Indonesian Journal of Electrical Engineering and Informatics (IJEEI)
Vol. 10, No. 3, September 2022, pp. 684~697
ISSN: 2089-3272, DOI: 10.52549/ijeei.v10i3.3879

On the Audio-Visual Emotion Recognition using Convolutional Neural Networks and Extreme Learning Machine

Arselan Ashraf[1], Teddy Surya Gunawan[2], Fatchul Arifin[3], Mira Kartiwi[4], Ali Sophian[5], Mohamed Hadi Habaebi[6]
[1,2,6]Department of Electrical and Computer Engineering, International Islamic University Malaysia, Malaysia
[3]Department of Electronic and Informatics Engineering, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia
[4]Department of Information Systems, International Islamic University Malaysia, Malaysia
[5]Department of Mechatronics Engineering, International Islamic University Malaysia, Malaysia

Article Info

Article history:

Received May 15, 2022
Revised Sep 18, 2022
Accepted Sep 19, 2022

Keyword:

Artificial Intelligence
Convolutional Neural Networks
Emotion Recognition
Human-Computer Interaction
Machine Learning

ABSTRACT

The advances in artificial intelligence and machine learning concerning emotion recognition have been enormous and in previously inconceivable ways. Inspired by the promising evolution in human-computer interaction, this paper is based on developing a multimodal emotion recognition system. This research encompasses two modalities as input, namely speech and video. In the proposed model, the input video samples are subjected to image pre-processing and image frames are obtained. The signal is pre-processed and transformed into the frequency domain for the audio input. The aim is to obtain Mel-spectrogram, which is processed further as images. Convolutional neural networks are used for training and feature extraction for both audio and video with different configurations. The fusion of outputs from two CNNs is done using two extreme learning machines. For classification, the proposed system incorporates a support vector machine. The model is evaluated using three databases, namely eNTERFACE, RML, and SAVEE. For the eNTERFACE dataset, the accuracy obtained without and with augmentation was 87.2% and 94.91%, respectively. The RML dataset yielded an accuracy of 98.5%, and for the SAVEE dataset, the accuracy reached 97.77%. Results achieved from this research are an illustration of the fruitful exploration and effectiveness of the proposed system.

Corresponding Author:

Teddy Surya Gunawan,
Department of Electrical and Computer Engineering,
International Islamic University Malaysia,
53100 Kuala Lumpur, Malaysia.
Email: tsgunawan@iium.edu.my

1. INTRODUCTION

Over the past decade, recognizing emotion has acquired expanding interest as it possesses a significant part in creating models proficient for perceiving, understanding, communicating, and responding to emotions [1, 2]. Furthermore, it signifies a key to improving the interactive experience of human-computer interaction applications in various proficient fields. It can be performed by employing a multimodal emotion recognition technique, as intrinsic human-human communication does. As a result, there is a need to investigate multimodal techniques to determine which channels provide valuable data for automatic emotion recognition. Human feeling can be conveyed through various biological and audio-visual channels, for example, voice inflection, facial appearances, temperature, body motions, pupil dilations, cerebrum signals, and pulse. Identifying human sentiments is among the significant components of empowering robots to interact with people. The perceived emotion will be considered for deciding the appropriate automated response [3]. The subsequent investigation has various applications in computing, mechanical technology, healthcare, gaming, security [4], and many

more. In fact, a few issues can influence the execution of a computer vision method. For instance, various subjects may express a similar feeling non-indistinguishably. Moreover, various viewpoints bring about inconsistent portrayals of the feeling. Besides, the presence of impediments and changes may delude the recognition technique. If the feeling perceived is dependent on voice, the surrounding commotion and the contrasts between voices of various subjects are enormous components that may influence recognition performance. To precisely perceive feelings, humans utilize both audio and visual signs.

Emotion recognition research can now benefit from deep learning techniques like Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and so on, owing to the increasing computational power of computers. Multi-modular emotion recognition research is separated into feature layer fusion and decision layer fusion based on the phase of consolidating multi-model data. Both fusion methods make use of deep learning algorithms. Additionally, decision layer fusion does not necessitate an exact timing match between the voice signal and the face appearance signal, making it easier to address feature reliability concerns. According to [7], people use coverbal signs to highlight their speech inference, which incorporates body, finger, arm, head, and facial appearances like gaze, movements, and speech metrics. About 93% of human interaction is performed through nonverbal methods, including facial appearances, body movements, and voice tone. Automated Facial Expression Recognition (FER) comprises face discovery, feature extraction, and expression recognition [8]. Face detection from the input video or image data plays an important part in the Facial Expression Recognition. Once the face is identified from the input data samples, it is subjected to various image processing steps. The next important step is extracting the facial information to distinguish the considered emotion. The features can be extracted from the entire face or specific facial regions like eyes, nose, eyebrows, etc.

Feature extraction can be done using texture filters like the Gabor filter. With the progress in the sphere of machine learning and image processing, modern models use deep learning techniques like convolutional neural networks for both feature extraction and classification [9]. In the case of speech emotion recognition (SER), it is essential to have good audio features. So, the process starts with pre-processing the input speech signal and then feature extraction. The feature extraction methods for speech signals are Mel-Frequency Cepstral Coefficients (MFCC), Mel-Spectrogram, and Teager Energy Operator (TEO). The deep learning classifier then trains the model over these features and results in recognition of speech emotions. For the audio-visual emotion recognition model, the feature data from both modalities need to be fused either before classification or at the phase decision level [10].

Many researchers have designed deep learning-based facial emotion recognition models due to the huge success of deep learning and in particular convolutional neural networks for image classification and other machine vision challenges [11,12]. Figure 1 from LENS.ORG shows a graphical analysis of scholarly works done in the sphere of visual emotion recognition.

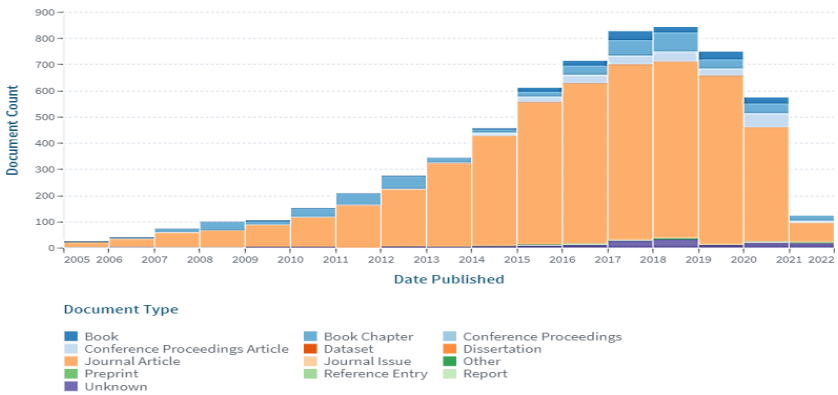


Figure 1. Scholarly works in the field of visual emotion recognition.

