Journal of Information Systems and Telecommunication
Vol.11, No.3, July-September 2023, 209-221
http://jist.acecr.org
ISSN 2322-1437 / EISSN:2345-2773

Recognition of Facial and Vocal Emotional Expressions by SOAR Model

Matin Ramzani Shahrestani[1], Sara Motamed[2*], Mohammadreza Yamaghani[3]

[1].Department of Computer Engineering, Rasht Branch, Islamic Azad University, Rasht, Iran
[2].Department of Computer, Fouman & Shaft Branch, Islamic Azad University, Fouman, Iran
[3].Department of Computer, Lahijan Branch, Islamic Azad University, Lahijan, Iran

Received: 23 Oct 2022/ Revised: 04 Dec 2022/ Accepted: 10 Jan 2023

Abstract

Todays, facial and vocal emotional expression recognition is considered one of the most important ways of human communication and responding to the ambient and the attractive fields of machine vision. This application can be used in different cases, including emotion analysis. This article uses six basic emotional expressions (anger, disgust, fear, happiness, sadness, and surprise), and its main goal is to present a new method in cognitive science, based on the functioning of the human brain system. The stages of the proposed model include four main parts: pre-processing, feature extraction, feature selection, and classification. In the pre-processing stage, facial images and verbal signals are extracted from videos taken from the enterface'05 dataset, noise removal and resizing is performed on them. In the feature extraction stage, PCA is applied to the images, and the 3D-CNN network is used to find the best features of the images. Moreover, MFCC is applied to emotional verbal signals, and the CNN Network will also be applied to find the best features. Then, fusion is performed on the resulted features and finally Soar classification will be applied to the fused features, to calculate the recognition rate of emotional expression based on face and speech. This model will be compared with competing models in order to examine the performance of the proposed model. The highest rate of recognition based on audio-image was related to the emotional expression of disgust with a rate of 88.1%, and the lowest rate of recognition was related to fear with a rate of 73.8%.

Keywords: Emotion Recognition; Facial and Vocal Emotional Expressions; Cognitive Model; Soar Model.

1- Introduction

Emotion recognition is the task of automatically discovering the precise emotional ideas of a person, which is one of the most important and challenging problems in the field of human-machine communication. Various computational models of emotional learning have been studied in the literature to understand the emotional expression of people. Since the brain is the main organ of the human central nervous system, computational models inspired by the human brain have a high ability to recognize and classify the pattern [1]. Emotion analysis thought mining, and analyzing subjectivity are related fields of research that use different techniques derived from Natural Language Processing (NLP), Information Retrieval (IR), and Structured and unstructured Data Mining (DM). Most parts of available data throughout the world are unstructured (like text, speech, verbal and visual), which leads to significant research challenges [2]. Ekman et al. have examined facial changes in the form of these muscles' activity, and have collected some of them in the form of the Facial Action Coding System (FACS) [3]. In this system, action units refer to a change in the face that, firstly, can be done alone, and secondly, is indivisible. The limitation of this system is that the expression of action units is only based on local specifications [4]. A new type of facial feature that is commonly used is classified into two main categories of appearance and geometric features. Geometric features refer to the shape and location components of the face, such as eyes, eyebrows, lips, etc., and display the appearance feature, and facial texture, such as wrinkles, ridges, and dimples. Geometric features are obtained based on the alignment results of face components through the Active Appearance

✉ **Sara Motamed**
motamed.sarah@gmail.com

Model (AAM) method [8]. Meanwhile, the representation of local binary patterns and local phase quantization are two representations of appearance-based feature extraction methods [9-15].

Existing research, in general, uses two types of dynamic and static classification. Dynamic classifiers (Hidden Markov Model (HMM)) use several video frames and perform classification by analyzing time patterns from the analyzed areas, or extracted features. Static classifiers classify each frame in a video into a category of facial expressions, based on specific video frame results. In general, these methods are such that first some features of the images are extracted, then they are classified into a classification system, and as a result, one of the emotional categories will be selected. Automatic emotion recognition from face video images encountered many challenges, including finding the face in the image, localizing the dimensions of the eyes, nose, and mouth, revealing the changes in the face and its components, during a certain period of time, and also establishing the relationship between these changes with the person's emotional expression. Each of these issues has its own variation, depending on environmental and personal conditions. For example, at the time detecting a face and finding the exact location of its components, the composition, and makeup of the facial appearance, such as wearing glasses and the angle of the head, results in several problems, each of which has been the subject of extensive independent research. In order to identify emotion from video images, Yacoob and Black model locally the facial action in different areas [16].

In [41], a novel approach is proposed to utilize two face images. In the proposed method, the face component displacements are highlighted by subtracting neutral image from emotional image; then, LBP features are extracted from the difference image. The proposed method is evaluated on standard databases and the results show a significant accuracy improvement compared to DLBPHS.

In [42], we propose an Affine Graph Regularized Sparse Coding approach for resolving this problem. We apply the sparse coding and graph regularized sparse coding approaches by adding the affinity constraint to the objective function to improve the recognition rate. Several experiments has been done on well-known face datasets such as ORL and YALE. The first experiment has been done on ORL dataset for face recognition and the second one has been done on YALE dataset for face expression detection. Both experiments have been compared with the basic approaches for evaluating the proposed method. The simulation results show that the proposed method can significantly outperform previous methods in face classification. In addition, the proposed method is applied to KTH action dataset and the results show that the proposed sparse coding approach could be applied for action recognition applications too.

In [43], focuses on facial expression to identify seven universal human emotions i.e. anger, disgust, fear, happiness, sadness, surprise, and neu7tral. Unlike the majority of other approaches which use the whole face or interested regions of face, we restrict our facial emotion recognition (FER) method to analyze human emotional states based on eye region changes. The reason of using this region is that eye region is one of the most informative regions to represent facial expression. Furthermore, it leads to lower feature dimension as well as lower computational complexity. The facial expressions are described by appearance features obtained from texture encoded with Gabor filter and geometric features. The Support Vector Machine with RBF and poly-kernel functions is used for proper classification of different types of emotions. The Facial Expressions and Emotion Database (FG-Net), which contains spontaneous emotions and Cohn-Kanade(CK) Database with posed emotions have been used in experiments. The proposed Method was trained on two databases separately and achieved the accuracy rate of 96.63% for spontaneous emotions recognition and 96.6% for posed expression recognition, respectively.

In [44], introduces a new classification method using multi-constraints partitioning approach on emotional speech signals. To classify the rate of speech emotion signals, the features vectors are extracted using Mel frequency Cepstrum coefficient (MFCC) and auto correlation function coefficient (ACFC) and a combination of these two models. This study found the way that features' number and fusion method can impress in the rate of emotional speech recognition. The proposed model has been compared with MLP model of recognition. Results revealed that the proposed algorithm has a powerful capability to identify and explore human emotion.

