Egyptian Informatics Journal 22 (2021) 167–176

Contents lists available at ScienceDirect

Egyptian Informatics Journal

journal homepage: www.sciencedirect.com

A 3D-convolutional neural network framework with ensemble learning techniques for multi-modal emotion recognition

Elham S. Salama [a,*], Reda A. El-Khoribi [a], Mahmoud E. Shoman [a], Mohamed A. Wahby Shalaby [a,b]

[a] Faculty of Computers, and Artificial Intelligence, Cairo University, Giza, Egypt
[b] Smart Engineering Systems Research Center (SESC), Nile University, Giza, Egypt

ARTICLE INFO

Article history:
Received 29 December 2019
Revised 19 March 2020
Accepted 24 July 2020
Available online 13 August 2020

Keywords:
Electroencephalogram
Facial expressions
Multi-modal emotion recognition
Deep learning
Transfer learning
Data augmentation
Ensemble learning techniques

ABSTRACT

Nowadays, human emotion recognition is a mandatory task for many human machine interaction fields. This paper proposes a novel multi-modal human emotion recognition framework. The proposed scheme utilizes first the 3D-Convolutional Neural Network (3D-CNN) deep learning architecture for extracting the spatio-temporal features from the electroencephalogram (EEG) signals, and the video data of human faces. Then, a combination of data augmentation, ensemble learning techniques is proposed to get the final fusion predictions. The fusion of the multi-modalities in the proposed scheme is carried out using data, and score fusion methods. Hence, three human recognition approaches are built to achieve the proposed goal. They are namely EEG-based emotion recognition approach, face-based emotion recognition approach, and fusion-based emotion recognition approach. For the EEG approach, the 3D-CNN is used to get the final predictions of the EEG signal. For the face approach, the Mask-RCNN object detection technique in combination with OpenCV libraries are first utilized to extract the exact face pixels with emotional content. Then, the Support Vector Machine (SVM) classifier is utilized to classify the 3D-CNN output features of the face chunks. For the fusion-based emotion recognition approach, two fusion techniques are experimented; bagging, and stacking. It is found that the stacking technique gives the best accuracy, and achieves recognition accuracies 96.13%, and 96.79% for valence, and arousal classes respectively using the grid search ensemble learning technique due to transferring the weights from the EEG, and the face approaches to the fusion-based emotion recognition approach. The proposed approach outperforms recent works in multi-modal emotion recognition field.

1. Introduction

Emotions play an important role in human life. They affect their physiological, and psychological state. Emotions can be used for rating customers' impression for the environment in restaurants as stated by authors in [1]. They presented a rating system based on facial expression recognition with pre-trained deep convolutional neural network (CNN) models. The food, and the environment are supposed to be rated in their system. Their system consists of Android mobile application, a web server, and a pre-trained AI server. Shanok et al. [2] utilized an emotion recognition for children with autism disorders. Their results showed that the TD children were more proficient on the emotion recognition systems overall, whereas ASD children recognized familiar expressions more precisely than unfamiliar ones. Emotion recognition is the ability of someone to identify what other people feel from moment to moment and understand the connection between his feelings and expressions.

Ekman et al. [3] defined six basic emotions namly happiness, sadness, surprise, fear, and anger. He proved that human perceive these emotions regardless of their cultures. Emotions could be expressed using two orthogonal dimensions: valence, and arousal as stated by Feldman et al. [4]. He stated that each individual could express his emotions in different way than others. This difference is clearly noted when someone asked to express periodic emotions. Valence ranges from pleasant to unpleasant, and arousal ranges from calm to excited. In the proposed work, authors intend to classify the input instance into its binary combinations of valence, and arousal; low/high valence or low/high arousal.

Mehrabian et al. [5] stated that the overall impression could be expressed by facial expression with 55%, by vocal part with 38%,

* Corresponding author.
E-mail address: elham.shawky@fci-cu.edu.eg (E.S. Salama).

Peer review under responsibility of Faculty of Computers and Artificial Intelligence, Cairo University.

https://doi.org/10.1016/j.eij.2020.07.005

and by semantic content with 7%. So, the human emotion is multimodal in nature. Emotion recognition works in literature are partitioned into two main areas: Some works use one modality for recognizing emotions such as face regions [6], speech [7], or electroencephalogram (EEG) signals [8]. Other researchers built multi-modal emotion recognition systems using a combination of different emotion modalities. Combining knowledge form different sources to produce more accurate information is called fusion [9]. There are three common fusion methods; data level, feature level, and score level. Data level combines the input raw data from different sources to produce more informative data. Then, one set of features is extracted for the combined data, and one classification system is built to group the extracted features. Feature level creates a combined features from multiple input features extracted from different data sources. Then, one classifier is also used to model the relation between the combined features. In score fusion, a separate classification system distributed the data from each source, and categorize them into its related classes. Then, the scores from each classification system are combined to get the final scores. The proposed work adopted the data, and score fusion methods in one compact framework.

Over decades massive research work have been developed using artificial intelligence (AI), and deep learning techniques to solve efficiently different complex problems [10–14]. Recently, there are several human emotion recognition schemes have been developed based on AI, and deep learning techniques [8,15,16]. It is seen from the literature that these recent techniques can improve performance dramatically. One of the most common deep learning approaches is the 3D-CNNs architicture which is mainly proposed to model the temporal correlation in long sequences such as speech, and EEG signals as stated by Maturana et al. [17]. Authors in [8] proved the superiority of 3D-CNN architicture in recognizing emotions from multi-channel EEG signals. They employed the DEAP dataset to achieve 87.44% accuracy for valence, and 88.49% for arousal. Long sequences have temporal relationships between its segments, and neglecting these temporal information will affect the robustness of emotion recognition systems. The 3D-CNN architicture models the temporal dependencies in long sequences by applying 3D-convolution operations over the input segments. The 3D-CNN architecture is adopted for the proposed goal of this work.

The main drawback of any recognition system is the lack of data which may affect its generalization to unseen samples. This can be solved by augmenting the data (increasing the number of samples), or using a model that is already trained on a task, and re-use it on another related task (transfer learning). It is well known that data augmentation avoids system over-fitting problem. In addition, transfer learning has several benefits such as saving processing time, requiring less data and enhancing robustness of developed system. Therefore, in the proposed multi-modal emotion recognition framework, transfer learning and data augmentation phases are employed in one compact fusion system to tackle the problem of lacking data efficiently.

This paper is organized as follows: In Section 2, previous related works are discussed. In Section 3, the main goal of the proposed approach, and the three main components for face, EEG, and fusion recognition approaches are discussed. Evaluation of the proposed approach on DEAP benchmark, comparison against state-of-the-art approaches, discussion, and analysis for the experimental works are presented in Section 4. In Section 5, conclusion for the proposed works, and future works are presented.