It is demonstrated that CNNs can recognize emotions with great precision and reach state-of-the-art results. Deep learning was used to construct a model of facial expressions for stylized animated characters in [13]. They trained a network to imitate the expression of the human face, animated face, and conversion of photographs of humans into animated visuals. In [14], they introduced a FER neural network with two convolution layers, one max-pooling layer, and four "inception" layers, or sub-networks. According to [9], CNN and 2-Cross validation technique was used to construct a visual-based emotion detection model that achieved 80 % and 83.33 % accuracy on the validation sets 1 and 2, respectively. In [15], they combined CNN with ImageNet transfer learning to discern emotions from static photos. On the ER sub-challenge dataset of static facial expressions, the authors attained 55.6 %. According to [16], they proposed an image-based emotion

recognition system based on LBP, GMM, and SVM. Using the CK+ database, the system attained an accuracy rate of 99.9% [17]. Based on the work of [18], using the eNTERFACE database, an IDP, and an extreme learning machine-based emotion recognition system from images achieved an accuracy of 84.12%.

Due to the importance of SER in developing human-computer interaction and artificial intelligence systems, it has become the subject of many recent types of research and surveys. Figure 2 from LENS.ORG shows various scholarly works done in the sphere of speech-based emotion recognition.

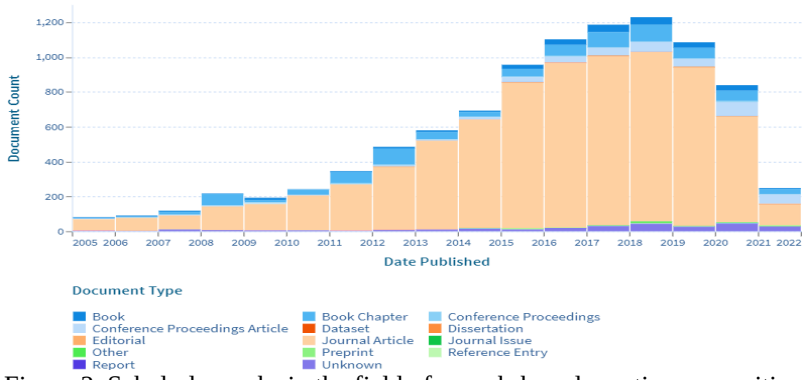


Figure 2. Scholarly works in the field of speech-based emotion recognition

According to [19], the development of the SER system from 2000 to 2017 was analyzed from three perspectives: database, feature extraction, and classifier. The content of the research comprises database and feature extraction, although only traditional machine learning approaches are evaluated as classification tools. In [20], a year later, a review was conducted on the discrete method in SER using deep learning. Deep Neural Networks (DNN), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Autoencoders are the deep learning methods discussed in this study, along with some of their limitations and functions. Based on [21], a brief review of the value of speech emotion features and data sets, noise reduction, and the importance of various classification algorithms (such as SVM and HMM) is carried out. In [22], a deep neural network (DNN) is used to determine the emotional probability of each portion of the speech. These probabilities are utilized to generate speech-level characteristics sent to the extreme learning machine classifier. The method used the interactive emotional binary motion capture database [23]. This strategy yielded an accuracy rate of 54.3%. In [24], a DBN and linear regression method was employed to detect musical emotion. The technique yielded a 5.41 % error rate in a Mood Swings Lite music database. CASESD was used to investigate DBNs and the SVM in [25]. The SVM had an accuracy of 84.54%, whereas the DBNs had a 94.6% accuracy. In [26], the authors developed a DL system in the form of CNN. The input to the system was the spectrogram of voice sound, fed into convolutional neural networks (CNNs). For the IEMOCAP database, they had a 64.78% accuracy rate.

Multimodal emotion recognition combines typically video and audio [27] because they are received in a non-invasive approach and are more expressive. Figure 3 from LENS.ORG shows various scholarly works done in the sphere of audio-visual-based emotion recognition.

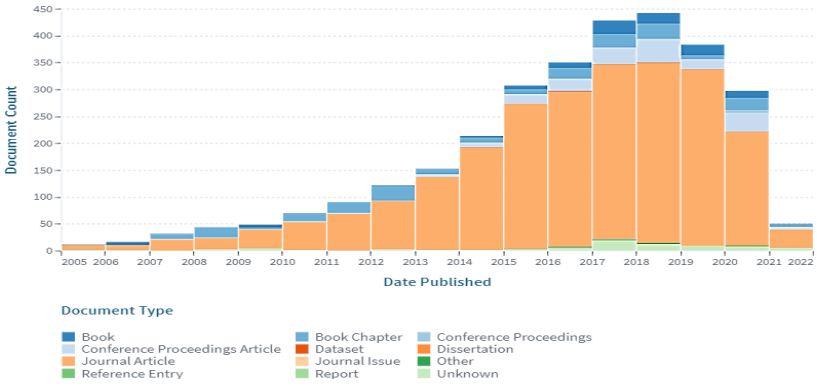


Figure 3. Scholarly works in the field of audio-visual-based emotion recognition

According to [28], they used an advanced hybrid deep learning-based CNN to extract audio records and a 3D-CNN for visual features, in addition, to a deep belief network for fusion and an SVM for classification. To detect emotions, [29] utilized an audio-visual database. Video and audio data were fed into CNNs and DBNs and were trained to recognize emotions. The precision for their version reached 47.67 %. Using two databases, the authors achieved 57.9% and 59.1% accuracy rates. According to [30], multidirectional regression and ridgelet transform skills were used to construct an audio-visual emotion identification model. The records were categorized using the ELM. They achieved an accuracy rate of 83.06 % with this method. Emotion recognition uses rhythmic and layout features in audio and visual data as defined in the model [31]. Using the eNTERFACE database, the technique had a 77% accuracy rate. In [32], an emotion recognition model was proposed that relied on audio and face data. A 3D DBN version of the classifier was employed. The eNTERFACE database had a 66.54 % accuracy rate on this model. In [33], audio features from speech signals were fused at the decision level with dense features from CNN-based features from image frames to recognize emotions. EmotiW 2015 database attained an accuracy of 54.55 %, and the CK+ database got 98.47 %, respectively. Using pre-trained models, an emotion recognition system was developed in [34]. The audio signal was fed into a CNN using a Mel-spectrogram, while the video signal was fed through a 3D CNN via face frames.

This research offers an audio-visual emotion recognition modal that employs two deep networks for feature extraction and fusion. These two networks ensure that the characteristics are fused with fine nonlinearity. A support vector machine is used to complete the final classification. This paper's contributions are as follows: the proposed model has been trained using a large amount of data with the help of data augmentations; feature extraction has been more robust with the inclusion of Convolutional Neural Networks; the usage of interlaced derivative pattern images and the local binary pattern picture in the three-dimensional CNN, along with the typical gray intensity image of keyframes; this gives the CNN different informative patterns of keyframes for feature extraction; utilization of Extreme Learning Machine (ELM) results in a promising increase in the accuracy of the modal. The following is a breakdown of the paper's structure. The second section contains the proposed methodology for the development of an audio-visual emotion recognition system. The experimental data, results, and benchmarking are presented in Section 3. Finally, section 4 concludes the paper.

