero, N. López, and J. Castro, "An approach to emotion recognition in single-channel EEG signals: A mother child interaction," in *Proc. J. Phys., Conf.*, vol. 705, 2016, Art. no. 012051.
- [92] W. Zhang, F. Wang, Y. Jiang, Z. Xu, S. Wu, and Y. Zhang, "Cross-subject EEG-based emotion recognition with deep domain confusion," in *Intelligent Robotics and Applications (ICIRA)* (Lecture Notes in Computer Science), vol. 11740. Cham, Switzerland: Springer, 2019, pp. 558–570.
- [93] Z. Yin, Y. Wang, L. Liu, W. Zhang, and J. Zhang, "Cross-subject EEG feature selection for emotion recognition using transfer recursive feature elimination," *Frontiers Neurobot.*, vol. 11, p. 19, Apr. 2017, doi: 10.3389/fnbot.2017.00019.
- [94] Y. Cimtay and E. Ekmekcioglu, "Investigating the use of pretrained convolutional neural network on cross-subject and cross-dataset EEG emotion recognition," *Sensors*, vol. 20, no. 7, p. 2034, Apr. 2020.
- [95] *Empatica E4*. Accessed: Apr. 9, 2020. [Online]. Available: <https://www.empatica.com/en-int/research/e4/>
- [96] V. Gupta, M. D. Chopda, and R. B. Pachori, "Cross-subject emotion recognition using flexible analytic wavelet transform from EEG signals," *IEEE Sensors J.*, vol. 19, no. 6, pp. 2266–2274, Mar. 2019.
- [97] Z. Yin, Y. Wang, L. Liu, W. Zhang, and J. Zhang, "Cross-subject EEG feature selection for emotion recognition using transfer recursive feature elimination," *Frontiers Neurobot.*, vol. 11, p. 200, Apr. 2017.
- [98] F. Yang, X. Zhao, W. Jiang, P. Gao, and G. Liu, "Multi-method fusion of cross-subject emotion recognition based on high-dimensional EEG features," *Frontiers Comput. Neurosci.*, vol. 13, p. 53, Aug. 2019.
- [99] P. Pandey and K. R. Seeja, "Subject independent emotion recognition from EEG using VMD and deep learning," *J. King Saud Univ.-Comput. Inf. Sci.*, pp. 53–58, Nov. 2019. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S131957819309991>

- [100] K. Yan, L. Kou, and D. Zhang, "Learning domain-invariant subspace using domain features and independence maximization," *IEEE Trans. Cybern.*, vol. 48, no. 1, pp. 288–299, Jan. 2018.
- [101] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011.
- [102] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2960–2967.
- [103] Y. Shi and F. Sha, "Information-theoretical learning of discriminative clusters for unsupervised domain adaptation," in *Proc. 2012 Int. Conf. Mach. Learn. (ICML)*, Edinburgh, U.K., Jun./Jul. 2012, pp. 1275–1282.
- [104] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 2066–2073.
- [105] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, Jul. 1998.
- [106] Y. Huang, J. Yang, S. Liu, and J. Pan, "Combining facial expressions and electroencephalography to enhance emotion recognition," *Future Internet*, vol. 11, no. 5, p. 105, May 2019.
- [107] V. Rozgic, S. N. Vitaladevuni, and R. Prasad, "Robust EEG emotion classification using segment level decision fusion," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Vancouver, BC, Canada, May 2013, pp. 1286–1290.



YUCEL CIMTAY received the Ph.D. degree in electrical and electronics engineering from Ankara University, Turkey. He is currently a Postdoctoral Research Associate with the Institute for Digital Technologies, Loughborough University London. His research interests include signal and image processing, data science, and machine learning.



ERHAN EKMEÇCIOĞLU received the Ph.D. degree from the University of Surrey, U.K., in 2010. He was a Postdoctoral Researcher in 2014. Since 2014, he has been with the Institute for Digital Technologies, Loughborough University London, U.K., where he is currently a Senior Lecturer and the Chief Director with the Postgraduate Taught Program. His current research interests include affective computing, immersive media, multimedia processing, and applied machine learning. He serves as a Guest Editor for the IEEE Multimedia Communications Technical Committee publications and a Regular Reviewer for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, *Journal of Multimedia*, and *Journal on Image Processing*.



SEYMA CAGLAR-OZHAN received the master's degree from Hacettepe University, Turkey, in 2017, where she is currently pursuing the Ph.D. degree in education with the Computer Education and Instructional Technology Department. Her research interests include cognitive and emotional processes in e-learning environments, human-computer interaction, and cognitive science.

...