2. Related works

Human emotions can be expressed using several behaviours such as gesture [18], text [19], EEG signals [20], and facial expressions [21]. Authors in [22] improved the accuracy of recent emotion recognition systems by learning the spatial, and temporal features of EEG signals using CNN models which require no feature engineering, and achieved the best recognition performance on temporal, and frequency combined features. Authors in [23] proposed a human emotion classification using a hierarchical bi-directional Gated Recurrent Unit (GRU) network from continuous EEG signals which is able to learn more significant feature representation from the EEG sequence. Noroozi et al. [18] defined a complete framework for automatic emotional body gesture recognition, and introduced a person detection, and comment static, and dynamic body pose estimation methods. Pan et al. [24] proposed an EEG-based brain-computer interface (BCI) system for emotion recognition in patients with disorders of consciousness (DOC). Common spatial pattern (CSP), and differential entropy (DE) features in the delta, theta, alpha, beta, and gamma frequency bands were employed to classify the EEG signals.

Chu et al. [25] proposed a multi-level algorithm that combines the spatial, and temporal features. The spatial representations are extracted using a CNN network. While the temporal dependencies are modeled by Long-Short-Term-Memory (LSTM) network. The outputs of the CNNs, and LSTM networks are further aggregated into a fusion network to produce a per-frame prediction. Hasani et al. [26] extracts the spacial, and temporal relations within the face frames in video sequence using a 3D-Inception-ResNet network followed by an LSTM unit. The Facial landmark points are set as inputs to this architicture. Graves et al. [27] adopted two types of LSTM (bidirectional LSTM, and unidirectional LSTM) to model the temporal dependencies within the image sequences. Their experiments proved that the bidirectional network provides a significant performance compared to the unidirectional LSTM. Jain et al. [28] employed an LSTM, and CNN architicture to get the final prediction for facial expressions. First, the background is abstracted, and isolated from the foreground images. Then, the texture patterns are relevant key features are extracted. Finally, the relevant features are extracted to be introduced later to the LSTM-CNN network.

Many emotion related works use one source to recognize emotions. However, the perception of emotion is multi-modal in nature. Therefore, several works that combine different source of emotions are implemented recently, but achieved accuracy that needs further improvement to reach robust recognition performance. Liao et al. [29] proposed two multi-modal fusion methods for emotion recognition which ate sum rule, and product rule. The input signals are EEG, and facial expression. For EEG detection, emotion states are detected by support vector machines (SVM). The face region is detected using the AdaBoost algorithm. The neural network classifier is adopted for facial expression detection. Emotion recognition results based on fusion of both EEG, and facial expression detections show that the accuracies of the two multi-modal fusion detections are 81.25%, and 82.75%, respectively which are higher than that of facial expression (74.38%) or EEG detection (66.88%). Castellano et al. [30] combined facial expressions, speech, and body gestures information for multi-modal emotion recognition, and achieved accuracy of 78.3% using the Bayes network algorithm which outperforms single modality recognition systems.

Deep neural networks are explored in multi-modal fusion in several active works [31,30]. Gunes et al. [32] combined the face, and body modalities. Performance evaluation showed that bi-modal fusion outperformed the classification done using the facial modality alone. Baltrusaitis et al. [33] developed a system that inferred emotions from upper body gestures in addition to facial expressions, including head, and shoulder motions. A multi-level Dynamic Bayesian Network (DBN) models the emotional state depending on the probabilities of the gestures. It is possible to improve any module, or add features to it, but the limitation is

properties like intensity, offset, and onset of the emotional state are not considered which if included could improve the performance further.

As illustrated, the recent multi-modal emotion recognition systems still need further improvement in their performance. The objective of the proposed approach is to investigate combining facial expressions with the EEG signals to recognize person's emotions using deep learning approaches.

3. The proposed approach

The main purpose of the proposed work is to investigate the effectiveness of deep learning approaches in combination with the ensemble learning techniques for multi-modal emotion recognition. The proposed approach combines the face data from the input video, and the EEG signals. In the proposed approach, two phases are created: in the first phase, two 3D-CNN based classifiers are trained to classify the EEG signals, and the video data into their binary classes. This results in two trained models one for each modality. In the second phase, a third model is created based on the fusion chunks from the EEG, and face modalities. The complete framework of the proposed work that contains the two phases is shown in Fig. 1. The detailed description of the main building blocks of the proposed framework shown in Fig. 1 will be explained in the below subsections.

The 3D-CNN architicture is the extension of traditional Convolutional Neural Networks (CNN) deep learning architecture. The 3D-CNN deep learning architicture is proposed to model the temporal dependencies, and the spatial correlation between long duration sequences using the 3D-convolution operation. The 3D-convolution operation produces a set of 3D-feature maps using 3D-filters. The 3D-Convolution Operation is shown in Fig. 2.

The adopted 3D-CNN architecture consists of six basic layers. The input layer. Then, two convolution layers that results in 3D-feature maps. Each convolution layer is followed by a max-pooling layer. The max-pooling layer down-samples the 3D-feature maps from previous layer to decrease the time needed to process volumes with huge dimensions. The last layer is a fully-connected layer to extract the final features. The dimension of the input volume is 5*32*128, where 5 is the number of consecutive frames which models the temporal information, 32, 128 are the height, and the width of the input frame from the EEG, and the face domains respectivly. For the EEG input, 32 represents the number of channels, and 128 represents the samples in the

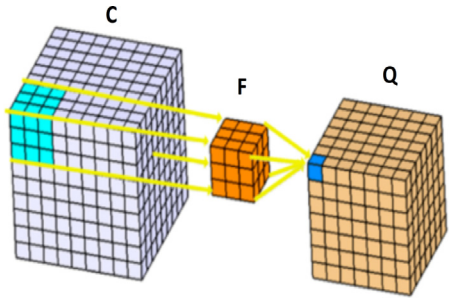

Fig. 2. The 3D-Convolution Operation: C is the input volume, F is convolution filter, and Q is the output of the convolution operation.

frame segment. For the face input, 32, and 128 are the height, and the width of the face frame. In the proposed network, the convolution filter has shape of 3*3*3: where 3, 3, and 3 are its height, width, and depth respectively. The number of feature maps in the first layer is 8. The max-pooling layer has a resolution of 2*2*2. The number of feature maps of the second convolution layer is set to 16. The proposed 3D-CNN network is shown in Fig. 3. Below is a detailed description of the main three approaches for achieving the main goal of the proposed framework.

3.1. EEG-based emotion recognition approach

Usually, the EEG data from every signal is recorded from different channels. The data from every channel is segmented into small segments (frames). The input chunk to the 3D-CNN model is created by combining 5 consecutive frames from 32 EEG channels. The temporal domain data is captured from the number of frames appended together in the EEG chunk. A sample EEG chunk is shown in Fig. 4.