2. RESEARCH METHODOLOGY

Based on the preceding recent related study, it is concluded that the previous systems were insufficient in terms of accuracy. As a result, we propose an emotion recognition system that will work effectively with data and its augmentation. Figure 4 depicts the proposed emotion recognition system's overall block diagram. Speech and video are the two types of input to the system. Before categorization, speech and visual signals are analyzed independently and then merged at a later step. Each of these modalities has two fundamental processes before fusion which are pre-processing and CNN-based deep networks

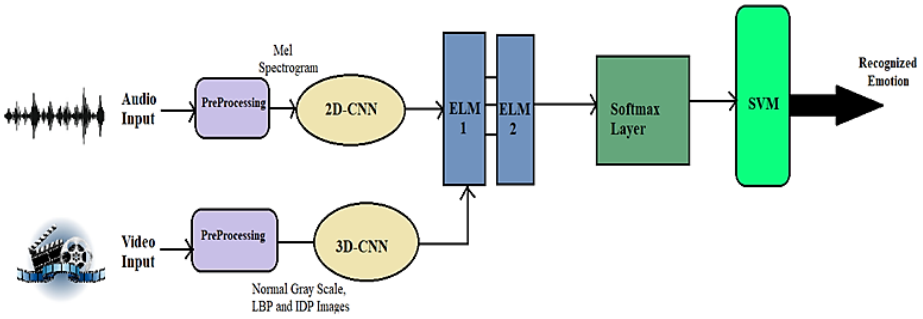


Figure 4. Block diagram of the proposed model

As seen in Figure 4, the input, i.e., audio and video samples, are subjected to pre-processing. In the case of pre-processing, various techniques are implemented, including division and overlapping, multiplication of frames, Fourier transforms, and much more for the audio input and face detection, histogram equalization, image cropping, resizing, frame selection, and extraction for the video input. The output is then fed to the respective CNNs, i.e., 2D-CNN for audio and 3D-CNN for video. Two Extreme Learning Machines follow it. Their result is transferred to the Softmax layer, and finally, SVM is used for the classification.

2.1. Input Video Pre-processing

In Figure 5, the pre-processing steps done on the video input for the proposed system are presented.

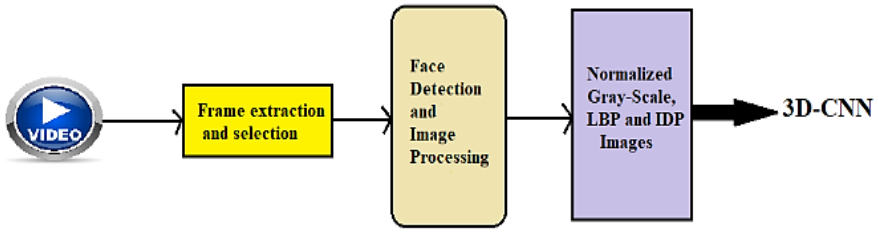


Figure 5. Video Pre-processing block diagram

The initial step is to choose a few crucial frames from a 2-second video clip. Figure 6 depicts the critical frame selection process. First, calculate the histograms of the image frames acquired from video in a window containing 2k+1 frames, k=3. Next, the difference between the subsequent frame histograms is calculated using the chi-square distance as shown in Eq. (1). In that sequence, the frame having the slightest difference is selected as the mainframe. Viola-Jones face detection algorithm [35] is used to clip the face region before generating the histograms. The histograms are created by cropping the facial photos. If no face was found in a frame, it was skipped over for further processing. Once the keyframe has been picked, the frame is turned to a grayscale image. The image is subjected to mean normalization. The grayscale image is also used to calculate the LBP and IDP images. As a result, each keyframe yields three images LBP, grayscale, and IDP.

$$X^2 = \frac{1}{2}\sum_{i=1}^{n}\frac{(x_i - y_i)^2}{(x_i + y_i)^2} \quad (1)$$

where x_i and y_i, are the subsequent histogram frames.

Four frames relocate the window after the keyframe is detected, and another keyframe is chosen. The method is continued till the video section is completed. Sixteen keyframes for CNN are selected every 2 seconds of the video segment. The images acquired from mainframes are subjected to sampling at a resolution of 225 × 225 pixels.

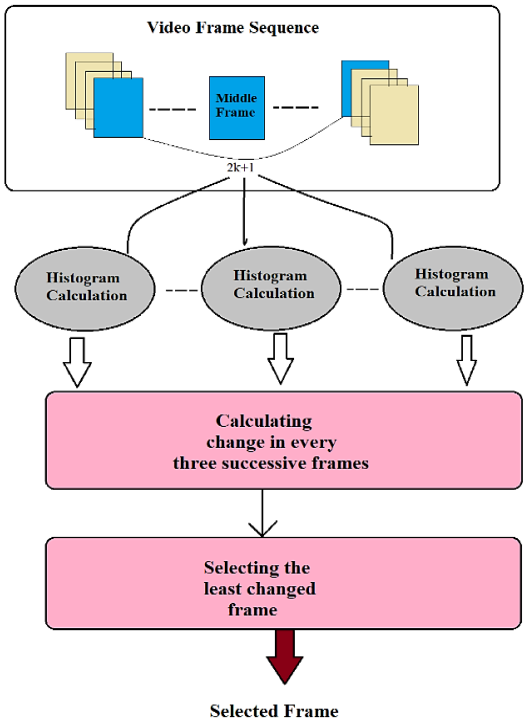


Figure 6. Flow chart for selecting video frames

2.2. Input Audio Pre-processing

For the audio pre-processing, the first thing to obtain is a Mel-spectrogram from the voice signal. It can be obtained by following the methods outlined as follows:

- Divide the signal into 40-millisecond frames, with 50 percent overlap between each frame.
- Use a Hamming window to multiply the frames.
- Convert the time-domain segment to the frequency-domain segment using the Fourier transform on a windowed frame.
- Filter the frequency-domain signal using 30 band-pass filters.
- Suppress the dynamic range by using the logarithm function on the filter outputs.
- Arrange the yields from the previous stages by framing to make the prompt for Mel-spectrogram.