Kumari et al., (2015), stated that the facial expression recognition system has many applications, and is not just limited to understanding human behavior, revealing mental disorders, and expressing human structure. Two popular methods for automated FER systems are geometry-based and appearance-based [37]. The emotional expression of the face is divided into six categories: anger (combination of lowering of the eyebrows, raising of the upper eyelid, narrowing of the eyelids, tightening of the lips), disgust (combination of the wrinkle of the nose, lowering of the corners of the lips, lowering of the upper lip), fear (combination of raising the eyebrow, raising the upper lip, narrowing of the eyelids, widening of the lips, dropping the jaw), happiness (combination of raising the chin, closing the corners of the lips), sadness (combination of raising the eyebrow, closing the corner of the lip), surprise (combination of raising the eyebrow, raising the upper eyelid, dropping the jaw) [7,17]. The main advantage of the deep learning neural network approach in facial emotion recognition is its capacity to learn a large

amount of data. Secondly, recent neural network architectures allow end-to-end training from the representation stage to the classification stage, and finally, they have shown high efficiency compared to non-neural network systems and can learn the limits of complex decision-making in classification. The only disadvantage of these networks is their lack of interpretability. Identifying features play an important and influential role in forecasting results. Moreover, to train deep learning neural networks, many data sets are needed for better performance. This article will use the combination model of deep learning neural networks and the Soar model to solve these problems.

Soar is a symbolic cognitive architecture, which implements problem-solving as behavior in line with the goal, and includes research through the problem and learning from its results [5]. This architecture is used for a wide range of applications, such as routine tasks and solving open problems, which are common for artificial intelligence, and also for interacting with the outside world, either in simulation or in reality [6]. In Soar architecture, the problem is solved by decomposing the goal into hierarchical sub-problems. Procedural memory stores previous states of problem-solving, while semantic memory deals with known evidence.

The proposed model of this article consists of the main parts of pre-processing, feature extraction, feature selection, and classification. First, pre-processing stage, facial images, and verbal signals are extracted from videos, and noise removal and resizing are performed on them. The output of this stage will move to the next stage, and in the feature extraction stage, PCA and MFCC will be applied for face image and verbal signal, respectively. In the best features selection stage, 3D-CNN face images and CNN verbal signals will be applied to the obtained features. Finally, and in the fusion stage, fusion is performed on the best features, and the final vector will be sent to the Soar classifier, to calculate the recognition rate of emotional expression based on facial and verbal emotional expressions.

In the feature extraction stage, the reason for choosing PCA on face images is to use different components to achieve more details and simplify them, and choosing MFCC on verbal signals leads to the necessary compression and understanding of the verbal source, and contributes to the necessary identification and diagnosis. Moreover, in the feature selection stage, the reason for choosing 3D-CNN on face images is that no dimension is removed from the images, and considering the temporal information in dynamic images, it leads to more efficient and better feature selection. CNN networks are used to select the best verbal features because they are very resistant to changing the image's scale and size, and even in noisy images, they provide a suitable response to the output. On the other hand, fusion leads to obtaining a uniform vector for single and optimal use in decision making. This is why we use the fusion method in this paper before using the Soar classifier. This article is organized as follows: Section 2 cognitive science models are introduced. In Section 3, the proposed method is introduced. Experimental results and conclusions are presented in Sections 4 and 5, respectively.

2- Recognition of Facial and Vocal Emotional Expression

The verbal signal is the fastest and most natural way of communication. Accordingly, speech is used as a fast and efficient method for human-computer interaction. So far, many efforts have been made to verbal recognition. Despite many advances in this field, there is a long gap between natural human-computer interactions. The main reason is the computer's inability to understand the user's feelings. Therefore, in the last few years, emotion recognition by speech is one of the challenging issues in the field of verbal processing. In addition, emotion recognition from speech can be used to extract useful meanings from speech, as well as, is used to recognize the user's personality, which is used in many organizations, including political, military, etc. Emotion recognition from speech has various applications, which have been addressed from different aspects. Among its applications are the development of automatic verbal recognition systems, text-to-speech conversion, driver's mental states report, computer games [12], diagnosis tools in medical sciences [13], for disabled people or autistic children, in order to communicate with others [14], and also as an application in mobile phones [15]. Considering the importance of recognizing facial and vocal emotional expressions, this article presents a practical method inspired by the human brain system.

3- Proposed Model

In order to recognize the emotional verbal and facial expressions, this article has used the fusion of appropriate features resulting from signals and emotional images, and then applied Soar classification. Considering Fig. 1, the procedure sequence of the proposed model includes four main parts: preprocessing, feature extraction, feature selection, and classification. In the first steps, in order to extract suitable features, first pre-processing, and then PCA and MFCC will be applied to face images and verbal signals, respectively. The output of the extracted features is entered into CNN and 3D-CNN so that more suitable features are selected. Before applying the classifier, fusion is applied to all the selected features, and finally, the Soar

classifier will be applied in order to examine the emotional verbal-visual recognition rate.

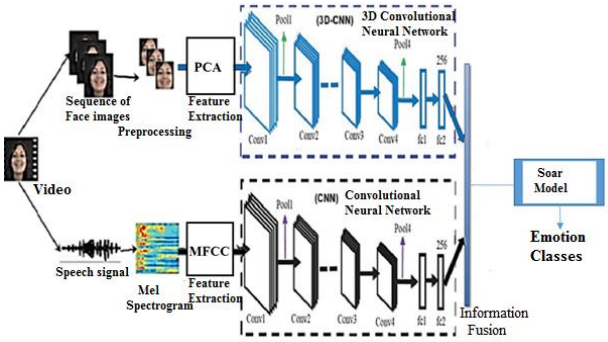

Fig. 1. Outline of the proposed method

According to Fig. 1, the first step is to read the videos from the enterface'05 dataset [30]. To separate the verbal signal and extract the image frames from the video, the separation algorithm with a certain frame rate, and also the separation of the speech from the video will be applied to all the samples. Face images and verbal signals are stored in specific folders, and then pre-processing operations are applied to the frames of images and resulted in speech from the video. In the following, PCA and MFCC are used to extract features from face images, and for the verbal signal, respectively. PCA is an accurate quantitative method to achieve this simplification. This method produces a new set of variables, called principal components. MFCC also compresses the information related to the verbal system, based on understanding, to a small number of coefficients, which this simplification greatly contributes to overall recognition. Moreover, video samples are divided into three categories: training, validation, and testing. In the end, the image frames of each video sample are sent to the 3D-CNN network in the form of a four-dimensional matrix, including the number of sequences of frames of each video, width, height, and the number of images channels [19]. The resulted speech from the video is also sent to CNN for feature selection. The reason for selecting 3D-CNN is that this method is can store data in the entire network, and also can deal with incomplete knowledge, and also tolerate high error. Therefore, it is expected to select appropriate features. The reason for selecting CNN, in addition to the above reasons and speech will be two-dimensional. Finally, in order to classify the obtained features, the Soar method has been used. In Soar, all goal-oriented symbolic tasks are formulated in problem spaces. A problem space consists of a set of states and a set of operators. States represent situations, and operators refer to actions that, when applied to states, lead to other states. Each functional context contains a goal, roles for a problem state, a state, and an operator. Problem-solving is driven by decisions that lead to the selection of problem spaces, states, and operators for the respective context roles. According to a goal, a problem space should be selected, in which the achievement of the goal can be pursued. Then, an initial state must be selected, which represents the initial state. Then, to apply in the initial state, an operator must be selected [24]. This process continues until a sequence of operators is discovered, which transforms the initial state into a state in which the goal is achieved [31].