Koelstra et al. [34] developed a database for human emotion analysis using physiological signals (DEAP). It consists of 32-channel EEG signals, and 12 peripheral physiological signals. DEAP dataset rate is 512 Hz, and pre-processed to have a sample rate of 128 Hz. This produces a small number of samples which may affect the performance of any machine learning system to generalize to unseen samples. A data augmentation operation is employed to increase the number of samples. For augmenting the EEG signals, a Gaussian noise signal w is first generated randomly. Then, both the noise signal, and the original EEG signal are mathematically

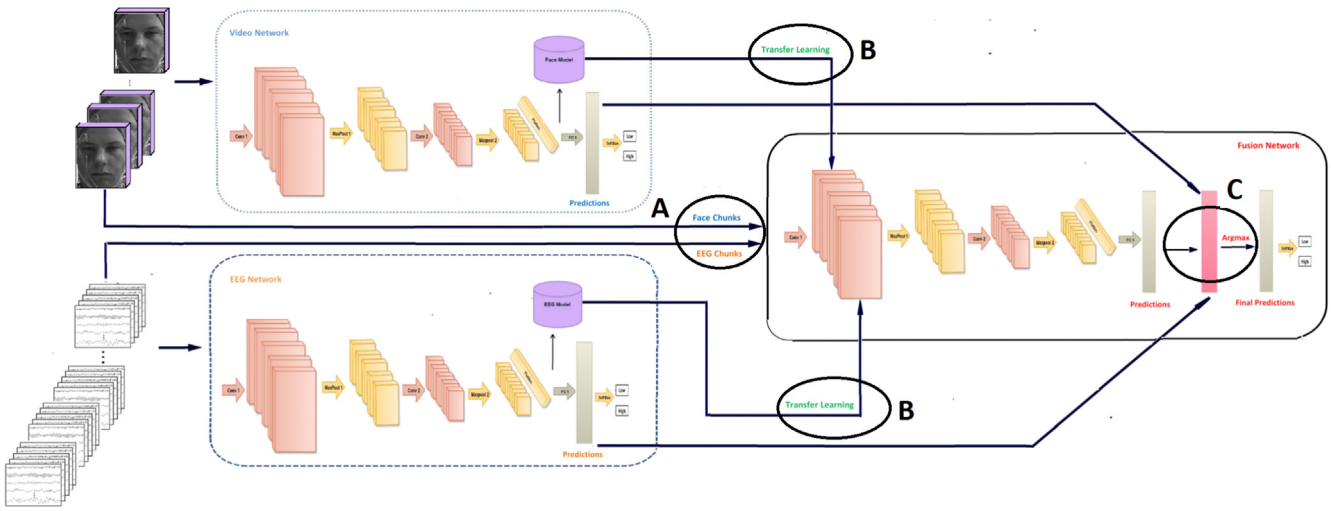

Fig. 1. The framework of the proposed approach using stacking, or bagging techniques for the fusion-based emotion recognition approach.

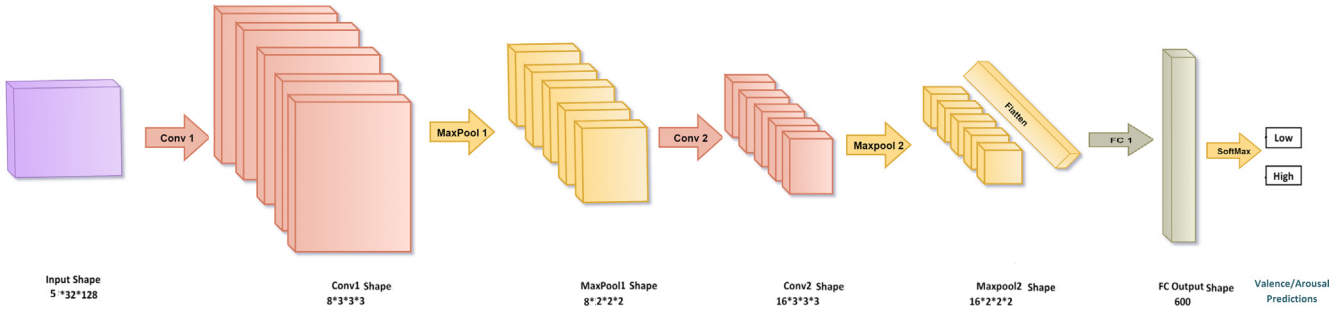

Fig. 3. The proposed 3D-CNN Network.

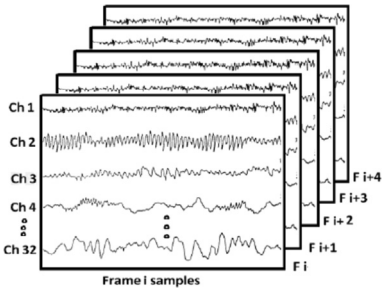

Fig. 4. The 3D-shape of the EEG chunk.

added to create a new noisy version. The augmentation step is applied in the training phase only. During the testing phase, clean versions of the EEG signals are used. The probability density function (P) of a Gaussian random noise signal w is defined by:

$$P(w) = \frac{e^{\frac{(w-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}} \quad (1)$$

Where mu, and sigma are the mean, and the standard variation of the noise signal respectively, and w is the noise signal.

3.2. Face-based emotion recognition approach

Facial expression is considered as a main source of human emotion as stated by Khatri et al. [35]. For predicting emotions based on facial expression in the proposed face system, the following operations were taken: The frame rate of the input video is 30 frames-per second (fps). This produces 1800 frames in total since the duration of one video is one minute (60 s) in the DEAP dataset. To minimize the processing time, only 60 frames are taken per video by taking a frame from each 30 consecutive frames. A data augmentation step is further added to tackle the problem of lacking data. The selected augmentation types are chosen such that they not affect the emotion related features in the image, and preserve the shape of face expressions appear clearly without any distortion. Only 3 types of augmentation are adopted in addition to the original frame; flipping, updating the color, and adjusting the brightness. Flipping left to right means reflecting the frame around the vertical axes. Updating the color means converting the original (Red-Green–Blue) RGB pixel values to (Hue-Scale-Variance) HSV domain. Then, update the hue, saturation, and variance pixel values with some values. Finally, reconvert the updated HSV to RGB again. Updating the brightness means adding a constant value (delta) to the input pixel values. Delta with value 0.2 is chosen in the proposed work. Then, the Mask-RCNN instance segmentation technique is used to remove the background region in the input face frame. The pre-processing steps included for the face system are shown in Fig. 5.

The Mask-RCNN is proved to give enhanced performance for instance segmentation [36]. For the input face frame, the Mask-RCNN returns a class label, and a bounding box around each object. The Mask-RCNN is trained on the COCO dataset. It is labeled for objects such as person, cars, and others (face object is not included). The Mask-CNN is tested on the DEAP dataset. This results in person region in the input image frame. Then, the OpenCV [37] is applied to extract the face pixels from the masked frames. OpenCV, or Open Source Computer Vision library, is a C++ library which is originally developed for image processing, and computer vision. Finally, five consecutive resulting frames are appended in one chunk to model the temporal information in the video data. The final predictions of face chunks are calculated using the SVM classifier of the 3D-CNN output features produced by the last fully connected layer. The outputs predictions are the status of valence, and arousal.

3.3. Fusion-based emotion recognition approach

Input fusion aims to combine data from multiple sensors to achieve improved accuracy, obtain a lower detection error probability, produce a higher reliability, and more specific inferences than could be achieved by the use of a single sensor data alone. The proposed fusion framework has three main stages that affect the recognition accuracy; stages A, B, and C as shown in Fig. 1.