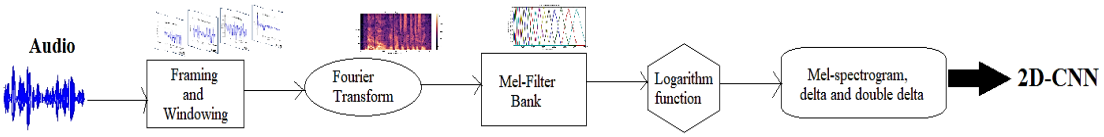


Figure 7. Audio pre-processing block diagram

Figure 7 depicts the proposed system's speech signal pre-processing processes. The Mel spectrogram is CNN's input. As the audio signal is of two seconds duration, we process the signal. As a result, the Mel-spectrogram has a size of 30 × 100. Conventional discourse attributes can achieve great acknowledgment execution with loud audio data; however, noisy information is often neglected. Deep learning models then remove highlights with a severe level of nonlinearity and encode signal vacillations. As a result, the CNN models are used in the proposed system. Images are required as input for CNN models.

In most cases, images have three channels. We derive (delta) and (double delta) coefficients from the Mel-spectrogram with a window size of three to be consistent with this depiction. Subsequently, the Mel-spectrogram picture (in grayscale), its delta picture, and the two-fold/double delta picture are identical to the three channels. A discourse signal's delta and two-fold delta coefficients encode relative fleeting data.

2.3. CNN Architectures

CNN is a fantastic feature learning and extraction method since it learns the features of neighborhood and spatial surfaces by employing convolution and nonlinearity tasks [36]. The DCNN consolidates lower-level features to reflect more significant-level features. There are various DCNN models, every one of which is valuable somehow or another. CNN for the discourse input and the video input in this proposed framework are unique; in the case of video, a 3D CNN is used, and for the audio input, 2D CNN is utilized.

2.3.1. 3D-CNN for video

A pre-trained model for the 3D CNN is used, as shown in [37]. This model was created to recognize sports actions. Eventually, the model was applied in different visual processing operations, including recognizing video emotions. Eight convolution layers and five max-pooling layers make up the architectural model of this study. The final two layers, each with 4096 neurons, are fully connected. A layer of SoftMax follows the fully connected layers. The model is taken care of 16 key casings that have been scaled to 225 × 225 pixels. To utilize the pre-prepared 3D CNN model, all the convolution layer and pooling layer loads from the model in [38] are utilized. The softmax layer is then supplanted with the number of feeling classes in our framework. Following that, a new softmax layer to fine-tune the model and use a backpropagation approach to update all the weights. The configurational view of the proposed CNN architecture till 4 convolutional layers is shown in Figure 8.

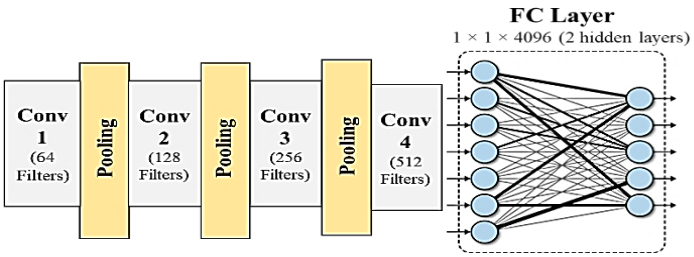


Figure 8. 3D-CNN Architectural diagram

2.3.2. 2D-CNN for audio

For the audio module, 2D-CNN is proposed for the development of an emotion recognition framework, as represented in Figure 9. There are three pooling layers and four convolution layers. A completely associated neural organization with two secret layers establishes the last layer. The yield of the total associated layer is given a softmax work. The softmax's yield is then gone through the ELM-based combination. There are 64 channels with a 7×7 size in the 2D CNN's convolution layer 1, 128 channels with a 7×7 size in the next layer, and 256 channels with a 3×3 size in the third layer. Eventually, in the fourth convolution layer, there are 512 channels, each sized 3×3. The max pooling is used in the first layer of the proposed architecture, whereas normal pooling is used in the following two layers. Each hidden layer in the FC networks has 4096 neurons. After the last yield layer, a softmax layer is utilized to create even dissemination for the yield esteems.

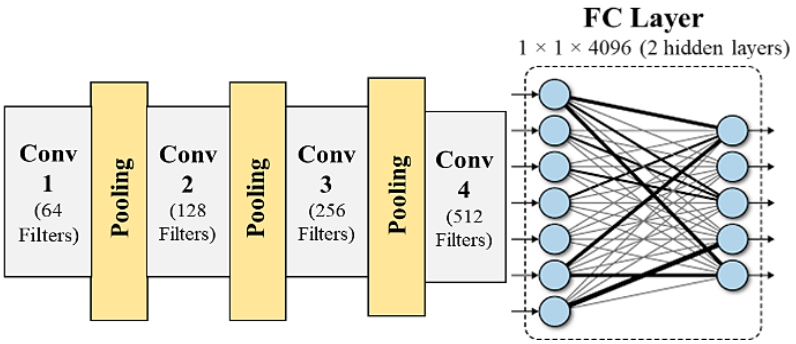


Figure 9. 2D-CNN Architectural diagram

2.3.3. fusion based on Extreme Learning Machine

Due to their widespread success, object recognition systems benefit from deep learning models, particularly convolutional neural networks (CNNs). Due to overfitting, the insufficiency of training data results in poor performance. The backpropagation approach necessitates many hyperparameter adjustments to effectively train a CNN, which makes it very slow. Extreme Learning Machine (ELM) is used to learn important CNN features and conduct quick and accurate classification to overcome these drawbacks. In [39], an efficient learning approach was proposed for single hidden layer feedforward neural network and extreme learning machine. The hidden nodes' input weights are generated randomly in ELM, and the output weights of SLFN are produced using the pseudoinverse operation of the hidden layer output matrix. Compared to conventional CNN, ELM has many advantages, including eliminating the need for weight adjustments during training, faster learning, and elimination of overfitting.

Two ELMs consecutively for the mix of scores from audio and video portions are utilized in the proposed emotion recognition framework, as displayed in Figure 10. Except for the last yield layer (softmax), the yields of completely associated networks (FCs) are utilized as contributions to the primary ELM in the proposed strategy. To keep up with the organization's sparsity, the number of hubs in the ELM's secret layer is equivalent to the number of classes multiple times (50). The primary ELM is based on gender (2 classes), while the subsequent ELM is gender-dependent on feelings. ELM-1 includes 100 hidden layer neurons since it has two yield classes. We erase the ELM-1's yield layer and supplant it with the ELM-1's trained hidden layer as the input to ELM-2. ELM-2 constitutes 300 hidden layer neurons if the emotion classes present are six.

