International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Classroom Student Emotions Classification from Facial Expressions and Speech Signals using Deep Learning

Archana Sharma, Vibhakar Mansotra

Abstract*: Classroom environment is a competent platform for the students to learn and improve their understanding of the subject. An instructor's primary responsibility lies in managing the students in a way they feel interested and focused during the class. With the aid of automated systems based on artificial intelligence, an instructor can get feedback on the students' attention span in the class by monitoring their emotions using learning algorithms that can prove to be effective to improve the teaching style of the instructor that can in turn have positive effects on the class. In this paper, we propose an LSTM recurrent neural network trained on an emotional corpus database to extract the speech features and convolutional neural networks trained on the FER2013 facial emotion recognition database were used to predict the speech and facial emotions of the students respectively, in real-time. The live video and audio sequence of the students captured is fed to the learned model to classify the emotions individually. Once the emotions such as anger, sadness, happiness, surprise, fear, disgust and neutral were identified, a decision-making mechanism was used to analyze the predicted emotions and choose the overall group emotion by virtue of the highest peak value achieved. This research approach has the potential to be deployed in video conferences, online classes etc. This implementation proposal should effectively improve the classification accuracy and the relatability of the detected student emotions and facilitate in the design of sophisticated automated learning systems that can be a valuable tool in evaluating both the students and the instructors. The adapted research methodologies and their results are discussed and found to perform suggestively better than the other research works used in the comparison.

Keywords: **deep learning, emotion recognition, convolutional neural network, face recognition, speech emotion, recurrent neural network**

I. INTRODUCTION

Human interactions have evolved over the years through sharing of information and conveying emotions through verbal communications. With the current technological advancements, the communication between humans and machines have become the focal point of interest and observations. Machines, unlike humans, lack the feelings and empathy to read the emotions or sentiments of the human beings with whom they interact with. However, the manual analysis of human emotions can be tedious task to perform and with the availability of humungous amount of valuable data to examine the behavior of the human behavior and sentiments, it has become practically impossible to work on those data and try to make any meaningful sense of it by working on it manually. This demands the need to use machines to take over the role of the humans in evaluating the emotion data and classify the types of emotions to understand the mental state of a human being.

As emotions convey the psychological condition of the human mind, analyzing them through various means like facial appearances, vocal recordings and written thoughts can prove beneficial in understanding them. This phenomenon is presently utilized by the researchers to develop systems that can classify human sentiments that can aid them in venturing into newer research domains of human psychoanalysis, security and health-care applications. Most of the automatic recognition of human sentiments use facial expressions as the most popular way of interpreting emotions, which formed the basis for our previous research work [1]. There are sophisticated methods that use two-dimensional facial features that can narrow down the unique facial points and look for deformity in those facial points to determine the emotions. Even though it has proven to be effective in classifying emotions, there are certain downsides to this scheme. Since the expressions of the face are captured by a video camera before they can be analyzed frame by frame for those facial points that can be the precursor for the determination of emotions, the posture of the person like the prominence of the face in the frame, the angle at which it faces the camera, and how well the face movement is captured also becomes significant in the final classification accuracy achieved for the emotions. This has prompted us to build on our previous research work by including the speech features in addition to the facial features, thus integrating the video and audio channels to improve the performance of the emotion detection system. Despite the latest advancements in the field of facial and audio emotion identification, learning representations for natural speech segments and facial image extractions poses a substantial challenge in deploying them in noisy, unrestrained environments. Using neural networks to mimic the human brain's learning capacity of identifying faces and recognizing voices can work well in this context.

Manuscript published on 30 September 2019
* Correspondence Author
Dr. Archana Sharma*, Department of Computer Science, Government M.A.M College, Cluster University of Jammu, Jammu, India. Email: archana.35188@yahoo.com
Dr. Vibhakar Mansotra, Department of Computer Science and IT, University of Jammu, Jammu, India. Email: vibhakar20@yahoo.co.in

Since any neural network must be trained with the proper dataset comprising the facial expressions and speech audios of the humans, acquiring those huge datasets with the labelled emotions can be difficult sometimes.

First, by labelling the naturalistic speech segments and facial images were extremely difficult and mostly in the case of large datasets, For speech emotion corpus database [2], most datasets consists of different audio files recorded by different speakers and every emotion is labelled in a sequence of time and for facial image database (FER2013) [3], the images and labels represent the emotions of different people. Secondly, the labeled datasets can deteriorate from misplaced annotations that needs proper revision. Training those datasets on supervised techniques like Convolutional and Recurrent Neural Networks (CNN, RNN) can prominently improve the classification accuracy a great deal.