Stage A

At this stage, the face chunks, and EEG chunks are combined to create the fusion chunks. The EEG chunk, and the face chunk of the same 5 s are combined together in one chunk to create the fusion chunk. This is achieved by appending the first frame of the EEG chunk, and the first frame of the face chunk above each other, then combining the consecutive frames in the same manner to create the fusion chunk with depth 5, height 64, and width 128. The duration of the EEG signal is 63 s. The first 3 s were removed since they have no emotion content. In the EEG signals, the original number of samples in the DEAP dataset are 8064 samples for 63 s. Each second (frame) has 128 samples (This is the frame rate of the EEG signals in the DEAP dataset). Since the first 3 s are removed which mapped to 384 samples (3*128), the remaining 60s seconds have 7680 samples (8064–384). This mapped to 60 frames in 60 s (60*128 = 7680). The samples from each 5 consecutive frames (seconds) are combined in one chunk. This produces 12 chunks (60/5 = 12) for each video. For the face video, the frames were 1800 from one video, after changing the frame rate, the number of frames became 60 frames. To create chunks, 5 consecutive frames are combined, which results in 12 chunks during the 60s of each video (60/5 = 12). This means that in the fusion stage, each

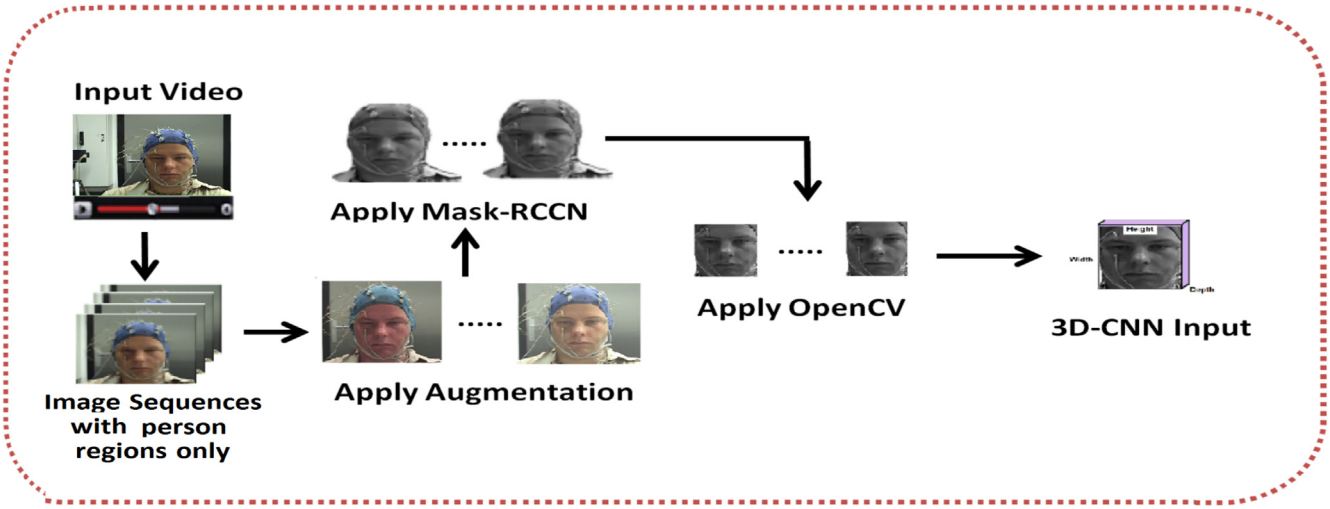

Fig. 5. The pre-processing steps for the face-based emotion recognition system.

chunk in the EEG domain is combined with its corresponding chunk in the face domain since the same number for chunks (12) exists in the 60 s. Fig. 6 describes the fusion process on the time scale of the input video, and signal.

Stage B

Every 3D-CNN system consists of input data, initial weights, 3D-CNN network. The 3D-CNN which is trained in the training part of the input data, and the initial weights to give a trained model. Then, this trained model is tested on the testing part of the input data to give the final predictions of this system. One of the most effective strategies in the deep learning techniques is to replace the initial weights with some weights from a previously trained model to make use of its learnt information, and save time training from scratch without any emotional background. This strategy is applied in the proposed fusion system which starts training with the initial weights of the EEG, and face systems. This is the main idea of transefer learning which transfers the knowledge from a task to another related task with the aim to reduce the generalization error, and achieve an improvement in the performance in terms of percentages of accuracy, and processing time.

Stage C

The final scores are achieved by building a third model that works on the fusion chunks separately. The prediction of this third model is the fusion scores. A further step is added to get the maximum predictions of the three models which are considered as the final fusion scores. Ensemble learning is the art of combining a diverse

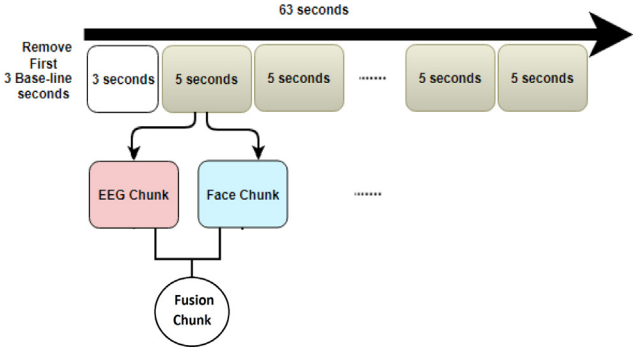

Fig. 6. The fusion process between EEG, and face on the time scale.

set of learners (individual models) together to improvise on the stability, and improve the overall performance of the system. There are several common types of ensemble learning techniques in machine learning including stacking [38], bagging[39], and boosting [40]. For the proposed work, two different score fusion methods are adopted; stacking, and bagging.

3.3.1. Stacking

Each of the proposed face, and EEG emotions systems produces a trained model, and final predictions for valence, and arousal classes. Stacking creates a model to learn from previous trained models. Then, the output score vector is determined by the Map rule which gives the hypothesis that has the maximum probability from the three trained models; EEG, face, and stacking models [41].

$$P = \arg\max\langle S(i), F(i), E(i)\rangle \quad (2)$$

where S, F, and E are the predictions of stacking, face, and EEG emotion recognition approaches, i denotes the current index of prediction of valence, and arousal. While P is the final combined predictions of the three emotion systems.

3.3.2. Bagging

It is a technique that creates many sampled dataset from the original training data in order to reduce the variance. Then, a separate classifier is built for each split. Finally, the results of these multiple classifiers are combined using the Map rule function that get the maximum prediction from the output predictions of the trained models. There are many ways for splitting for the whole data into subsets. The original data may be splitted using k-fold cross validation which splits that data into equal sized subsets, but prevents samples to be repeated over subsets. Another possible method for splitting, random sub-sampling with replacement which allows samples to be repeated in different subsets. Bagging techniques help to reduce the variances error, and avoid overfitting, improve the stability, and the accuracy of machine learning algorithms.