Emotion recognition is a problem with many parameters and close to each other, and since the Soar architecture has not been used in this field and this model has many free parameters, it can be a suitable model for prediction. Also, this model is inspired by the human brain system, and we believe that any model inspired by the human biological system can work properly.

3-1- Pre-Processing

According to Fig. 1, the first step is to read the video from the enterface'05 dataset and pre-process the images. The separation of image and voice will be performed using call and separation functions in PYTHON. To extract the frames of face images from the video, the separation algorithm with a certain frame rate will be applied to all samples. In these tests, the frame rate of the images is 10 (that is, it extracts 10 frames per second). The images are stored in specific folders, and then, pre-processing operations are applied to the resulted frames from the video. This operation includes face detection, face alignment, and image resizing. The image size of each video frame is changed to 3x100x96 resolution. Moreover, the Mel spectrogram function has been used for verbal conversion, and the general algorithm has been used to calculate the Mel spectrogram.

3-2- Feature Extraction from Images and Speech by PCA and MCCF

PCA is an accurate quantitative method to achieve this simplification. This method results in a new set of variables called principal components. The PCA function is used to find the principal components. PCA is a quantitatively accurate method to achieve this simplification. This method produces a new set of variables called principal components. The reason for choosing PCA is to use different components to achieve more details and simplify them. Fig. 2 shows an example of feature extraction from an image using PCA.

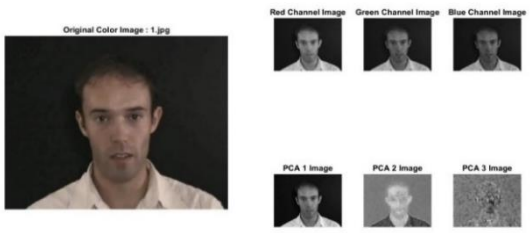

Fig. 2. An example of feature extraction from an image by PCA

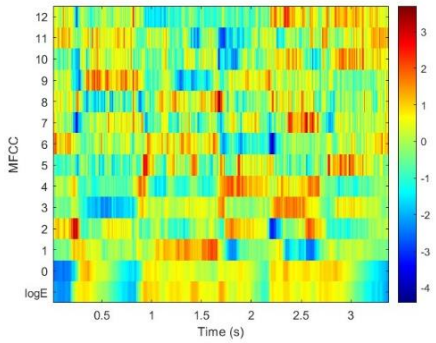

Fig. 3. Example of feature extraction from speech file images by MFCC

Mel-frequency spectrum coefficients are popular features extracted from verbal signals for use in recognition tasks. In the source-filter model of speech, brain coefficients are understood as a filter representative (verbal system). The frequency response of the speech system is relatively smooth, while the spoken verbal source can be modeled as a shock train. As a result, the verbal system can be estimated by the spectral coverage of a verbal segment. The motivational idea of the Mel frequency envelope coefficients is to compress the information about the verbal system (smooth spectrum) into a small number of coefficients based on cochlear perception. Although there is no exact standard to calculate coefficients, the basic steps are outlined in the diagram. Fig. 3 shows an example of feature extraction from speech using MFCC.
According to Fig. 3, the horizontal axis shows the time, and the vertical axis is the calculated value for 12 filters, in which the signal input energy in the time and frequency domain results in a distance between these filters, and different colors.

3-3- Selection of Image and Speech Frame Features using 3D-CNN and CNN

3D-CNN (convolutional neural network) model is used to represent the features of the facial emotional expression in the video [21]. Fig. 4 shows the architecture of the facial expression recognition model using 3D-CNN.
of the facial expression recognition model using 3D-CNN.

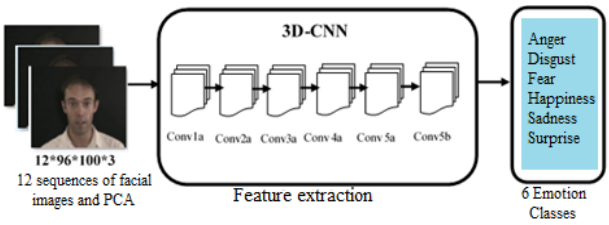

Fig. 4. The architecture of facial expressions recognition model, using 3D-CNN [20]

According to Fig. 4, 12 3D-sequences, with a size of 96*100*3, are entered into the 3DCNN as a feature selection step, and Conv and Pool layers are applied to them. This function and Leaky ReLU have been used due to the popularity of the RelU function in transfer functions, and its operation is shown according to Eqn. 1 and Eqn. 2.

$$RelU \Rightarrow \max(0, x) \quad (1)$$

$$Leaky\ RelU \Rightarrow \max(0.1x, x) \quad (2)$$

3-4- Information Fusion

Information fusion is a set of activities that by using information from several resources obtain more and more accurate information about a subject [32]. The need for such activities is evident in several aspects. In classification problems, considering the structure of the problem and the selected solution, the information fusion is performed in four levels: data fusion, features, classification, and decision making. If different sensors measure the same quantities of the same phenomenon (such as seeing an object from different angles, or from a combination of basic colors), the fusion process is a process at the data level. Moreover, the fusion process can be delayed to the feature level. Many psychological studies have shown theoretically and experimentally, the importance of integrating information from several modalities (for example, emotional verbal, and facial expressions in this design) to obtain a coherent representation and inference of emotion [32-33]. As a result, in recent years, many studies have been conducted on the recognition of human verbal-visual emotion [34-35]. In addition, there are studies in the field of other resources, such as information fusion of facial expressions and head movement [31] and also the fusion of facial expressions and body movement or behavior [36], most of which are independent of the verbal text. In 1971, Friesen and Ekman identified six distinct emotional categories; each characterized by a specific facial expression. These six categories, which include happiness, sadness, anger, surprise, fear, and disgust, are common to all human beings of all nationalities and are called basic emotions [38]. Cohen et al. (2000) have dealt with the automatic face recognition of live video, and to this end, they used dynamic programming methods of pattern pairing and the

Hidden Markov Model (HMM). This has presented existing methods and a new HMM structure for automatic human face recognition from video sequences. The advantage of this structure is that both segmentation and face recognition have been performed automatically using HMM, increasing the differentiation between different face classes. In this work, the dependent and independent expression of the face was investigated, and the average total ratios of detecting the dependent emotional expressions of the person's face, using the special sense of one-dimensional HMM and multi-level HMM, were 78.49 to 82.46%, respectively [6]. Fusion has different levels: information fusion at the feature level, information fusion at the classification level, and information fusion at the decision-making level. According to the performed investigations, we decided to focus on fusion at the feature level due to the combination of the product of the independent features of verbal and facial expression.