Hence considering all these challenges, we ought to think whether to learn a representation of the emotional facial and speech content from an unlabeled audio-visual database such as eNTERFACE [4]. As we can see with the recent emergence of large-scale audio-visual datasets of human voice and face, it is easier to find those unlabeled voice and face datasets from many research institutions and university repositories. In our research work, we have focused our attention to proposing a technique that adapts the visual information along with the recorded audio signals from the unlabeled and live-recorded video signal data. We try to drive at the conjecture that there is a direct correlation between the facial expressions of the speaker and his/her speech signals. We propose this new approach in a classroom arrangement to determine the emotional state of the students when listening to a lecture, from their facial and voice modulations. This can help us in designing intelligent systems that can read the psychological condition of the students better and for instructors, this can act like a supervising tool to help them in reframing their lecturing skills.

II. Related Works

Multimodal emotion which includes facial and speech classification is being used in a vast number of applications. There are many ways of approach to build this application. Here we shall see some of the related works where multimodal based emotions were used. At first, we have considered some related works based on the face and speech emotions individually and later the multimodal emotion systems.

Chao ma et.al [5] proposed a facial emotion recognition system in an online learning environment. This study captured a student facial image through webcam, judges the student learning emotions and provide feedback to the lecturer. This is a one-to-one correspondence i.e. a student having a class with lecturer. A convolutional neural network model was used to train and identify the facial expression of the student. The main drawback of this system is that it cannot recognize multiple persons at the same time.

In another work, Wang et.al. [6] proposed the Fourier parameter model which was used to identify speaker-independent speech emotion recognition system. Mel-Spectrum cepstral coefficients [7] and Fourier parameter feature extraction [8] techniques were used to extract the features from the audio files and Support vector machine [9] and Bayesian Classifiers [10] were used to identify the emotions of a subject by using two different databases such as a German database and a Chinese elderly emotion database. They had achieved moderate success with their results, and it would have been better, had neural networks been used.

Another research work was proposed by Sara Zhalehpour et.al [11] for audio-visual emotion recognition. In this study, three different methods such as Maximum Dissimilarity-based (MAX-DIST) [12], Clustering-based (DEND_CLUSTER) [13] and Emotion Intensity [14] based methods were used to extract the facial features and identify the faces from a selected peak video frame. Mel-frequency Cepstral coefficients (MFCC) [7] and Relative Spectral Features (RASTA) [15] techniques were used to extract the features of audio. eNTERFACE [4] and Baum-la were the two different audio-visual databases used to train the Support Vector Machine [9] that predicts the emotion of a subject. This model, however, does not allow multiple persons to be captured in a video or recorded in an audio which could be considered as a serious drawback.

Samuel Albanie et.al [16] had proposed an audio-visual emotion recognition system that used a cross-modal distillation process [17] to exploit the relationship between visual and audio emotions. This model is related in some way to our proposed method like, a teacher to train a single student and teacher should be able to identify the emotions of a student from audio and video. A convolutional neural network model was used to train on the voxceleb [18] database. This model also represents a single person emotion recognition.

A recent work by Stavros Petridis [19], had proposed an audio-visual speech recognition model that classifies the emotion of a student from audio and video signals. It is basically a speech recognition model, but along with speech, they had used a visual frame of face while speaking. It identifies the lip reading of a person with an audio. A lip-reading database was used to train the convolutional neural network and LSTM recurrent neural networks.

In a different research approach, Chao et.al. [20], had proposed an emotion identification system with a temporal alignment and perception attention. Temporal alignment was used for feature-level fusion mechanism and perception alignment was used to better prediction. An LSTM-RNN was used to train EmotiW2015 database for audio and visual based features. Principle component analysis was used extract the features of vocal data. To improve the recognition rate of vocal and visual data, decision-level model was used on this dataset. Satisfactory result was achieved in this study with the perception attention technique and decision level model.

III. Existing System Frameworks

In this paper, we have intentions to review on those research works, and the system frameworks implemented done thus far in the field of speech and facial emotion recognition in a classroom environment. For this purpose of investigation, we have selected four previous research works that focus on the analysis on student emotions in a classroom setting.