Fig. 7 illustrates the proposed stacking, and bagging ensemble learning techniques. For the stacking technique, the input fusion data is splitted using the cross validation technique with K = 5 to have 0.8 of the original data for training, and the remaining for testing. This produces fusion predictions that are combined later with the face, and the EEG predictions using the Map rule that consider the indeces of maximum predictions from the three predictions. For the proposed bagging, the original data is splitted into

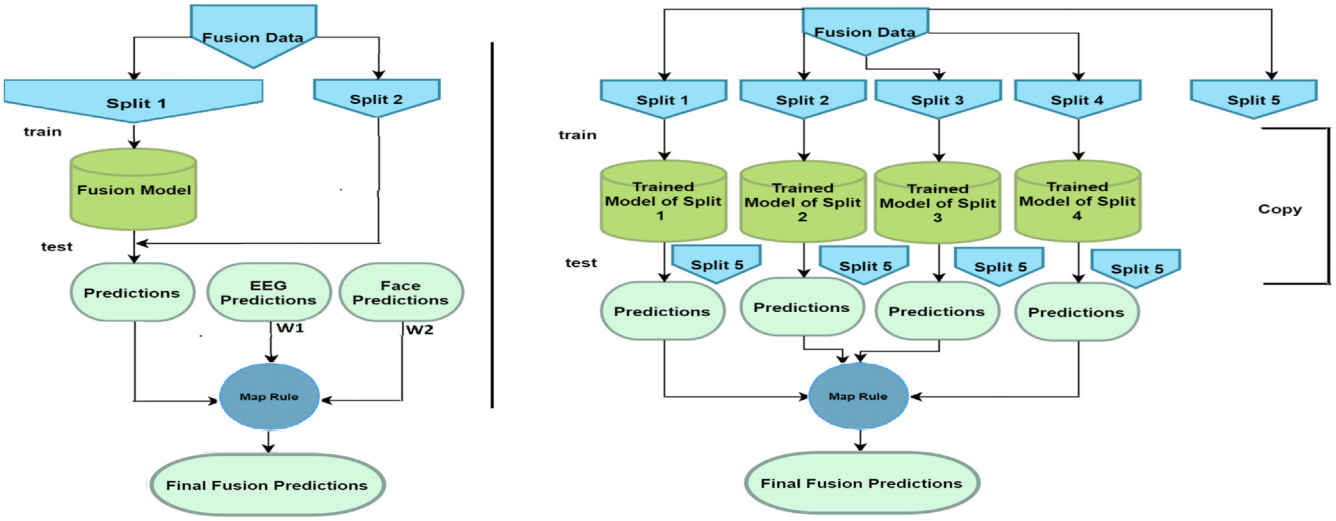

Fig. 7. The proposed stacking (left), and bagging (right) techniques.

5 different splits reserving one split for testing. For each remaining split, a separate model is created producing four trained models. These four trained models are tested using the fifth split which produces four prediction vectors. The final fusion prediction is produced by the Map rule.

Three different stacking methods are adopted for the proposed work; they are namely: model averaging ensemble, weighted sum ensemble, and grid search ensemble. Model averaging is an ensemble learning approach where each model contributes with an equal amount to the final prediction. In the weighted-sum ensemble, and grid search ensemble, the contribution of each ensemble model can be weighted by a coefficient(weight) to indicate its expected performance. A product for all predictions of any model is multiplied by its weights to produce a new updated weighted prediction vector with length N.

$$x = w_i * v_i \quad for \quad all \quad i \in N \tag{3}$$

The main difference between the proposed weighted sum ensemble, and grid search ensemble methods is the way of finding the best values of weights. In the weighted sum ensemble, the weight values are determined experimentally. While, in the grid search ensemble [42], the weight values are determined by selecting the values that minimize the recognition error after starting with initial weight values. Grid search is an automatic algorithm that iterates over the different values of hyper-parameters (weights) to update them, and get the best value of hyper-parameter that minimized the error. Grid search takes the model that you would like to train, and the hyper-parameter values. Then it calculates the mean square error of the hyper-parameter values allowing you to choose the best value. Initially, it starts up with one value for the hyper-parameter, and train the model. Then, used different values to train the model. The process is continued, until all set of values of the hyper-parameter is finished. Each model produces an error, the hyper-parameter value that minimizes the error is selected.

4. Experimental works

4.1. Experimental setup and Results

Several emotion related databases are built to evaluate multimodal recognition systems such as DEAP, SEED, and MAHNOB-HCI databases [43]. DEAP dataset is chosen to evaluate the proposed work since it has the longest duration signals, and contains more multi-modal emotion signals than others [43]. For evaluating the proposed approach, the DEAP dataset which is developed by Koelstra et al. [34] is utilized to model the state of a participant by his Physiological signals. Physiological signals were recorded from 32 participants. the frontal face video data is recorded for only 22 participants. The EEG signals, and peripheral signals were recorded at a sampling rate of 512 Hz.

A set of pre-processing operations are applied to improve the output EEG signals including minimizing the sampling rate to 128 Hz, averaging it to the common reference, and the eye artifacts were removed. Also, a band-pass filter from 4 to 45 Hz is applied to the EEG signals. In addition, the EOG artifacts is removed from the EEG signals. Music videos are used to elicit different emotions, and produce EEG signals, and video data for each participant in the DEAP dataset after performing self-assessment for each music video on a scale between 0 to 9 for arousal, valence, like/dislike, and dominance. Only the valence, and the arousal classes are adopted in the proposed work. To convert these scales to labels, the valence scale of 1–4 was mapped to low, and 5–9 to high, respectively. The arousal scale of 1–4 was mapped to low, and 5–9 to high, respectively. In order to save time required to process a huge number of pixels in the proposed deep learning model, and have similar input to the 3D-CNN model in the fusion system, the input frame is resized to small ratio. The video data, and EEG segments are resized to 32*128 height, and width respectively. The number of features extracted from each system is 600, 600, and 200 for the EEG, the face, and the fusion systems respectively. The number of features are chosen experimentally.

Table 1 lists the experimental results of the proposed framework in terms of the average accuracy over 22 users from the DEAP with both the EEG, and the face data. The classification accuracies of the face, and the EEG detection approaches are also listed. In addition, the results from the two proposed ensemble learning techniques are shown in details. For the face part, three main experiments are conducted: the face system without augmentation using the 3D-CNN classifier, the face system without augmentation using the SVM classifier, and the face system with augmentation using the SVM classifier. The SVM is tested as a classifier to enhance the system classification results. The SVM classifier proved to give better recognition results than the 3D-CNN one. In the SVM experiment, the SVM classifier works on the features extracted from the 3D-CNN deep learning architicture. The augmentation process proves t give better accuracy than without

Table 1
The average accuracy: stacking, and bagging.

Systems with details		Valence	Arousal
Face + 3D-CNN		65.43%	69.17%
Face + SVM		70.28%	71.45%
Face + SVM + Augment		74.93%	75.15%
EEG + 3D-CNN		79.31%	77.97%
EEG + 3D-CNN + Augment		82.32%	84.12%
Bagging + Augment	Sub-sample	86.22%	86.22%
	K-Fold	85.49%	87.75%
	Average Ensemble	84.46%	86.49%
	Weighted Sum Ensemble	84.83%	86.64%
	Grid Search Ensemble	84.40%	87.43%
Stacking + Augment + Transfer Learning (Proposed Fusion Work)	Average Ensemble	95.98%	96.31%
	Weighted Sum Ensemble	95.38%	95.15%
	Grid Search Ensemble	**96.13**%	**96.79**%

augmentation experiments. For the EEG part, two main experiments are conducted namely EEG with, and without augmentation. Both experiments are conducted using the 3D-CNN classifier. The number of augmentation times are three that mapped to the augmentation times in face part: flip, change the color, and adjust the brightness. Therefore, three noisy versions of the input EEG signal, in addition to the original sample data. Fig. 8 shows the fusion of EEG segments with the face frames in the case of augmentation. In addition, it describes a sample fusion chunk with its dimensions.