Using Fusion for data classification provides results with high accuracy, sensitivity, specificity, accuracy and consistency. This leads to improved modeling metrics such as MCC (Matthews Correlation Coefficient), which is reliable, robust and effective. Also, the amount of rating error is reduced. It extracts better, more important, useful, richer, valuable, relevant, and meaningful information and improves the signal-to-noise ratio (compared to irrelevant information) and provides promising, better, and reasonable results, resulting in a decision. It gets better [39].

3-5- Classification by Soar learning Model

To recognize emotion from facial expressions in the video, the Soar learning model is used [5, 6]. Soar is used as a method based on brain architecture and pattern-based learning. In this method, first, patterns are learned on the basis of deep learning and memory mechanisms, and then, are recognized on the basis of the learned model.

Soar is a symbolic cognitive architecture, which implements problem-solving as behavior in line with the goal, and includes research through the problem and learning from its results. This architecture is used for a wide range of applications, such as routine tasks and solving open problems, which are common for artificial intelligence, and also for interacting with the outside world, either in simulation or in reality. In Soar architecture, the problem is solved by decomposing the goal into hierarchical sub-problems. Procedural memory stores previous states of problem-solving, while semantic memory deals with known evidence [5].

The working memory of an agent is used to evaluate its current position. To this end, this memory uses the perceptual information received through its sensors, as well as the information stored in the long-term memory. Working memory is also responsible for making motor commands or selective actions by the decision-making process module, selecting operators, and tracking potential bottlenecks. In Soar design, decision-making is controlled by the pattern matching part of production rules. An important issue in Soar decision-making is the selection of operators used to change the problem state. Multiple operators can be selected for a problem, and in the case of conflict, it can be resolved through a preference structure. The preference structure can select operators according to different criteria. By going through these options, the seriousness and difficulty of the criteria are reduced, and they show a trend of more adaptability and more innovation in generating ideas; therefore, the preference structure can reflect the problem-solving style as a kind of superiority, which an operator may choose, and move along different problem-solving paths [6].

Algorithm: Generalized SOAR

Input: A, q_0, p_0 and an integer order n

Output: the orthonormal matrix Q_n

1. $\beta = \|q_0\|$
2. $\begin{bmatrix} q_1 \\ p_1 \end{bmatrix} = \frac{1}{\beta}\begin{bmatrix} q_0 \\ p_0 \end{bmatrix}$
3. For $j = 1, 2, \cdots, n-1$
4. $\quad \begin{bmatrix} q_{j+1} \\ p_{j+1} \end{bmatrix} = A\begin{bmatrix} q_j \\ p_j \end{bmatrix}$
5. $\quad$ For $i = 1, 2, \cdots, j$
6. $\quad\quad h_{ij} = q_i^T q_{j+1}$
7. $\quad\quad \begin{bmatrix} q_{j+1} \\ p_{j+1} \end{bmatrix} = \begin{bmatrix} q_{j+1} \\ p_{j+1} \end{bmatrix} - h_{ij}\begin{bmatrix} q_j \\ p_j \end{bmatrix}$
8. $\quad$ End
9. $\quad h_{j+1\ j} = \|q_{j+1}\|$
10. $\quad$ If $h_{j+1\ j} \cong 0$, stop (breakdown)
11. $\quad \begin{bmatrix} q_{j+1} \\ p_{j+1} \end{bmatrix} = \frac{1}{h_{j+1\ j}}\begin{bmatrix} q_{j+1} \\ p_{j+1} \end{bmatrix}$
12. End
13. $Q_n = [q_1 \quad \cdots \quad q_n]$

Fig. 5. Soar architecture algorithm [5]

Q is the canonical response matrix consisting of <mode, state> pairs, which are performed according to the execution of different steps. And the inputs are the states that are determined based on the input matrix A with the matrix P. The details of the method were described in [40]. The decision-making process in Soar is the same as in other systems: it involves matching and firing rules, which is, in fact, a context-dependent representation of knowledge. These conditions describe the current situation of the agent, and the body of rules describes the actions that lead to the creation of structures related to the current conditions in the working memory [22]. Fig. 5 shows the architectures of the classic and current versions of Soar. The classical Soar architecture includes two types of memory: (1) a long-term symbolic memory, which is

encoded using production rules, and (2) a short-term working memory, which is encoded as a graph structure, to make possible the representation of objects in detail and relationships between them [6]. Soar has a deep learning component, which is related both to procedural knowledge and working memory (the latter through the evaluation component). Reinforcement learning in Soar is relatively simple. This adjusted action selection is based on environmental numerical rewards. This evaluation has been applied to all goals and sub-goals in the system, which enables quick identification of good and bad operators for specific conditions [23]. Evaluations lead to evaluations of how goals are achieved, which in turn affects the feelings of the agent. These emotions express the intensity of emotion and thus act as an intrinsic reward for reinforcement learning. Currently, emotions in Soar only connect reinforcement learning with working memory [24]. Fig. 5 shows the Soar algorithm.

4- Results

Python software is used to simulate the proposed model. A validity test was used in all experiments using the K-Fold method and K=10 [25].

4-1- Enterface'05 Dataset Verbal-Visual Emotion

The verbal-visual database enterface'05 was developed in 2005 and included a number of video files [26]. This database consisted of 44 people, each of whom expresses 6 basic emotions (anger, disgust, fear, happiness, sadness, surprise) with five different scenarios and sentences. In total, the database contains 1320 samples of verbal-visual clips, whose verbal signal was recorded with a sampling rate of 48000 Hz, with a resolution of 16 bits and one channel. The average length of each video is 3-4 seconds, and the database language is English. Fig. 6 shows an example of arranged images of different facial expressions in the enterface'05 database [27].

Fig. 6. Some examples of arranged images of different expressions in the enterface'05 database

4-2- Results' Analysis

4.2.1 Analyzing the Results of Facial Emotional Expression

In order to emotion recognition from the facial emotional expressions in the video, in the first step, the database is prepared, and then, the pre-processing operation is performed. In the database preparation phase, first, the database is divided into 2 training and testing sections. So that 70% of the database is considered for training, and 30% for testing. Separation of the image formation is done using call and separation functions in PYTHON. In order to feature extraction, PCA has been used. The complete set of principal components is the same as the principal set of variables. But it is common that the sum of the variances of several principal components exceeds 80% of the total variance of the original data. By examining plots of these new multivariate, researchers often gain a deeper understanding of the driving forces that produced the original data. To use PCA, actual measured data to be analyzed are required [28]. As shown in Fig. 4, a sequence of 10 frames of color face images, each with a size of 96x100, is fed to the 3D convolutional neural network, and the output is equal to the number of basic emotion classes. Table 1 shows the number of layers, the filters of each layer, and the size of each filter in the structure of the 3D-CNN model.