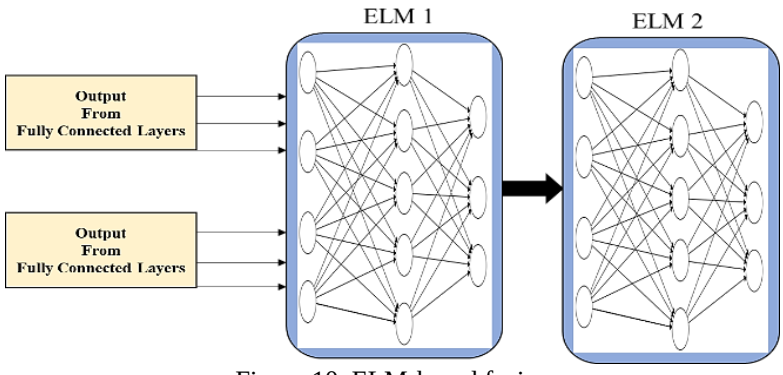


Figure 10. ELM-based fusion

If the ELM has L hidden nodes with an activation function ψ_q, input weight is represented by w_q, bias is represented by σ_q, and output weight by α_q, the following output function is acquired as in Eq. (2).

$$y_l(x) = \sum_q \alpha_q \psi_q(w_q \times x + \sigma_q); \quad q \in \{1, L\} \tag{2}$$

Eq. (3) and (4) are used to find the best output weights, where P denotes the number of samples for training. M represents the matrix (output) $[\varphi(x\_1), \varphi(x\_2), \ldots, \varphi(x\_p)]^T$, the identity matrix is represented by I, and $\varepsilon > 0$ denotes the regularisation coefficient. ELM's two layers give nonlinearity to the combination in a manner that rushes to figure yet is significant. It ought to be referenced that a combination dependent on deep organizations is already recorded in writing [29], yet this kind of combination is computationally rigorous, while our recommended strategy is not. The two-stage ELM achieves emotion recognition by relying upon gender orientation in a general sense, which upgrades accuracy.

$$\breve{\alpha} = \left[M^T M + \frac{I}{\varepsilon}\right]^{-1} M^T N; \quad P > L \tag{3}$$

$$\breve{\alpha} = M\left[M M^T + \frac{I}{\varepsilon}\right]^{-1} N; \quad P \leq L \tag{4}$$

2.4. Support Vector Machine Classifier

The softmax work is utilized to transform the ELM 2 yield scores into probabilities. The SVM gets the likelihood circulation of the ELM combination's yields as the contribution, as displayed in Figure 11. SVM changes the input proportion into an elevated proportion, permitting two classes' examples to be isolated by a straight plane. It should be underlined that SVM is utilized as the framework's classifier. On the contrary, CNN is used to extricate highlights from voice and video samples.

Moreover, extreme learning machines are utilized to intertwine the highlights. SVM is a paired classifier that utilizes a kernel capacity to extend input information into a high-dimensional space, isolating information from two classes by a hyperplane. The prime objective is to find the appropriate hyperplane with the most detachment from the vectors. In our framework, we utilize the SVM to make utilization of its incredible ability to arrange numerous sorts of information.

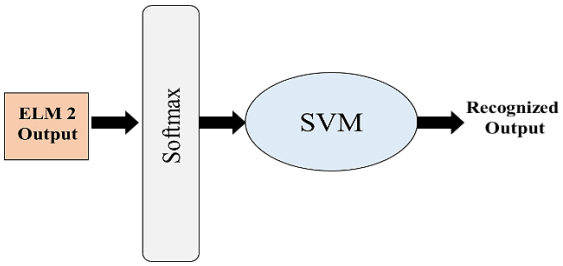


Figure 11. The flow of data to the SVM classifier

3. RESULTS AND DISCUSSION

A description of the databases utilized in this work and various experimental setups and results were presented in this section.

3.1. Audio-Visual Emotion Datasets

Three audio-visual datasets, in particular eNTERFACE'05, Ryerson Multimedia Lab (RML) [40], and Surrey Audio-Visual Expressed Emotion (SAVEE) [41], was utilized to assess the proposed approach to emotion recognition. eNTERFACE'05 data set contains tests from 42 individuals from different nations. They were all English speakers. Guys made up 81% of the members, while females made up 19%. The examples were recorded at 25 frames per second with a scaled-down DV computerized camcorder with an 800,000-pixel res. (FPS). Uncompressed sound system voice signals at a recurrence of 48,000 Hz in a 16-cycle design were recorded utilizing a high-quality receiver. The amplifier was around 30 cm underneath the subject's mouth and out of the camera's vision. A dark foundation was set behind the subjects to work with simple face recognition and tracking, as displayed in Figure 12.

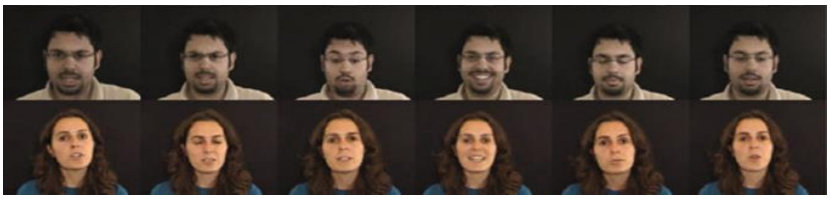


Figure 12. Sample images from the eNTERFACE'05 dataset

Each subject carried six diverse, passionate states, including happiness, disgust, fear, anger, surprise, and sadness. The Ryerson Multimedia Lab built the RML database. It has 720 audio-visual emotional expression samples—anger, contempt, fear, happiness, surprise, and sadness. The accounts were made with a computerized camera in a tranquil and brilliant climate with a basic foundation. The accounts included eight subjects who talked in various dialects, including English, Mandarin, Urdu, Punjabi, Persian, and Italian, in varied English and Chinese dialects. The examples were recorded at a recurrence of 22,050Hz utilizing 16-bit single-channel digitization. The casing rate was set to 30 frame-per-second. Each clip was recorded somewhere in the range of three and six seconds.

Figure 13. Sample images from the RML dataset

Figure 13 shows an example of photographs from the RML database. The SAVEE data set incorporates accounts of four males aged 27 to 31 who carried on six real feelings: anger, happiness, disgust, fear, sadness, and surprise, and also included neutral. The information base contains 480 local British English expressions, including 60 examples for each emotional state referenced. The members were recorded while passionate, and text prompts were streaked on a screen before them. Ten subjects rated the samples under auditory, visual, and audio-visual conditions. Figure 14 shows a selection of photographs from the SAVEE database.

Figure 14. Sample images from the SAVEE dataset

3.2. Experimental Setup

All calculations and computations, including network training and testing, were done on the laptop using the hardware combinations listed in Table 1.