In the fusion part, two main ensemble learning techniques are experimented namely bagging, and stacking. Both of them are conducted using the augmentation process. An additional experiment is conducted using the stacking technique with the augmentation process, and transferring the weights from the EEG, and the face trained models to the fusion approach. From the experimental results, the EEG modality gives better accuracy than the face one. Consequently, in the proposed stacking technique, larger weight values are set to the EEG modality than the face one. the weight values that are chosen in the proposed work are 0.9, and 0.4 for EEG, and Face modalities respectively. The weight values for the three proposed stacking methods are shown in the below equation such that W is the weight vector with length two which mapped to the EEG, and the face weight values respectivly. The type parameter represents the type of the current working stacking method; average, weighted sum, or grid search. As can be concluded from Table 1, the grid search method is the best method comparing to the weighted sum, and average ensemble method. This is due to using weight values based searching for the values that minimize the error, while the weighted sum method uses weight values that are chosen by hand, and not many values are tested in this method.

$$W = \begin{cases} [1,1] & \text{if } type = Average \\ [0.9,0.4] \quad experimentally & \text{if } type = Weighted \\ [0.9,0.4] \quad initially & \text{if } type = GridSearch \end{cases} \tag{4}$$

4.2. Comparison with other literature

The proposed stacking technique is compared with one baseline technique (bagging), and five state-of-the-art approaches [44–48]. All compared works use the same DEAP dataset in their experimental work. Fig. 9 shows the comparison of the proposed work with related works in literature. A detailed description of the works in literature that we compare with is listed below. The final cell in the figure represents the proposed fusion results from the stacking fusion method with augmentation in combination with transferring learning stage. The difference in accuracy between the proposed system, and the compared works is also provided at the end of this section.

Tang et al. [44] has suggested two models to classify emotions from EEG signals, and peripheral physiological signals. The first model is referred as Bi-modal Deep De-noising Auto-Encoder (BDDAE) which is an extension of the original De-noising Auto-encoders (DAE). The second model is the Bi-modal-LSTM which models the temporal information in input features from data generated from EEG signals, and eye movement data. Bi-modal-LSTM obtains better performance on both arousal, and valence classification tasks. Their performance for valence, and arousal are 83.82%, and 83.23% respectively. Liu et al. [48] demonstrated a Bi-modal Auto-Encoder (BDAE) to model the shared representation of EEG

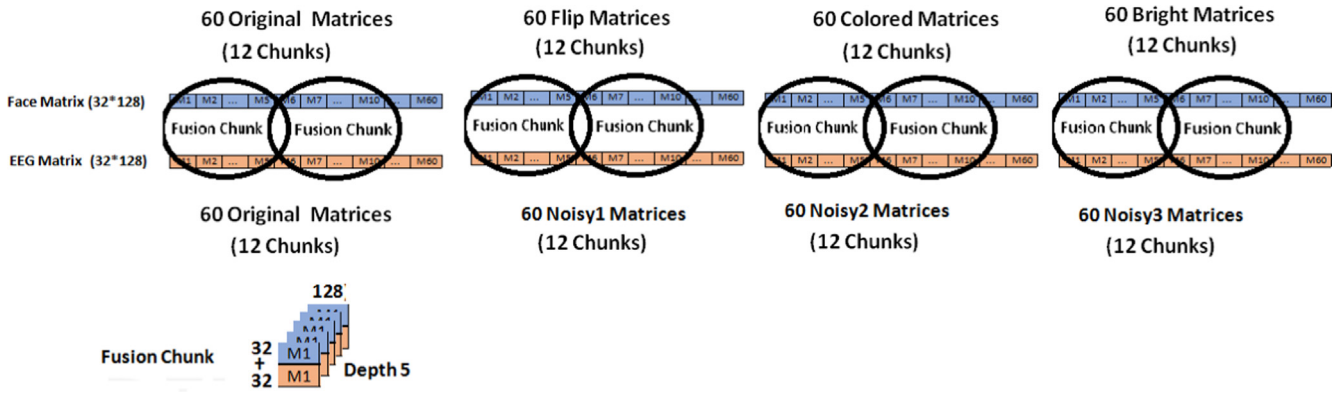

Fig. 8. The fusion process of the EEG signals and the face frames in the augmentation case.

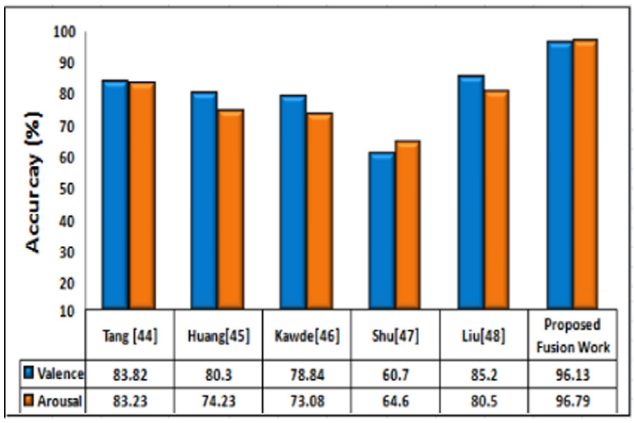

Fig. 9. Comparison of the proposed fusion work (stacking with augmentation, and transfer learning) with the related works in literature.

features, or eye movements which is later classified by a linear SVM classifier to enhance the performance of multi-modal emotion recognition models. Their proposed approach achieved 85.2%, and 80.5% for valence, and arousal binary classes respectivly. Shu et al. [47] proposed restricted Boltzmann machine (RBM) to model the dependencies among multiple physiological signals specifically, the visible nodes of RBM represent EEG signals, and peripheral physiological signals. The RBM generates new representation from multiple physiological signals. Then, the support vector machine is adopted to recognize users' emotion states from the generated features. This provided 60.7%, and 64.6% classification accuracy for valence, and arousal respectivly.

Kawde et al. [46] presented the deep learning network (DLN) to determine the correlation between features from EEG, EMG, and EOG signals. It is introduced into two semi-supervised algorithms, namely the Stack Auto-Encoder (SAE), and Deep Belief Network (DBN). Decision fusion is used to give more precise explanation for all signals. Their classification accuracy for valence, and arousal classes are 78.84%, and 73.08% respectivly. Huang et al. [45] investigated the use of facial expression, and EEG signals in a combined fusion system. For face system, CNN architecture is used to detect the state of valence, and arousal. For EEG detection, support vector machine classifier. All experiments are conducted using DEAP dataset. Their Face, and EEG fusion system are 68.00%, and 70.00% for two dimension classes: valence, and arousal.