Table 1. The structure details of the used layers in the 3D-CNN model

32 Filters 3*3*3	conv1
2 steps	pool1
64 Filters 3*3*3	conv2
2 steps	pool2
128 Filters 3*3*3	conv3
2 steps	pool3
1024	Fully Connected
256	Fully Connected
Classes Count=6	Softmax

The evaluation indices in this article are accuracy rate, precision rate, and recall rate [29]. The proposed model applied 3D-CNN to select the best features from the emotional expressions of facial images, and the results are shown in Fig. 7.

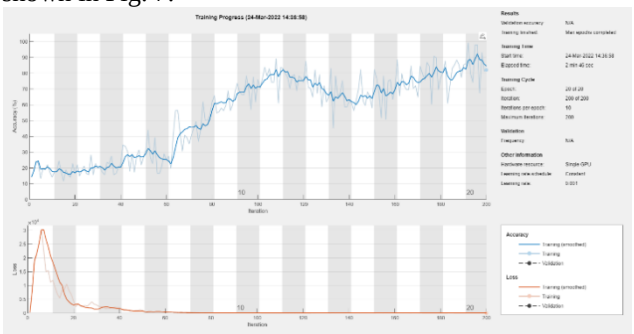

Fig. 7. The Soar output on the 3D-CNN features selected from images

According to Fig. 7, the training in the network started with very low values close to an accuracy of 20%, and

these values increased with the considered number of iterations, which finally reached the final value of 82% in the amount of 200 iterations. In the bottom part of the figure, the values of missing information or failed goals can be seen, which were very high in the initial iterations, and in the following, the value was close to zero. According to the proposed method, the number of input frames to this stage can be different, and Table 2 shows the recognition value in different numbers of frames.

Table 2. Comparison of recognition accuracy according to different sequences of facial expressions images

Count sequences of facial expressions images	accuracy
6	57%
8	62%
10	62.5%
12	63%

As shown in Table 2, by using 10 sequences of images, the recognition accuracy has reached 62.5, which is the highest value compared to 6 and 8 sequences. Of course, this accuracy is only 0.5% different compared to 8 sequences, and if the speed and quantity of identification are considered, 8 sequences can be considered, and in the case of increasing the quality index, 10 sequences can be considered. 12 sequences were considered in the proposed method. After performing the pre-processing operation, the number of sample frames can be less or more than the frames required for processing. If it is less, we copy several copies of the last frame as much as the difference with the required number of frames, and if it is more, we select the coefficients of the frames.

For the proposed facial expression recognition model using 3D-CNN, the learning values are determined as follows. The size of each batch is 32, the number of iteration steps is 500, the learning rate is 0.001, and momentum rate is 0.0001. In this method, the early stopping technique is used during training. To examine the used layers in this model, we have displayed the image of the middle layers for each frame. The facial expression recognition test was performed on the enterface'05 database for 1005 training video samples and 250 test video samples in the form of cross-validation. The results of this test are shown in Tables 3 and 4.

According to Table 3, by training the model and memorizing the detection method in the proposed Soar model, the test data set centered on each of the data set emotions has been presented to the model, and the results have been recorded. Considering the way of selecting the test set, and completely different examples from the training set, "disgust" and "surprise" have shown the highest and lowest recognition rates, respectively. For example, out of 100% of test samples with "disgust " emotion, 86% were correctly recognized, and 14% of the rest were recognized in other emotions. According to Table 4, "disgust " and "happiness" have the highest recognition accuracy in recognizing facial expressions, and "anger" and "fear" have the lowest recognition accuracy.

Table 3. Confusion matrix of facial expression recognition on the enterface'05 database

	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	47%	7%	15%	6%	10%	15%
Disgust	2%	86%	5%	0%	5%	2%
Fear	7%	7%	55%	5%	17%	10%
Happiness	2%	12%	5%	67%	5%	10%
Sadness	5%	2%	14%	2%	67%	10%
Surprise	10%	5%	17%	10%	15%	44%

Table 4. Confusion matrix of facial expression recognition on the enterface'05 database

	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	77%	3%	7%	3%	4%	6%
Disgust	1%	91%	3%	1%	2%	2%
Fear	1%	4%	81%	5%	7%	2%
Happiness	2%	3%	2%	90%	2%	1%
Sadness	2%	4%	2%	4%	87%	1%
Surprise	2%	4%	6%	2%	2%	84%

According to Tables 3 and 4, on average, the recognition percentage of facial expressions for 6 basic emotions, for the test data set from the enterface'05 database, was 73%. The recognition rate of the above two tables, which are specified for each of the emotions, is added together and divided by the number of emotions in the two tables, which, based on this calculation, and the overall recognition rate presented in the simulation, the value of 73% has been recorded by the proposed model as the result of the overall recognition of emotions.

Fig. 8 and Fig. 9 show the recognition accuracy and loss error function for different iteration stages and the validation data set in the training stage of the 3D-CNN network for recognizing facial expressions on the enterface'05 database. The horizontal axis in this figure is the number of iteration steps, the vertical axis in Fig. 8 is the recognition accuracy and in Fig. 9, the loss error. According to the figure, the program stopped after about 130 apks.

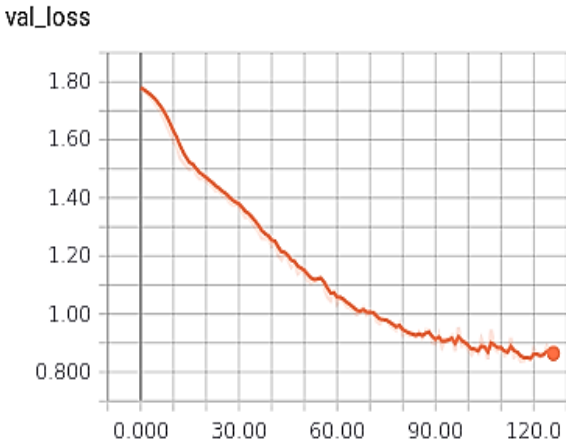

Fig. 8. The loss error of the validation data set, for the number of different iterations in the training phase

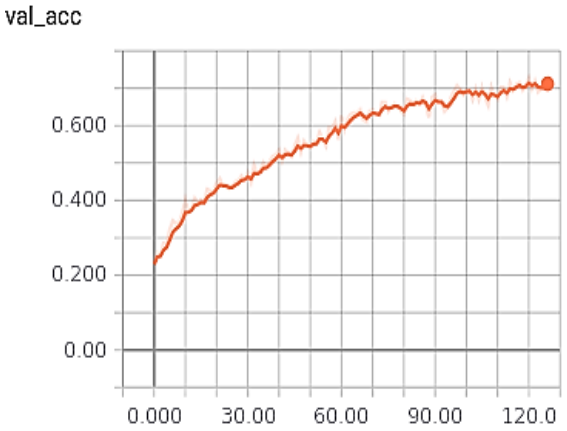

Fig. 9. Recognition accuracy of the validation data set, for the number of different iterations in the training phase

4.2.2 Analyzing the Results of Verbal Emotional Expression

In the part of verbal expression recognition, the Mel spectrogram function follows the general algorithm to calculate the Mel spectrogram, which is explained below. In this algorithm, the audio input is first buffered in frames of the number of samples (windows). Frames are overlapped by the number of overlap length samples. The specified window is applied to each frame, and then the frame is converted to a frequency domain representation with the number of FFTLength points. The frequency domain representation can be specified in magnitude or power by Spectrum Type. If Window Normalization is adjusted to true, the spectrum is normalized by the window. Each frame of the frequency domain display passes through a Mel filter bank. The output spectral values of the Mel filter bank are summed, and then the channels are concatenated so that each frame becomes a column vector of NumBands-Element numbers. MFCC has been used to extract features in verbal expression.