Table 1. Hardware Specifications

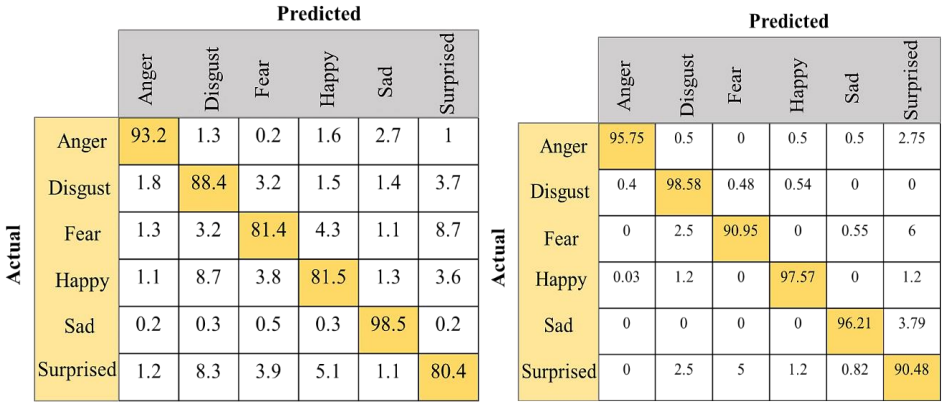
Computer	**HP PAVILION 15-BC408TX**
CPU	Intel Core i7-8750H (8th Gen)
RAM	8 GB DDR4 RAM
HDD	1TB
GPU	NVIDIA GeForce GTX 1050
Graphics Memory	4GB

The CNN models were prepared to utilize Stochastic gradient descent having batch size = 100, learning rate = 0.001, momentum = 0.9, and a weight decay =0.00005 as the preparation boundaries. The weights in the last layer were initialized utilizing Gaussian distribution with a mean equal to 0 and standard deviation equal to 0.01. As stated previously, weights of different layers have been taken from a pre-prepared

recognition system. During the training, there were 9000 iterations. The number of epochs used was 50. A 50% dropout was utilized in the last two completely linked layers to reduce overfitting. The samples in each dataset were divided into training and testing/validation, accounting for 70% and 30% of the total.

3.3. Experimental Results for the eNTERFACE'05 dataset

An accuracy of 87.2% was achieved for the eNTERFACE'05 database. eNTERFACE database includes 6 different sentences. As a result, the eNTERFACE database's accuracies were sentence independent. However, the results acquired from the eNTERFACE database were not highly encouraging. Therefore, with this dataset, we employed a four-fold cross-validation technique. For implementing augmentation, we rotated facial images at angles (4°, 8°, 16°, and 32°) and added white Gaussian noise to the audio signal at SNR = 15 dB, 20 dB, 25 dB, and 30 dB, respectively. By implementing augmentation, the highest accuracy acquired was 94.91% which was much more promising as compared to the earlier one without augmentation.



Actual \ Predicted	Anger	Disgust	Fear	Happy	Sad	Surprised
Anger	93.2	1.3	0.2	1.6	2.7	1
Disgust	1.8	88.4	3.2	1.5	1.4	3.7
Fear	1.3	3.2	81.4	4.3	1.1	8.7
Happy	1.1	8.7	3.8	81.5	1.3	3.6
Sad	0.2	0.3	0.5	0.3	98.5	0.2
Surprised	1.2	8.3	3.9	5.1	1.1	80.4

Actual \ Predicted	Anger	Disgust	Fear	Happy	Sad	Surprised
Anger	95.75	0.5	0	0.5	0.5	2.75
Disgust	0.4	98.58	0.48	0.54	0	0
Fear	0	2.5	90.95	0	0.55	6
Happy	0.03	1.2	0	97.57	0	1.2
Sad	0	0	0	0	96.21	3.79
Surprised	0	2.5	5	1.2	0.82	90.48

Figure 15. Confusion Matrix of eNTERFACE without augmentation and with augmentation respectively

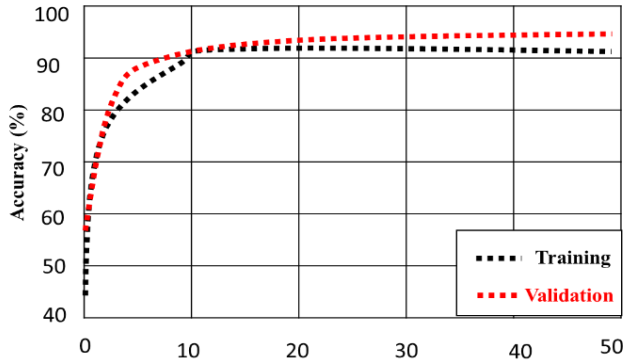


Figure 16. Plot for training and validation accuracy for augmented eNTERFACE database

Figure 15 depicts the system's confusion matrix utilizing the eNTERFACE database without and with augmentation. Accuracy is represented by the numbers (%). The diagonal dark-shadowed numbers represent individual emotion recognition accuracies. Lastly, Figure 17 displays the accuracies obtained from training and validation. The number of epochs used for the eNTERFACE database augmentation has also been shown. As seen, our system performs with better accuracy on the validation data.

3.4. Experimental Results for RML Dataset

An accuracy of 98.5% was obtained for the RML database. The database contains six emotions: fear, sadness, anger, disgust, surprise, and happiness. The fusion results were classified with promising output, depicting the proposed modal's better feasibility. Since the system's accuracy on the RML database was satisfactory, no augmentation was applied. The system's confusion matrix is depicted in Figure 17 using the RML database. The diagonal highlighted numbers show individual emotion recognition accuracies in (%).

Actual \ Predicted	Anger	Disgust	Fear	Happy	Sad	Surprised
Anger	98.33	0.67	1	0	0	0
Disgust	0	100	0	0	0	0
Fear	1	2	96.67	0	0	0.33
Happy	0	0	0	100	0	0
Sad	1	0	1	0	97.50	0.50
Surprised	0	0	0	1	0.67	98.33

Figure 17. Confusion Matrix on RML database

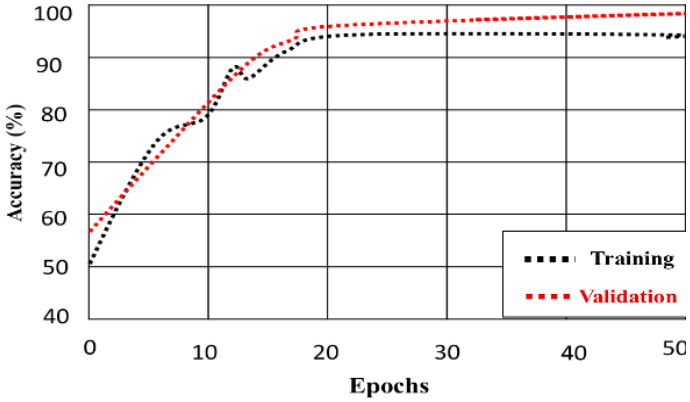


Figure 18. Training and Validation accuracy for RML database

Figure 18 illustrates the training and validation accuracies and the number of epochs utilized when working with the RML database. As can be observed, with the validation dataset, our system performs better. The findings of this study show that successful research in audio-visual emotion identification can be accomplished.

3.5. Experimental Results for SAVEE Dataset

SAVEE database attained an accuracy of 97.77 %. Sadness, surprise, anger, disgust, fear, happiness, and neutrality are the emotions recorded in the SAVEE database. The proposed modal is superior to the prior models, as per the extremely promising fusion results. Since the SAVEE database's accuracy was encouraging and was found adequate, therefore no augmentation was performed.