The selection of these papers were done based on the following criteria: year of publishing of the research work; we have primarily considered the most recent research paper, the methodology used in the work; we have selected papers that have used different methodologies for features extraction and classification of emotions, and finally the physical design and implementation; how the hardware setup consisting of cameras, microphones and the processing unit have been implemented in a classroom setup, the student population and the computing power of the processing unit used in facial and speech image processing. For our review, we have considered Egils et al. [21] proposal on audio-visual emotion recognition system, where there is no involvement of classroom environment. Another work, we have taken for our review was originally presented by Shiqing Zhang et al. [22] who used a hybrid deep learning model for identifying the emotions from speech and facial expressions. They had used the 3D features based deep learning model with three publicly available databases such as eNTERFACE, RML and BAUM-1s audio-visual dataset. Our third paper for the review was presented by Panagiotis Tzirakis et al. [23] for the design of an end-to-end deep learning model for emotional face and speech using CNN and RECOLA audio-visual dataset of the AVEC2016 research challenge in 2016. The final research work considered here was by Seng et al. [24]. A thorough search for the literature that combines both facial and speech emotions in a classroom environment did not yield any results based on the search keywords we used and we can safely assume there is not any comprehensive implementation of one such system in a classroom setting which we have considered as the focus of our study and proposal.

A. *Dataset*

The research work [21] presented a cross-corpus estimation using three databases such as eNTERFACE [4], SAVEE and RML. These three databases contained the video sequences with the speech audio signals of the speakers. It has 1166 video sequences recorded from 42 subjects. Each subject was asked to listen to six different stories eliciting different emotions. SAVEE database has high quality video recordings captured from different speakers of British English accent under seven different emotion categories. This data has been validated by ten participants under audio, visual and audio-visual conditions. RML database is the best performance database in the recent advancement of vocal-visual emotion identification, that uses 720 vocal-visual samples which portrays seven different emotions. These audio-visual samples were captured in a closed room along with an electronic equipment such as digital video camera and high-quality microphone, with a modest background. Each subject performed six types of expressions via face and voice. This database is limited to specific language and culture .

The research work discussed in [22] had used three different databases such as eNTERFACE [4], RML and BAUM-1's. As in the above section, we have discussed about eINTERFACE and RML databases. Another database, BAUM-1 has 1222 video samples from thirty-one Turkish subjects to classify six different emotions as well as boredom and contempt. To obtain spontaneous audio-visual expressions, emotion elicitation by watching films is employed. The size of original video frames used was *720×576×3*.

Another work [23], had proposed a Remote Collaborative and Affective (RECOLA) [25] Emotion database to predict the spontaneous and natural emotions of a human. Four modalities were included in the corpus; audio, video, electro-cardiogram and electro-thermal activity. In total, 9.5 hours of multimodal recordings from forty-six French-speaking participants were recorded and annotated for five minutes each, performing a collaboration task during a video conference.

The final research work [24], had evaluated the two most popular databases such as eINTERFACE and RML. We have already discussed on this in the previous section.

B. *Pre-processing of Facial Expressions Database*

In previous works, authors had used a frame extraction approach which means while recording, the continuous video must be divided into frames per second and apply pre-processing to that frame. For pre-processing, face detection and feature extraction are the most crucial stages for classifying emotions of a human.

B.1 Face Detection

In research works [21], [22] and [24], the authors have proposed the Voila-Jones [30] face detection algorithm to identify and crop out the faces from a respected video frame and save it to an image. The Voila-Jones algorithm returns a bounding box containing the region, which is cropped out from a video frame, resized and passed it to the feature extraction process.

In a prior work [23], they had not used any face detection algorithm because of limitations of datasets they have used in their study. Different feature extraction methods were directly applied to the video frames and extract the multiple features of a face and fine tuning it to identify the emotions. This type of method is only applicable in one-on-one conversational video conference.

B.2 Feature Extraction

In research work [21], the authors had not used any specific feature extraction method. Once the face is cropped and saved, the facial images are randomly translated in X and Y directions in the range of 30x30 pixels. Once the general features are identified, a Convolution neural network with AlexNet architecture in Fig 1 has taken those general facial pixel wise features and train it accordingly.