The set of extracted verbal features is given to the CNN network to select features, and the output of this network finally indicates the basic emotion class of each verbal file sample. In this experiment, the size of each batch is 32 and the number of iteration steps is 500, the learning rate is 0.001 and the momentum value is 0.0001. In Table 5, the confusion matrix of the presented model for verbal emotion recognition, for 6 basic emotion classes, is shown on the enterface'05 database.

Table 5. Confusion matrix of verbal emotion recognition, using CNN, for 6 basic emotion classes on the enterface'05 database

	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	82%	7%	0%	3%	3%	5%
Disgust	12%	55%	10%	7%	7%	10%
Fear	2%	5%	60%	14%	14%	5%
Happiness	12%	2%	5%	69%	2%	10%
Sadness	5%	5%	10%	0%	71%	10%
Surprise	5%	10%	2%	15%	5%	63%

The horizontal axis in this figure is the number of iteration steps, the vertical axis in Fig. 10 is the recognition accuracy and in Fig. 10, the loss error. According to the figure, the program stopped after about 170 apps (Because from this apps onwards, the recognition accuracy of the validation data will no longer increase)

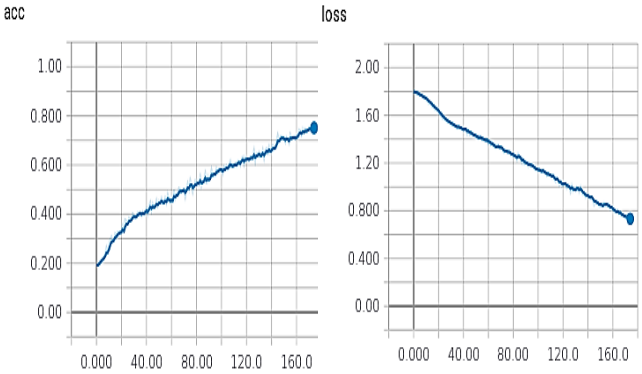

Fig. 10. Recognition accuracy and loss error, for different numbers of iterations in the training phase

4.2.3 Analyzing the Results of Facial and Verbal Emotional Expressions

Obtained features from facial expressions and verbal emotion have been used for fusion. Accordingly, 12 sequences of face images are given to the 3D-CNN network, and after applying different layers, a 256-character feature vector is obtained in the last fully

connected layer. Moreover, the verbal Mel-spectrogram images, in the size of 96x300x1 are given to the CNN neural network, and after applying different layers, a 256 feature vector is obtained from speech. Then, these features are combined and fed into the Soar classification model. This network, after feature fusion, results in the output of the final emotion class.

Feature fusion is the process of combining two feature vectors to obtain a single feature vector, which is more distinct than any of the input feature vectors. The architecture of the proposed emotion recognition model based on the fusion of two verbal-visual processes is shown in Fig. 1. According to this structure, information is obtained from two different paths. One is visual processing, in which after preprocessing, the facial images frames enter the visual cortex of the brain in the thalamus, and through PCA and then the 3D convolutional neural network, the best features of the face images are obtained and the other is audial processing, which after pre-processing operations such as framing and windowing, is converted into Mel-Spectrogram coefficients, and entered into CNN, to select the best features of the signals.

In order to address the advantage of the proposed facial expression recognition model, the efficiency of this model was compared with other methods that used the enterface'05 database. According to Table 6, the results of the proposed model are better than KNN and SVM. Even in order to examine the performance of the proposed model in the experiments, the feature selection part with 3D-CNN is also omitted in the calculations. Due to examining the recognition rate in Table 6, it can be seen that the proposed model provides higher recognition accuracy than other methods.

Table 6. Comparison of facial expression recognition accuracy, between different methods and the proposed model, for 6 classes on the enterface'05 database

Method	Accuracy
PCA+SVM	%64
My Method Without 3D-CNN	%65.5
PCA+KNN	%66.3
My Method	%73

The accuracy of emotion recognition in the proposed model, compared to other models in this field, was investigated and evaluated by researchers, and the results are shown in Table 7.

Table 7. Comparison of the effectiveness of the proposed model in bimodal emotion recognition with other methods performed on the enterface'05 database

Reference	Accuracy
Savestani et al. [28]	70.1%
Mansoorizadeh et al. [10]	71%
Bejani et al. [15]	77.7%
Zhalehpour et al. [24]	77.02%
My Method	81.7%

In order to show the improvement of the algorithms used in this thesis, Table 8 shows the recognition accuracy of all algorithms for different emotional expressions.

Table 8. Comparison of the recognition accuracy of the proposed models for different emotional expressions on the enterface'05 database

Suggested models	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Total Accuracy
Verbal emotional expressions	82%	55%	60%	69%	71%	63%	66.7%
Facial emotional expressions	47%	86%	55%	67%	67%	44%	62%
Fusion of features	85%	88.1%	73.8%	78.6%	81%	78%	81.7%

According to Table 8, the highest recognition rate is related to disgust and anger with a rate of 88.1 and 85%, and the lowest recognition rate is related to fear and surprise with a rate of 73.8 and 78%.

4.2.4 Results Analyze by Selecting the Feature

According to the proposed method, after the stages of feature extraction from facial expressions, using PCA, and from verbal expressions, using MFCC, a feature selection step using CNN and 3DCNN, respectively for verbal and facial expressions has been used, which then, fusion and final classification will take place on it; Table 9 shows the confusion matrix of emotion recognition based on this model, with feature selection.

Table 9. The confusion matrix of emotion recognition, on the basis of the fusion of verbal and visual features, using the Soar model by selecting the feature.

	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	85%	7.1%	2.3%	2.3%	0%	2.3%
Disgust	2.3%	88.1%	2.3%	2.3%	4.7%	0%
Fear	2.3%	0%	73.8%	7.1%	14%	2.3%
Happiness	7.1%	2.3%	4.7%	78.6%	2.3%	4.7%
Sadness	2.3%	0%	9.5%	0%	81%	9.5%
Surprise	2.3%	0%	4.7%	7.1%	7.1%	78%

According to Table 9, the recognition rate related to disgust and anger is 88.1%, and 85%, respectively, which is one of the highest emotion recognition rates, and on the other hand, the recognition rate of fear and surprise is recorded at 73.8% and 78%, respectively, which was the lowest recognition rate.