The best accuracy achieved from the proposed work is 96.13%, and 96.79% for valence, and arousal classes respectivly using the grid search stacking method with data augmentation, and transfer learning methods. The results from the proposed work are better than the results from the work developed by Huang et al. [45] who combined modalities similar to the proposed method; the EEG signals, and the face data, and reached fusion accuracy of 68.00%, and 70.00% for valence, and arousal classes respectivly. Their results may be worth since they use the decision level fusion only while our work make use of the data, and decision level fusion. In addition, their work neglect the temporal correlation between the EEG, and face segments. Our work is better than the work presented by Liu et al. [48] who combined the EEG signals, and eye movement data, and achieved 85.2%, and 80.5% for valence, and arousal binary classes respectivly using Bi-Modal Auto-Encoder Model which is the best recent fusion system for emotion recognition in literature. This may be due to working on the eye region only which neglects the remaining parts of the face which have rich emotional information. The proposed results proves the superior effect of transferring the knowledge from single modality systems to the fusion system compared to other works in literature.

4.3. Discussion

For the fusion of the facial expressions, and the EEG signals for emotion recognition, two methods are proposed in this study; stacking, and bagging. One data set containing facial videos, and EEG data is used to evaluate these methods. Significant results were obtained for both single modalities. Moreover, two fusion methods outperformed the single modalities.

In the proposed work, a significant improvement for the multi-modal fusion detection is found compared to single modality detection. The reason could be due to the fact that the facial expression detection has a fast, and strong but fluctuating response. While the EEG detection has a a stable response over the trial time. In addition, human could trick machine learning systems since they are able to know how to pretend via their facial expressions. The drawback of facial expression detection could be compensated using the EEG detection to a large extent. Thus, the facial expression detection, and EEG detection are complementary to each other, and the multi-modal fusion system should achieve higher accuracies using both detections than using one of the two detections alone.

Although most studies combine data at the feature level, the score fusion method is preferred for combining the information from different sources since it proved to have a strong reliable advantage of combining data. On the one hand, combining sources at the feature level is difficult to achieve in practice since different modalities may not be compatible (e.g., video data, and EEG signals in this study). In addition, feature level produces a high dimensional space that require further dimentionality reduction procedure to reduce the feature space. On the other hand, in the score level fusion, the knowledge from different sources can be applied separately. In this work, the facial expression, and EEG signals have their own capabilities, and limitations as mentioned above. This information can be used to improve the recognition performance. In addition, it is relatively easy to access, and combine the score generated by deep learning models pertained on facial expression, and EEG modalities. Thus, fusion at the score level is preferred in this study.

For the proposed fusion approach, data fusion, and score fusion levels are combined to get the final predictions. The EEG signals, and the facial expressions from the input participant are combined at the data level. Then, the combined data is set as an input to the proposed 3D-CNN deep learning architicture. Finally, the final predictions are calculated using the Map rule from the three output scores from the three systems; Face-based, EEG-based, and fusion-based systems. Our other highlight was solving this problem by pre-training our deep learning model before processing on the target data set. The number of samples extracted from a single object is limited. Therefore, it is challenging to train a complex model, such as the 3D-CNN, using only a small amount of training data without over-fitting. The data augmentation phase is added to avoid over-fitting. The proposed results proves the superior effect of transferring the knowledge from single modality systems to the fusion system. The accuracy is increased by 8.41%, and 9.36% for valence, and arousal classes respectivly compared to the proposed approach that does not transfer the weights between systems.

5. Conclusion

Information fusion technology by combining facial expressions, and EEG signals for recognizing human states is introduced in the proposed work. In this study, two fusion methods are explored for combining the facial expressions, and the EEG signals. In addition to implementing the widely used ensemble fusion methods in which different weights between two models, another novel fusion

method is explored by applying two different bagging techniques. Moreover, the proposed fusion methods outperform single modalities systems. From the experimental results, the 3D-CNN network is demonstrated to extract shared representations that have better performance in terms of accuracy. Based on the conducted comparative study, it is shown that the proposed scheme can achieve better performance in terms of arousal and valence binary classes since an end-to-end deep leaning framework is performed to map of the video data, and the EEG signals directly to emotion states instead of trying to manually extract features from the input frames, and model the temporal information in video, and EEG data.

Deep learning methods are often difficult to apply due to the limits of the samples. Therefore, it is challenging to train a complex model, such as a 3D-CNN using only a small amount of training data without over-fitting. Our other highlight is solving this problem by pre-training our single modality systems before processing on the target fusion task. This reduces the generalization error, and the overall time taken to develop, and learn a model is reduced. The main contributions of the proposed work that can be summarized as following: The proposed approach introduces a novel framework which investigates the use of the 3D-CNN for extracting related features in the EEG signals, and the video data with the combination of ensemble learning techniques. In addition, transferring the knowledge from a single modality system to the fusion system results in improvement in the overall time taken to develop, and learn a model, and increases the performance of the proposed fusion framework. Besides, The combination of EEG signals, and face information proves its superiority in comparison with other recent modalities combination. Finally, in the proposed work, data fusion, and score fusion are utilized on one compact fusion system. This study still has further issues that need to be addressed in the future. Most of recent emotion recognition systems are developed to be off-line, and huge processing are required to convert them to be on-line to simulate the human life, and facilitate the way people live. In addition, the DEAP dataset suffer from having many electrodes on different positions on the face, and this affect the process of face segmentation, and hence affect the performance of the proposed emotion recognition from the face region.