Actual \ Predicted	Anger	Disgust	Fear	Happy	Sad	Surprised
Anger	100	0	0	0	0	0
Disgust	5	90	1	0	4	0
Fear	0	0	96.67	0	1	2.3
Happy	0	0	0	100	0	0
Sad	0	0	0	0	100	0
Surprised	0	0	0	0	0	100

Figure 19. Confusion Matrix on SAVEE database.

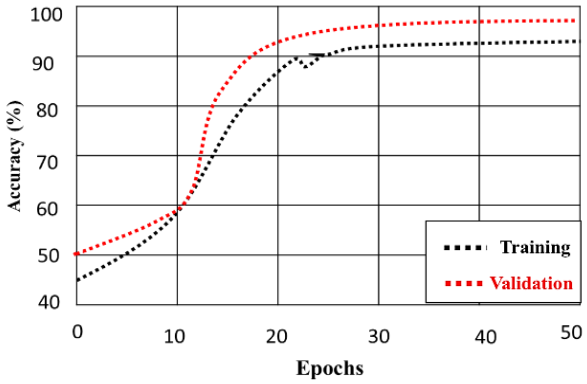


Figure 20. Training and Validation accuracy for SAVEE database

Figure 19 is a representation of the confusion matrix that was generated by the model utilizing the SAVEE database. The numbers indicate the accuracy (%). The numbers highlighted along the diagonal demonstrate the individual accuracy of each emotion. In addition, Figure 20 displays the accuracies obtained from training and validation. The number of epochs used for the SAVEE database has also been shown. As can be seen, our system performs better with the validation dataset. This study's findings demonstrate successful research in the sphere of audio-visual ER.

3.6. Benchmarking

This research's results are more promising than the benchmarked one, as shown in Table 2. This research has been directly benchmarked with Egils Avots et. al.'s work [41]. In addition to MFCC coefficients, benchmarked work used the most popular audio and spectral characteristics to characterize emotional speech. Using Viola–Jones face recognition and CNN (AlexNet) facial image emotion categorization, faces in keyframes are identified. Decision-level fusion is used for multimodal emotion recognition.

Table 2. Direct Benchmarking

Model	Technique used for Visual Input	Technique used for Audio Input	Fusion Technique	Comparison	Accuracy (%)		
Egils Avots et. al's model [14]	Viola-Jones and CNN (AlexNet)	Mel-frequency cepstral coefficients (MFCCs) and support vector machine (SVM)	Decision level fusion by transforming audio and video-based prediction output to a single normalized prediction of the sample (audio-video).	Direct comparison Using predicted labels	For eNTERFACE'05	50.2	
					For RML	69.3	
					For SAVEE	77.4	
Proposed Model	Viola-Jones and 3D-CNN	Mel-spectrogram and 2D-CNN	Decision level fusion by using two Extreme Learning Machine (ELM)	Using SVM	For eNTERFACE'05	Without Augmentation	87.2
						With Augmentation	94.91
					For RML	98.5	
					For SAVEE	97.77	

The research has also been indirectly benchmarked with other related works in terms of accuracy, as presented in Table 3. This proposed model outperformed the benchmarked research mostly due to the series of execution steps like the optimum algorithms and configurations opted, better image and signal processing techniques incorporated, and an optimum training environment.

Table 3. Indirect Benchmarking

S.No.	Model	Accuracy (%)
1.	Y. Kim et. al.'s model [42]	70-73
2.	M. Shamim Hossain and Ghulam Muhammad's model [30]	83
3.	M. Bejani et. al.'s model [31]	77

4. CONCLUSION

This paper proposed an audio-visual emotion recognition model. The deep learning technique was employed for the development of the modal. The audio signal was processed using a 2D-CNN, and the video was processed using a 3D-CNN. Audio and visual data were fused using an extreme learning machine. Two extreme learning machines were used for this work. This system incorporated three audio-visual databases: eNTERFACE, RML, and SAVEE. The data samples from each database were subjected to image and audio pre-processing. Since extreme learning machines add severe nonlinearity to the fusion, ELM-based combinations performed better. Results obtained from the proposed model were highly encouraging and premier. For the eNTERFACE dataset, the accuracy obtained without and with augmentation were 87.2% and 94.91%, respectively. The RML dataset yielded an accuracy of 98.5%, and for the SAVEE dataset, the accuracy came out to be 97.77%. Results achieved from this research illustrate fruitful exploration achievement in the field of visual emotion recognition. Future advancements in this subject appear to have much potential. More modalities can be taken into account in future inclusion strategies. The proposed technology can be implemented into any emotion-aware intelligent system for improved service to users or customers. Loads of the deep organization boundaries can be advantageously saved utilizing edge innovation for fast handling. The proposed framework can be evaluated using edge and distributed computing.

ACKNOWLEDGMENTS

The authors would also like to express utmost gratitude to the Kulliyyah of Engineering, International Islamic University Malaysia, for providing the KOE Postgraduate Tuition Fee Waiver Scheme and Universitas Negeri Yogyakarta for funding and facility for this research work.

REFERENCES

[1] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572-587, 2011.

[2] M. Morningstar, E. E. Nelson, and M. A. Dirks, "Maturation of Vocal Emotion Recognition: Insights from the Developmental and Neuroimaging Literature," *Neuroscience & Biobehavioral Reviews*, 2018.

[3] Fatemeh Noroozi, Marina Marjanovic, Angelina Njegus, Sergio Escalera, and Gholamreza Anbarjafari, "Audio-Visual Emotion Recognition in Video Clips," *IEEE Transactions On Affective Computing*, vol. 10, no. 1, pp. 60-75, 2019.

[4] Jean Phelipe de Oliveira Lima, & Carlos Maurí-cio Seródio Figueiredo. "A Temporal Fusion Approach for Video Classification with Convolutional and LSTM Neural Networks Applied to Violence Detection". *Inteligencia Artificial*, 24(67), 40–50. https://doi.org/10.4114/intartif.vol24iss67pp40-50, 2021.

[5] Poria, S., Cambria, E., Hussain, A., & Huang, G. B., "Towards an intelligent framework for multimodal affective data analysis," *Neural Networks*, vol. 63, pp. 104-116, 2015.

[6] Sahoo, S., & Routray, A., "Emotion recognition from audio-visual data using rule based decision level fusion," *Technology Symposium. IEEE.*, 2017.

[7] D. Bolinger and D. L. M. Bolinger, "Intonation and its Uses: Melody in Grammar and Discourse," *Stanford, CA: Stanford Univ. Press*, 1989.

[8] T. Wu, S. Fu, and G. Yang, "Survey of the facial expression recognition research," *in Proc. Int. Conf. Brain Inspired Cognitive Syst*, pp. 392-402, 2012.