The emotion recognition accuracy for all classes, using this proposed model, was 81.7% on average. The interesting point is that according to Table 9, in the corresponding enterface'05 database, the emotions "disgust" and "anger" have the highest recognition accuracy, and surprise and fear have the lowest recognition accuracy, with less than 80%; while, the accuracy of other emotions is more than 80%. Moreover, the overlap of fear and sadness is more so that 14% of fear samples are in sadness. In addition, according to the table recognizing "disgust" with 90% accuracy is easier than other emotions.

4.2.5 Results Analyze without Selecting the Feature

According to the proposed method, after the stages of feature extraction from facial expressions, using PCA, and from verbal expressions, using MFCC, a feature selection step using CNN and 3DCNN, respectively for verbal and facial expressions has been used, which then, fusion and final classification will take place on it; Table 10 shows the confusion matrix of emotion recognition based on this model, without feature selection, and accordingly, after feature extraction, there are fusion and classification.

Table 10. The confusion matrix of emotion recognition, on the basis of the fusion of verbal and visual features, using the Soar model without selecting the feature.

	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	82%	8.1%	2.3%	3.3%	0%	3.3%
Disgust	3.3%	86.1%	2.3%	3.3%	4.7%	0%
Fear	3.3%	0%	70.8%	8.1%	14%	3.3%
Happiness	7.1%	3.3%	4.7%	76.6%	3.3%	4.7%
Sadness	3.3%	0%	10.5%	0%	78%	10.5%
Surprise	3.3%	0%	5.7%	7.1%	7.1%	76%

According to Table 10, the recognition rate related to disgust and anger is 86.1%, and 82%, respectively, which is one of the highest emotion recognition rates, and on the other hand, the recognition rate of fear and surprise is recorded at 70.8% and 76%, respectively, which was the lowest recognition rate.

The emotion recognition accuracy for all classes, using this proposed model, was 79.2% on average. The interesting point is that according to Table 7, in the corresponding enterface'05 database, the emotions "disgust" and "anger" have the highest recognition accuracy, and surprise and fear have the lowest recognition accuracy, with less than 80%; while, the accuracy of other emotions is more than 80%.

5- Conclusions

Emotion plays an effective role in humans' non-verbal communication with each other. Automatic emotion recognition can lead to the construction of computer systems that have the ability to understand human emotion and appropriate responses to it. In these computers, natural interaction is with humans. That is, the computer receives and interprets the emotion expressed in facial expressions, and gives an appropriate response based on its judgment. This article presented a method to recognize emotion from people's faces in video images, extract image features due to receiving all discriminating features from PCA, extract verbal features due to compression of the verbal resource and its wide range, and extract MFCC is used appropriately. Then, in order to select the best features in verbal signals and for facial images, CNN and 3DCNN were used, respectively. Next, feature level fusion was performed on the output of the previous stage to combine and achieve the appropriate shape of speech feature vectors and facial expressions. Finally, by using Soar classification, the recognition rate of audio-visual emotional expression was calculated. The highest rate of recognition based on audio-image was related to the emotional expression of disgust with a rate of 88.1%, and the lowest rate of recognition was related to fear with a rate of 73.8%. Moreover, in order to examine the performance of the proposed model, the tests were performed again without selecting the features, the highest recognition rate was related to disgust with a rate of 86.1%, and the lowest was related to fear with a rate of 70.8%. The proposed model has recorded an average improvement of 4 percent compared to the closest model. In addition, compared to the well-known SVM and KNN machine learning models, the results showed that the proposed method had a 6.7% improvement in recognition.

Among the future works, I can mention face recognition using head and body movements, which are involved in the expression of emotions in a special way. Also, in the continuation of this research, other modalities such as text, head and body movements, thermal and infrared images and other physiological features such as EEG can be added to it.

References

[1] Senthilkumar, N., S. Karpakam, M. Gayathri Devi, R. Balakumaresan, and P. Dhilipkumar. "Speech emotion recognition based on Bi-directional LSTM architecture and deep belief networks." Materials Today: Proceedings 57 (2022): 2180-2184.

[2] Crisp, Nicholas H., Peter CE Roberts, Sabrina Livadiotti, A. Macario Rojas, Vitor Toshiyuki Abrao Oiko, Steve Edmondson, S. J. Haigh et al. "In-orbit aerodynamic coefficient measurements using SOAR (Satellite for Orbital Aerodynamics Research)." Acta Astronautica 180 (2021): 85-99.

[3] Vallejo, Carlos, Jun Ho Jang, Carlo Finelli, Efreen Montaño Figueroa, Lalita Norasetthada, Rodrigo T. Calado, Mehmet Turgut et al. "Efficacy and Safety of Eltrombopag Combined with Cyclosporine As First-Line Therapy in Adults with Severe Acquired Aplastic Anemia: Results of the Interventional Phase 2 Single-Arm Soar Study." Blood 138 (2021): 2174.

[4] Whittaker, Jackie L., Linda K. Truong, Trish Silvester-Lee, Justin M. Losciale, Maxi Miciak, Andrea Pajkic, Christina Y. Le et al. "Feasibility of the SOAR (stop OsteoARthritis) program." Osteoarthritis and Cartilage Open 4, no. 1 (2022): 100239.

[5] Laird, John Edwin, Keegan R. Kinkade, Shiwali Mohan, and Joseph Z. Xu. "Cognitive robotics using the soar cognitive architecture." In Workshops at the twenty-sixth AAAI conference on artificial intelligence. 2012.

[6] Stavros, J., and G. Saint. "SOAR: Chapter 18: Linking strategy and OD to sustainable performance." WJ Rothwell, JM Stavros, R. Sullivan, and A. Sullivan, Practicing organization development: A guide for leading change. San Francisco, CA: Jossey-Bass (2010).

[7] Lucey, Simon, Ahmed Bilal Ashraf, and Jeffrey F. Cohn. Investigating spontaneous facial action recognition through aam representations of the face. Vol. 2. INTECH Open Access Publisher, 2007.

[8] Chang, Ya, Changbo Hu, Rogerio Feris, and Matthew Turk. "Manifold based analysis of facial expression." Image and Vision Computing 24, no. 6 (2006): 605-614.

[9] Pantic, Maja, and Ioannis Patras. "Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences." IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 36, no. 2 (2006): 433-449.

[10] Guo, Guodong, and Charles R. Dyer. "Learning from examples in the small sample case: face expression recognition." IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 35, no. 3 (2005): 477-488.

[11] Anderson, Keith, and Peter W. McOwan. "A real-time automated system for the recognition of human facial expressions." IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 36, no. 1 (2006): 96-105.

[12] Whitehill, Jacob, and Christian W. Omlin. "Haar features for FACS AU recognition." In 7th international conference on automatic face and gesture recognition (FGR06), pp. 5-pp. IEEE, 2006.

[13] Pantic, Maja, and Ioannis Patras. "Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences." IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 36, no. 2 (2006): 433-449.