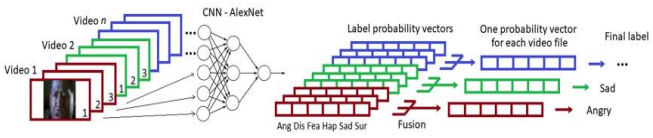

Fig. 1. AlexNet Architecture [20].

Zhang et al. in their work [22] did not propose any specific feature extraction method. The results of face detection methods can locate the center of two eyes in up-right face.

Then, it calculates the eye distance of facial images and normalize it to a fixed distance of 55 pixels. For a facial image, it is usually observed that its height is roughly three times longer than the eye distance, whereas its width is roughly twice. Consequently, based on the normalized eye distance, a resized RGB image of 150 x 150 pixels is finally cropped from each frame. To conduct a fine-tuning task, the cropped facial mage for each frame is resized to specified input of the proposed model.

In another work, Tzirakas et al. [23], had used pixel intensity from the cropped faces of a subject's video as an input to the Deep Residual Neural Network (DeepRNN) architecture of 50 layers. This network adopts the residual learning by stacking building blocks, where the input and output of the layer is the residual function to be learned and identify the mapping or a linear projection. A feature vector of bottleneck features will be gathered and applied to the deep residual architecture based convolutional neural network algorithm to identify the facial expressions of the students.

Seng et al. [24], in their groundbreaking work, had used Bi-Directional Principle Component analysis (BDPCA) [26] and Least-Square Linear Discriminant Analysis (LSLDA) [27] for dimensionality reduction and class discrimination. In their study, these two models were cascaded with the output of BDPCA being used as the input of LSLDA [27]. The extracted features are then forwarded to Convolutional Neural network.

C. *Pre-processing of Audio Emotions Database*

In most of the emotion classification using audio, especially speech signals, extracting of features must be the preliminary stage. Because, deep learning algorithms does not have an ability to learn from direct audio files. Pre-processing stage must be done on audio files in order to train it. In every research work that we have reviewed had involved a pre-processing stage.

Egils et al. in their research work [21], had proposed the "Mel-Frequency Cepstral Coefficients (MFCC), which are calculated for 400ms sliding window with size of 200ms". As we can understand from their work, every single audio file has several windows with MFCC and one feature vector. One feature vector consists of thirty-four parameters, where the first twenty-one represents the global audio features and the remaining thirteen coefficients represent the local MFCC. For MFCC feature extraction, we have used the following setup: pre-emphasis coefficient value of 0.97, twenty filter bank channels, thirteen cepstral coefficients, 300 Hz lower frequency limit and 3700 Hz upper frequency limit.

Whereas Zhang et al. [22] had used MFCC algorithm that extracts three channels of Mel-Spectrogram segments with size 64×64×3. By using a 25ms Hamming window and 25ms overlapping, the whole log Mel-Spectrogram with 64 Mel-filter banks from 20-8000Hz is obtained. Then, a context window of 64 frames is used to divide the whole log Mel-spectrogram into audio segments with size 64×64. A shift size of 30 frames is used during segmentation. Then, these audio features are given as an input to the neural network.

Tzirakas et al. [23], had combined both feature extraction and regression steps in one jointly trained model for predicting the emotions. Hence, they had segmented the raw waveform from audio files to 6-second-long sequences. At 16kHz sampling rate, time sequences correspond to a 96000-dimensional vector which is taken input to neural network.

Seng et al. [24] had proposed in their work, a Voice activity detector (VAD) [27] to pre-process the audio speech and eliminate the background noise and segment out the non-speech portions of the audio signal. The audio feature analyzer has been designed to extract features such as pitch, log-energy, TEO and ZCR from an audio signal. First, the speech signal is divided into certain number of frames by windowing and calculate the above features. Each feature represents group of emotions.

D. *Training and Feature Learning*

Egils et al. [21], had used a fusion based approach to train and predict the audio-visual emotion for a single frame and 400ms audio segments. To classify the emotions from an audio recording, Support Vector Machine (SVM) [9] algorithm was proposed and implemented. To classify the facial emotions, a convolutional neural network with AlexNet architecture was proposed and implemented. For testing, frame-based emotion prediction was transformed into a video-based prediction. For single prediction, each audio and video prediction have six score values which corresponds to predicted accuracy for a specific class. The sum of all probabilities will be unity and the highest probability is the predicted label.