References

[1] Chang WJ, Schmelzer M, Kopp F. et al. A deep learning facial expression recognition based scoring system for restaurants. In: International Conference on Artificial Intelligence in Information and Communication (ICAIIC). p. 251–4.
[2] Shanok A, Nathaniel NA Jones, Lucas NN. The nature of facial emotion recognition impairments in children on the autism spectrum. Child Psychiatry Human Devel. 2019;50(4):661–7.
[3] Ekman P, Friesen WV. Unmasking the face: a guide for recognizing emotions from facial expressions. 1st ed. En-glewood Cliffs, N. J: Prentice Hall; 1975.
[4] Feldman L. Valence focus, and arousal focus: Individual differences in the structure of affective experience. J. Personality, Social Psychol. 1995;69:53–166.
[5] Mehrabian Albert. Communication without words. Psychol. Today 1968:53–6.
[6] Ko B. A brief review of facial emotion recognition based on visual information. Sensors 2018;18(401).
[7] Elbarougy R. Speech emotion recognition based on voiced emotion unit. Int. J. Computer Appl. 2019.
[8] Salama Elham S, El-Khoribi Reda A, Shoman Mahmoud A, Mohamed A. Wahby Shalaby. EEG-Based emotion recognition using 3D-convolutional neural networks. Int. J. Adv. Computer Sci., Appl. 2018;9(8):329–37.
[9] Hall DL, Llinas J. An introduction to multi-sensor data fusion. Proc. IEEE 1997;85(1):6–23.
[10] M.A.W. Shalaby, N.R. Ortiz, H.H. Ammar. A Neuro-Fuzzy Based Approach for Energy Consumption, and Profit Operation Forecasting, in: Hassanien A., Shaalan K., Tolba M. (eds) Proceedings of the International Conference on Advanced Intelligent Systems, and Informatics 2019. AISI 2019. Advances in Intelligent Systems, and Computing, Vol(1058). 2020..
[11] Khaled K, Shalaby MAW, El Sayed KM. Automatic fuzzy-based hybrid approach for segmentation, and centerline extraction of main coronary arteries. Int. J. Adv. Comput. Sci. Appl. 2017;8(6):258–64.
[12] Shalaby MAW. Fingerprint recognition: a histogram analysis based fuzzy c-means multilevel structural approach Ph.D. dissertation. Concordia University; 2012.
[13] Kasim NAB, Mat NH Abd, Rahman Z Ibrahim, Abu Mangshor NN. Celebrity face recognition using deep learning, Indonesian. J. Electr. Eng., Computer Sci. 2018;12(2):476–81.
[14] Al A, Embaby MAW, Shalaby KM Elsayed. FCM-based approach for locating visible video watermarks. Symmetry 2020;12(3):339–57.
[15] Yang H, Han J, Min K. Distinguishing emotional responses to photographs, and artwork using a deep learning-based approach. Sensors 2019;19(24).
[16] López-Larrosa S, Sánchez-Souto V, Ha AP, Cummings EM. Emotional security, and interparental conflict: responses of adolescents from different living arrangements. J. Child, Family Studies 2019;28(5):1169–81.
[17] Maturana Daniel, Scherer Sebastian. VoxNet: A 3D-convolutional neural network for real-time object recognition. In: Proceedings of International Conference on Intelligent Robots, and Systems.
[18] Noroozi Fatemeh et al. Survey on emotional body gesture recognition. IEEE Trans. Affective Comput. 2018.
[19] H.O., Vong Anh, et al. Emotion recognition for vietnamese social media text, preprint, 2019..
[20] Yang Heekyung, Han Jongdae, Min Kyungha. A multi-column CNN model for emotion recognition from EEG signals. Sensors 2019;19(21).
[21] Shan Li, Weihong Deng. Deep facial expression recognition: a survey, 2018..
[22] Chen JX et al. Accurate EEG-based emotion recognition on combined features using deep convolutional neural networks. IEEE Access 2019;7:44317–28.
[23] Chen JX, Jiang DM, Zhang YN. A hierarchical bi-directional GRU model with attention for EEG-based emotion classification. IEEE Access 2019;7:118530–40.
[24] Pan J, Xie Q, Huang H, He Y, Sun Y, Yu R, Li Y. Emotion-related consciousness detection in patients with disorders of consciousness through an EEG-based BCI system. Front. Hum. Neurosci. 2018;12:198.
[25] Chu WS, Torre FD, Cohn JF. Learning spatial, and temporal cues for multi-label facial action unit detection. In: Proceedings of the 12th IEEE International Conference on Automatic Face, and Gesture Recognition. p. 1–8.
[26] B. Hasani, M.H. Mahoor. Facial expression recognition using enhanced deep 3D-convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision, and Pattern Recognition Workshops, pages 1–11, 2017..
[27] Graves A, Mayer C, Wimmer M, Schmidhuber J, Radig B. Facial expression recognition with recurrent neural networks. In: Proceedings of the International Workshop on Cognition for Technical Systems. p. 1–6.
[28] D.K. Jain, Z. Zhang, K. Huang, Multi angle optimal pattern-based deep learning for automatic facial expression recognition, Pattern Recognition Lett, pages 1–9, 2017..
[29] Huang Y, Yang J, Liao P, Pan J. Fusion of facial expressions, and EEG for multimodal emotion recognition. Comput. Intell., Neurosci. 2017;2:1–8.
[30] Castellano Ginevra, Villalba Santiago, Camuri Antonio. Recognizing human emotions from body movement, and gesture dynamics. In: Proceedings of International Conference on Affective Computing, and Intelligent Interaction. p. 71–82.
[31] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, Andrew Ng. Multi-modal deep learning, in: Proceedings of the 28th International Conference on Machine Learning (ICML-11), pages 689-696, 2011..
[32] H. Gunes, M. Piccardi. Affect recognition from face, and body: early fusion vs. late fusion, in: Proceedings of IEEE Int'l Conference on Systems, Man, and Cybernetics, Vol(4), 2005..
[33] Tadas Baltrusaitis, Daniel McDuff, Ntombikayise Banda, Marwa Mahmoud, Rana el Kaliouby, Peter Robinson, et al. Real-time inference of mental states from facial expressions, and upper body gestures, in: Proceedings of the IEEE International Conference on Automatic Face, and Gesture Recognition, and Workshops, pages 909-914, 2011..
[34] Koelstra Sander, Muhl Christian, Soleymani Mohammad, Lee Jong-Seok, Yazdani Ashkan, et al. DEAP: a database for emotion analysis using physiological signals. IEEE Trans. Affective Comput. 2012;3(1). 18-13.
[35] Jyoti Kumaria R, Rajesha KM Pooja. Facial expression recognition: a survey. Int. J. Computer Sci., Inform. Technol. 2014;5(1):149–52.
[36] Kaiming He, Georgia Gkioxari, Piotr Dollár, Ross Girshick, Mask-RCNN, Computer Vision, and Pattern Recognition, pages 1–12, 2018..
[37] Hai-Wu Lee, Fan-Fan Peng, Xiu-Yun Lee, Hong-Nian Dai, Ying Zhu, et al, Research on face detection under different lighting, in: IEEE International Conference on Applied System Invention (ICASI), 2018..
[38] Wolpert DH. Stacked generalization. Neural Networks 1992;5(2):241–59.
[39] Breiman L. Bagging predictors. Mach. Learn. 1996;24(2):123–40.
[40] Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning, and an application to boosting. J. Computer, Syst. Sci. 1997;55(1):119–39.
[41] Pitsikalis Vassilis, Katsamanis Athanassios, Papandreou George, Maragos Petros. Adaptive multi-modal fusion by uncertainty compensation. In: Proceedings of International Conference on Spoken Language Processing. p. 2458–61.
[42] Bergstra J, Bengio Y. Random search for hyperparameter optimization. J. Mach. Learn. Res. 2012;13:281–305.
[43] Shu Lin, Xie Jinyan, Yang Mingyue, Li Ziyi, Li Zhenqi, Liao Dan, Xiangmin Xu, Yang Xinyi. A review of emotion recognition using physiological signals. Sensors 2018;18(7):2074.
[44] Tang Hao, Liu Wei, Zheng Wei-Long, Lu Bao-Liang. Multi-modal emotion recognition using deep neural networks. In: Proceedings of International Conference on Neural Information Processing. p. 811–9.

[45] Huang Yongrui, Yang Jianhao, Liu Siyu, Pan Jiahui. Combining facial expressions, and electroencephalography to enhance emotion recognition. Future Internet 2019;11(5):105.
[46] Kawde Piyush, Verma Gyanendra K. Multi-modal affect recognition in V-A-D space using deep learning. In: Proceedings of International Conference Smart Technology Smart Nation. p. 890–5.
[47] Shu Yangyang, Wang Shangfei. Emotion recognition through integrating EEG, and peripheral signals. In: IEEE International Conference on Acoustics, Speech, and Signal Processing. p. 2871–5.
[48] Wei Liu, Wei-Long Zheng, and Bao-Liang Lu. Emotion recognition using multimodal deep learning, in: International Conference on Neural Information Processing, pages 521–529, 2016..