[9] Arselan Ashraf, Teddy Surya Gunawan, Farah Diyana Abdul Rahman, Ali Sophian, Eliathamby Ambikairajah, Eko Ihsanto, Mira Kartiwi, "Affective Computing for Visual Emotion Recognition Using Convolutional Neural Networks," *Advances in Robotics, Automation and Data Analytics: Selected Papers from ICITES 2020*, vol. 1350, p. 11, 2021.

[10] M.Shamim Hossain, Ghulam Muhammad, "Emotion Recognition Using Deep Learning Approach from Audio-Visual Emotional Big Data,," *Information Fusion*, vol. 49, pp. 69-78, 2019.

[11] Kaiming, H.; Zhang, X.; Ren, S.; Sun, J, "Deep residual learning for image recognition," *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778, June 2016.

[12] Pellejero, N. F., Grinblat, G., & Uzal, L. “Semantic analysis on faces using deep neural networks: AnÃ¡lisis semÃ¡ntico en rostros utilizando redes neuronales profundas”. *Inteligencia Artificial*, 21(61), 14–29. https://doi.org/10.4114/intartif.vol21iss61pp14-29, 2018.

[13] Deepali, A.; Colburn, A.; Faigin, G.; Shapiro, L.; Mones, B., "Modeling stylized character expressions via deep learning," *In Asian Conference on Computer Vision; Springer: Cham*, pp. 136-153, 2016.

[14] Ali, M.; Chan, D.; Mahoor, M.H, "Going deeper in facial expression recognition using deep neural networks," *In Proceedings of the IEEE 2016 IEEEWinter Conference on Applications of Computer Vision (WACV)*, 2016.

[15] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep Learning for Emotion Recognition on Small Datasets using Transfer Learning," *Proc. the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15)*, pp. 443-449, 2015.

[16] G. Muhammad, M. Alsulaiman, S. U. Amin, A. Ghoneim, and M. F. Alhamid, "A Facial-Expression Monitoring System for Improved Healthcare in Smart Cities," *IEEE Access*, vol. 5, no. 1, pp. 10871-10881, 2017.

[17] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," *in proc. IEEE Int. Conf. Autom. Face Gesture Recognition*, pp. 46-53, 2000.

[18] G. Muhammad and M. F. Alhamid, "User Emotion Recognition from a Larger Pool of Social Network Data Using Active Learning," *Multimedia Tools and Applications*, vol. 76, no. 2, pp. 10881-10892, 2017.

[19] Swain, M.; Routray, A.; Kabisatpathy, P., "Databases, features and classifiers for speech emotion recognition: A review," *Int. J. Speech Technol*, vol. 21, pp. 93-120, 2018.

[20] Khalil, R.A.; Jones, E.; Babar, MI; Jan, T.; Zafar, M.H.; Alhussain, T, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," *IEEE Access*, vol. 7, pp. 117327-117345, 2019.

[21] Tripathi, A.; Singh, U.; Bansal, G.; Gupta, R.; Singh, A.K, "A Review on Emotion Detection and Classification using Speech," *In Proceedings of the International Conference on Innovative Computing and Communications (ICICC)*, 2020.

[22] K. Han, D. Yu, and I. Tashev, "Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine," *Proc. INTERSPEECH* , pp. 223-227, 2014.

[23] C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335-359, 2008.

[24] E. M. Schmidt and Y. E. Kim, "Learning emotion-based acoustic features with deep belief networks," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 65-68, 2011.

[25] Zhang, W., Zhao, D., Chai, Z., Yang, L. T., Liu, X., Gong, F., and Yang, S, "Deep learning and SVM-based emotion recognition from Chinese speech for smart affective service," *Softw. Pract. Exper.*, vol. 47, pp. 1127-1138, 2017.

[26] Haytham M. Fayek, Margaret Lech, Lawrence Cavedon, "Evaluating deep learning architectures for Speech Emotion Recognition," *Neural Networks*, vol. 92, pp. 60-68, 2017.

[27] Y. Ma, Y. Hao, M. Chen, J. Chen, P. Lu, and A. Košir, "Audio-Visual Emotion Fusion (AVEF): A Deep Efficient Weighted Approach," *Information Fusion*, vol. 46, pp. 184-192, 2019.

[28] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, "Learning Affective Features With a Hybrid Deep Model for Audio–Visual Emotion Recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 3030-3043, 2018.

[29] S. E. Kahou, X. Bouthillier, P. Lamblin, et al., "EmoNets: Multimodal deep learning approaches for emotion recognition in video," *Journal on Multimodal User Interfaces*, vol. 10, no. 2, pp. 99-111, 2016.

[30] M. Shamim Hossain and Ghulam Muhammad, "Audio-Visual Emotion Recognition using Multidirectional Regression and Ridgelet Transform," *Journal on Multimodal User Interfaces*, vol. 10, no. 4, pp. 325-333, 2016.

[31] M. Bejani, D. Gharavian, N. Charkari, "Audio-visual emotion recognition using ANOVA feature selection method and multiclassifier," *Neural Computing Appl.*, vol. 24, no. 2, pp. 399-412, 2014.

[32] D. Jiang, Y. Cui, X. Zhang, P. Fan, I. Gonzalez, H. Sahli, "Audio visual emotion recognition based on triple-stream dynamic bayesian network models," *ACII*, pp. 609-618, 2011.

[33] H. Kaya, F. Gürpınar, A. A. Salah, "Video-based emotion recognition in the wild using deep transfer learning and score fusion," *Image and Vision Computing*, vol. 65, pp. 66-75, 2017.

[34] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, "Learning Affective Features With a Hybrid Deep Model for Audio–Visual Emotion Recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 3030-3043, 2018.

[35] LeCun Y, Bengio Y, Hinton G, "Deep learning.," *Nature*, vol. 521, pp. 436-444, 2015.

[36] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," *IEEE International Conference on Computer Vision (ICCV)*, pp. 4489-4497, 2015.

[37] M. Chen, P. Zhou, and G. Fortino, "Emotion Communication System," *IEEE Access*, vol. 5, pp. 326-337, 2017.

[38] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 1, no. 3, pp. 489-501, 2006.

[39] Wang Y., Guan L, "Recognizing human emotional state from audio-visual signals," *IEEE Trans. Multimed.*, vol. 10, no. 5, pp. 936-946, 2008.

[40] S. Haq and P.J.B. Jackson, "Multimodal Emotion Recognition," *In W. Wang (ed), Machine Audition: Principles, Algorithms and Systems*, pp. 398-423, 2010.

[41] Avots, E., Sapiński, T., Bachmann, M. et al. Audio-visual emotion recognition in wild. Machine Vision and Applications 30, 975–985,2019. https://doi.org/10.1007/s00138-018-0960-9.

[42] Y. Kim, H. Lee and E. M. Provost, "Deep learning for robust feature generation in audio-visual emotion recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3687-3691, 2013.