Zhang et al. [22], had implemented a feature fusion-based approach, where the audio and visual networks were trained individually with a fine-tuning scheme and replaced their fully connected layers with two new fully connected layers with target audio-visual emotion categories in Fig 2.

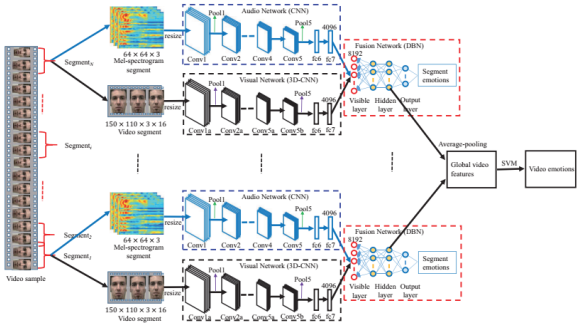

Fig. 2. Structure of the audio-visual emotion recognition model proposed by Zhang et al. [21].

The CNN and DBN neural networks have been implemented to train the audio and visual data. Once the fusion networks training is finished, a 2048-D joint feature representation can be computed and a linear SVM [9] was employed for emotion identification.

Tzirakas et al. [23], in their proposal of the classification model had designed a fine-tuned model with pre-trained ResNet-50 architecture implemented on the ImageNET dataset.

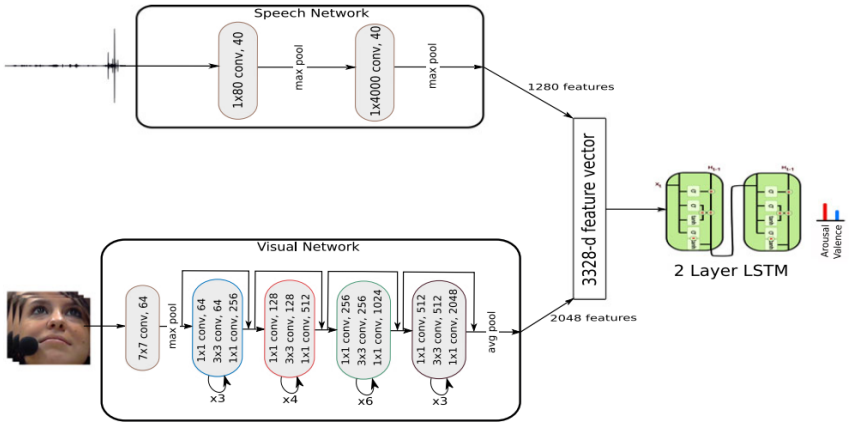

Fig. 3. Two-Layer LSTM Neural Network Model as proposed by Tzirakas et al. [22]

A two-layer LSTM neural network with 256 cells was implemented and trained the model for both audio and visual categories.

After training the audio-visual networks, LSTM layers were discarded and only the extracted features were considered as seen in Fig 3.

Seng et al. [24], had suggested a decision level mechanism which means, the audio and visual network were trained and predicted emotions individually and making the decisions based on the predicted emotion of both audio and visual networks. An optimized Kernel-Laplacian Radial Basis Function (OKL-RBF) [29] neural classification has used to classify emotions from facial expressions.

Feature level fusion mode in Fig 4. with combining both feature extraction and feature learning as we discussed in the feature extraction part.

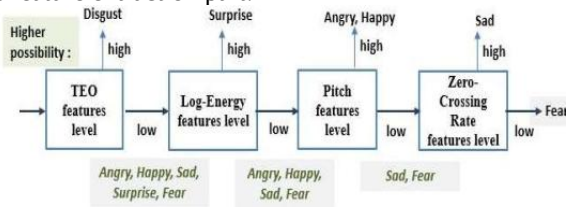

Fig. 4. Feature Level Fusion Model as proposed by Seng at al. [23]

IV. PROPOSED SYSTEM FRAMEWORK

The architecture of the proposed system implemented in a classroom environment as shown in Fig.5.