[14] Zhao, Guoying, and Matti Pietikainen. "Dynamic texture recognition using local binary patterns with an application to facial expressions." IEEE transactions on pattern analysis and machine intelligence 29, no. 6 (2007): 915-928.

[15] Dhall, Abhinav, Akshay Asthana, Roland Goecke, and Tom Gedeon. "Emotion recognition using PHOG and LPQ features." In 2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG), pp. 878-883. IEEE, 2011.

[16] Black, Michael J., and Yaser Yacoob. "Recognizing facial expressions in image sequences using local parameterized models of image motion." International Journal of Computer Vision 25, no. 1 (1997): 23-48.

[17] Soleymani, Mohammad, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. "A survey of multimodal sentiment analysis." Image and Vision Computing 65 (2017): 3-14.

[18] Rosenbloom, Paul S., John E. Laird, Allen Newell, and Robert McCarl. "A preliminary analysis of the Soar architecture as a basis for general intelligence." Artificial Intelligence 47, no. 1-3 (1991): 289-325.

[19] Livanos, Nicole. "Mobility for Healthcare Professional Workforce Continues to Soar." Journal of Nursing Regulation 10, no. 4 (2020): 54-56.

[20] Ngai, Wang Kay, Haoran Xie, Di Zou, and Kee-Lee Chou. "Emotion recognition based on convolutional neural networks and heterogeneous bio-signal data sources." Information Fusion 77 (2022): 107-117.

[21] Lu, Cheng, Yuan Zong, Wenming Zheng, Yang Li, Chuangao Tang, and Björn W. Schuller. "Domain invariant feature learning for speaker-independent speech emotion recognition." IEEE/ACM Transactions on Audio, Speech, and Language Processing 30 (2022): 2217-2230.

[22] Cecchi, Ariel S. "Cognitive penetration of early vision in face perception." Consciousness and Cognition 63 (2018): 254-266.

[23] Torfi, Amirsina, Seyed Mehdi Iranmanesh, Nasser Nasrabadi, and Jeremy Dawson. "3d convolutional neural networks for cross audio-visual matching recognition." IEEE Access 5 (2017): 22081-22091.

[24] Wu, Xun, Wei-Long Zheng, Ziyi Li, and Bao-Liang Lu. "Investigating EEG-based functional connectivity patterns for multimodal emotion recognition." Journal of neural engineering 19, no. 1 (2022): 016012.

[25] Jin, Qin, and Junwei Liang. "Video description generation using audio and visual cues." In Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, pp. 239-242. 2016.

[26] Zhang, Shiqing, Shiliang Zhang, Tiejun Huang, Wen Gao, and Qi Tian. "Learning affective features with a hybrid deep model for audio–visual emotion recognition." IEEE Transactions on Circuits and Systems for Video Technology 28, no. 10 (2017): 3030-3043.

[27] Neverova, Natalia, Christian Wolf, Graham Taylor, and Florian Nebout. "Moddrop: adaptive multi-modal gesture recognition." IEEE Transactions on Pattern Analysis and Machine Intelligence 38, no. 8 (2015): 1692-1706.

[28] Koromilas, Panagiotis, and Theodoros Giannakopoulos. "Deep multimodal emotion recognition on human speech: A review." Applied Sciences 11, no. 17 (2021): 7962.

[29] Pantic, Maja, and Leon JM Rothkrantz. "Facial action recognition for facial expression analysis from static face images." IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 34, no. 3 (2004): 1449-1461.

[30] Martin, Olivier, Irene Kotsia, Benoit Macq, and Ioannis Pitas. "The eNTERFACE'05 audio-visual emotion database." In 22nd International Conference on Data Engineering Workshops (ICDEW'06), pp. 8-8. IEEE, 2006.

[31] Farhoudi, Zeinab, and Saeed Setayeshi. "Fusion of deep learning features with mixture of brain emotional learning for audio-visual emotion recognition." Speech Communication 127 (2021): 92-103.

[32] Bloch, Isabelle. "Information combination operators for data fusion: A comparative review with classification." IEEE Transactions on systems, man, and cybernetics-Part A: systems and humans 26, no. 1 (1996): 52-67.

[33] Zhang, Yongmian, and Qiang Ji. "Active and dynamic information fusion for facial expression understanding from image sequences." IEEE Transactions on pattern analysis and machine intelligence 27, no. 5 (2005): 699-714.

[34] James, Alex Pappachen, and Belur V. Dasarathy. "Medical image fusion: A survey of the state of the art." Information fusion 19 (2014): 4-19.

[35] Chen, JunKai, Zenghai Chen, Zheru Chi, and Hong Fu. "Emotion recognition in the wild with feature fusion and multiple kernel learning." In Proceedings of the 16th International Conference on Multimodal Interaction, pp. 508-513. 2014.

[36] Teissier, Pascal, Jordi Robert-Ribes, J-L. Schwartz, and Anne Guérin-Dugué. "Comparing models for audiovisual fusion in a noisy-vowel recognition task." IEEE Transactions on Speech and Audio Processing 7, no. 6 (1999): 629-642.

[37] Kumari, Jyoti, Reghunadhan Rajesh, and K. M. Pooja. "Facial expression recognition: A survey." Procedia computer science 58 (2015): 486-491.

[38] Ekman, Paul, and Wallace V. Friesen. "Facial action coding system." Environmental Psychology & Nonverbal Behavior (1978).

[39] Nazari, Elham, Rizwana Biviji, Amir Hossein Farzin, Parnian Asgari, and Hamed Tabesh. "Advantages and challenges of information fusion technique for big data analysis: proposed framework." Journal of Biostatistics and Epidemiology 7, no. 2 (2021): 189-216.

[40] Su, Yangfeng, Jian Wang, Xuan Zeng, Zhaojun Bai, Charles Chiang, and Dian Zhou. "SAPOR: Second-order Arnoldi method for passive order reduction of RCS circuits." In IEEE/ACM International Conference on Computer Aided Design, 2004. ICCAD-2004., pp. 74-79. IEEE, 2004.

[41] Sadeghi, Hamid, Abolghasem-Asadollah Raie, and Mohammad-Reza Mohammadi. "Facial expression recognition using texture description of displacement image." Journal of Information Systems and Telecommunication (JIST) 2, no. 4 (2014): 205-212.

[42] Nikpoor, Mohsen, Mohammad Reza Karami-Mollaei, and Reza Ghaderi. "A new Sparse Coding Approach for Human Face and Action Recognition." Journal of Information Systems and Telecommunication (JIST) 1, no. 17 (2017): 1.

[43] Navraan, Mina, Nasrollah Moghadam Charkari, and Muharram Mansoorizadeh. "Automatic Facial Emotion Recognition Method Based on Eye Region Changes." Journal of Information Systems and Telecommunication (JIST) 4, no. 4 (2016): 221-231.

[44] Motamed, Sara, Saeed Setayeshi, Azam Rabiee, and Arash Sharifi. "Speech emotion recognition based on fusion method." Journal of Information Systems and Telecommunication 5 (2017): 50-56.