A. *Dataset Description*

As there are certain publicly available databases which contain basic human expressions and widely used in facial and speech emotion identification systems. There are different popular datasets available for multimodal systems such as audio and visual combination datasets. Though we have proposed an audio-visual emotion system, we do not use a combinational dataset that take the same video to extract the video and audio channels separately and analyze it, rather here we use different datasets for working on the facial and speech features. This may help us to have distinct conversations at a same time. In such a case, audio-visual datasets would not be enough for identifying the emotion and those datasets were mainly used for a one-on-one conversation where the speakers' faces were directly pointed towards the camera. Hence, the FER2013 [3] dataset and Emotional Corpus [1] databases were used in this study. FER2013 has 48x48 pixel size images of faces. The training set consists of 28,709 examples of human faces. This dataset predicts 7 types of facial expression (angry, sad, disgust, sad, happy, fear, surprise, neutral). Emotional Corpus [2] dataset has 14,503 speech utterances and it also predicts seven types emotion classes (angry, sad, disgust, sad, joy, fear, surprise, neutral). The main advantage of speech database is, it has group conversations. On all the many speech emotion datasets, this is only dataset which contains group conversations.

B. *Pre-Processing*

B.1 *Facial Emotion*

In facial emotional recognition, a pre-processing step is the crucial part to identify emotions of a human. It contains, face detection and feature extraction.

Face Detection

There are few most popular used for face detection which are Voila Jones [30], Haar Cascade and dlib face detector. We proposed a Voila Jones face detection algorithm to identify the emotions. As in case of multi-camera arrangement, it would be easy to identify the faces of the students in the classroom. Our proposed hardware assembly uses a single wide-angle and straight-ahead camera pointed towards the students. This algorithm crops the faces and resizes it to 150x150 pixel size images. The cropped faces will be highlighted by drawing a rectangular region around it. To remove the background and other edge-lying obscurities, the subject's face was must be cropped from the original image based on the positions of the eyes.

Feature Extraction

Feature extraction is an important stage in facial expression classification, as it is essential to reduce the time and storage space required. Feature extraction is process of dimensionality reduction which removes multi-collinearity and improves the interpretation of parameters of the neural network model. In our case, we have proposed a Haar Cascade feature extraction technique to analyze pixels in the images into squares by function.

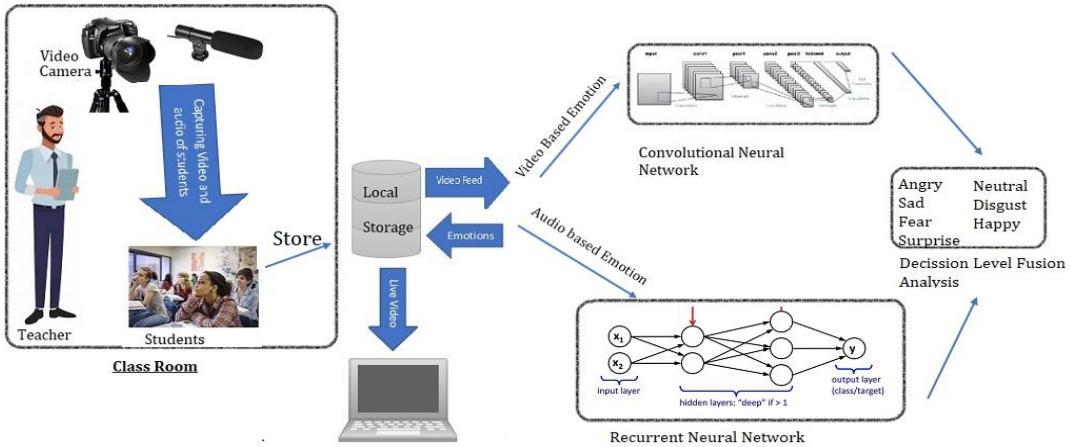

Fig. 5. Proposed CNN and RNN based Emotion Classification System Architecture.

After due to consideration of all the reviewed research works, this technique performed well in multimodal emotion recognition.

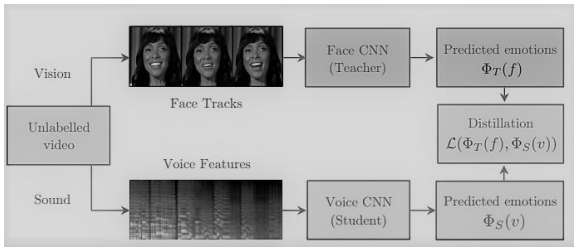

Fig. 6. Proposed Neural Network Model Flow Diagram

B.2 *Speech Emotion*

Feature extraction is the most crucial part to identify emotions based on speech. There are few popular algorithms for audio feature extraction such as Acoustics Feature Extraction, Pitch, Format Frequency and Mel Frequency Cepstral Coefficients (MFCC). Here, we have proposed the MFCC algorithm which is mainly based on the human peripheral auditory system. Mel is a unit to measure the frequency of a tone.

In this implementation, speech signal is divided into frames with the help of sliding window method. A Discrete Fourier Transform of frame signals will be considered to obtain the magnitude of spectrum. There are few steps in this algorithm to get the features of audio.

- Frame Blocking
- Windowing
- Applying FFT
- Mel-Frequency Rapping
- Cepstrum (Discrete Cosine Transform)

C. *Training and Feature Learning*

C.1 Facial Emotion Recognition

The recommended method here is to implement the CNN model for the recognition of facial emotions. For computer vision applications, convolutional neural networks are proved to better in recognition and classification-based tasks. Deep learning architectures do not learn directly from any kind of media data such as audio or images and it is the most common assumption. In the case of recognition, CNN algorithm first preprocess the data and transforms it into arrays or vectors of data and assign some labels to each vector of image which specifies the output.

Convolutional Networks (ConvNet) has basically consists of four types of layers; input, hidden layer, output and fully connected layers. Input layer consists of initial data such as input data, data size which are used to further processing of subsequent layers. Hidden layer consists of set weighted inputs produce an output through activation function. Output layers has produced outputs from hidden layers. Fully connected layer is the last layer of neural network and every layer in neural network is connected to fully connected layer and gathers all outputs and produce a single vector of values as final output. Each layer builds with specific layers namely Convolution (CONV), Normalization (NORM), Pooling layer (POOL) and Activation Function. Each layer does a specific job. Convolution layer consists of set of learnable filters. Normalization is used to accelerate the performance of training the neural network. Pooling layer is specifically designed to reduce the number of parameters and computation in the neural network. Max Pooling is the most widely used approach in CNN. Activation functions are complicated and introduce the non-linear mappings between input and output variables to our network. The pseudo code of the convolution neural network as shown below.

***Algorithm*:** *Convolutional Neural Network Model for training.*

Step-1: *Initialization*
Initialize the CNN parameters such as weights, number of batches, number of output features and epochs.

Step-2: *Input*
Load the FER 2013 dataset as training and test data.

Step-3: *Pre-processing*
Scale the training dataset images into 48 x 48 pixels.
Apply rotation, shifting and color modes if necessary.

Step-4: *Build the neural network model:*
Add 3 consecutive hidden layers and add fully connected layer with SoftMax function.
{C1→ C1-Norm→ReLU→Pooling}x 3 →{FC→Smx}

Step-5: *Train the model*

- Iterate the model by dividing the dataset as batches and apply CNN algorithm on each epoch.
- Update weights and bias in each iteration and calculate criterion loss and accuracy of the epoch.

Step-6: *Prediction and save the model.*

- Predict the outputs on test dataset and save the trained model for further predictions.
- Saved model will be stored in local drive.

C.2 Speech Emotion Recognition

Our idea is to implement a Recurrent Neural Network (RNN) [31] with Long- and Short-Term Architectures (LSTM) for training the speech data to classify the emotions.

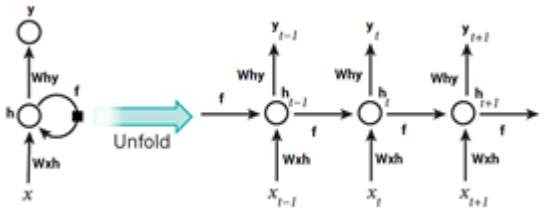

Fig. 7. A recurrent neural network and unfolding in time of the computation involved in its forward computation [30].

RNN is used in the form of sequential data information.

The other neural networks such as convolutional neural network and deep neural network assumes all input signal must be independent of each other. However, in many applications we usually treat all types of time-distributed signals. Therefore, these sequential data analysis approaches

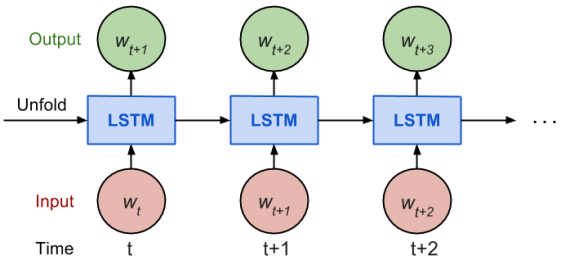

Fig. 8. RNN-LSTM Architecture

are widely used not only for language modelling but also for machine translation.

The LSTM cells are structures included in recurrent networks that preserve most relevant patterns from a chain of events in the input data. The implementation of RNN determines an output sequence through input sequence, and the structure of the LSTM cell allows to outperform the network by storing information from long-term context.

The LSTM network in Fig. 8 itself can decide what information can forget (through forget gate) and what information should remain (to update parameters). At this point, the output is computed considering the input after parameters updating. A gradient descendent algorithm was used as an optimization function which reduces the error LSTM cells in time sequences. This characteristic was used to learn temporal changes in the features extracted from the audio signals.

Algorithm: *Recurrent neural network*

Step-1: *Pre-Processing*

1. Feature Extraction:
   a. Frame Blocking
   b. Windowing
   c. Applying FFT
   d. Mel Frequency Wrapping
   e. Cepstrum
2. Convert Data into Tensors

Step-2: *Creating the Network.*

Add layers multiple RNN layers with sequential hidden layers.

Step-3: *Training the Network with the below parameters*

Num_epochs=100.
Optimization = Adam
Criterion = Cross Entropy

Step-4: *Learning decision*

Calculation of criterion loss on train and test datasets over the epochs.

Step-5: *Making Predictions*

Test on individual images.
Evaluate trained model on the test set.

V. RESULTS AND DISCUSSION

We have reviewed few prior research works done on visual and vocal emotion classification in a classroom environment and the comparable results are presented in Table. I.

TABLE I
RESULTS OF CLASSIFIED EMOTIONS ACROSS REVIEWED RESEARCH WORKS

Authors	Database	Result (%)
Egils Avots et al. [20]	SAVEE	94.33
	RML	60.20
	eNTERFACE	48.31
	AFEW	94.68
Shiqing Zhang et al. [21]	RML	80.36
	eNTERFACE	85.97
	BAUM-1	54.57
Panagiotis Tzirakis et al. [22]	RECOLA in Speech	71.5
	RECOLA in Facial	43.5
Seng et al. [23]	eNTERFACE	86.67
	RML	90.83

The authors in the reviewed research works have tried to approach the work using different neural network algorithms with mixed results. This is mainly due to the differences in the feature extraction schemes used in their work and the dataset used in their studies also had a huge impact on their results. Seng et al. [23] approach using Radial Basis Function Neural Networks yielded better results compared to other reviewed here as seen from the table above.

TABLE II
PROPOSED SYSTEM EMOTION CLASSIFICATION RESULTS

Image	Classification Result	Audio	Classification Result
	Happy: 95% Surprise: 4.1% Neutral: 0.9%		Happy: 80% Angry: 9.3% Disgust: 0.7%
	Happy: 96.6% Surprise: 3.4%		Surprise: 68.2% Happy: 21.1% Angry: 10.7%
	Sad: 86.4% Neutral: 9.7% Fear: 3.9%		Sad: 56.4% Neutral: 34.7% Fear: 10.9%

Our approach is unique in the way we use different datasets for video and audio feature extraction, and we have used RNN-LSTM network for emotion classification. This is targeted towards multiple students in a classroom setting and we hope to achieve comparative results with this proposed system architecture.

The expected results are believed to have better classification accuracy compared to the reviewed works and the predicted results are represented in the Table II.

VI. CONCLUSION

Emotion classification has been a major challenging aspect in the field of computer vision and artificial intelligence. In this paper, we have presented a detailed overview of few research works done in a field of multimodal (speech and visual) emotion recognition, as discussed from the comparisons of various datasets and methodology frameworks. Based on the assessment of other similar works and our proposed work, we have come to understand that every implementation in analyzing emotions has its own merits and demerits depending upon the methodologies they choose to use in their work and the databases available at their disposal. After considerable research on the techniques and methodologies applied thus far for emotion classification, we have proposed our paper with an amended technique that can compensate for the failings of the past frameworks. The preprocessing stage is vital in extracting the right set of features; hence we have opted for Voila-Jones proposed feature extraction technique for facial features and MFCC based feature extraction scheme for vocal features. Whereas for the emotion classification, neural network implementations such as RNN and CNN was used to learn from the datasets, and the trained network model was used to speed up the learning rate. This system with decent processing capability can categorize various sentiments in real-time environment compared to the other discussed works in this paper. LSTM based RNN was used to handle the issue with the model being sensitive to outliers. This model must function better in recognizing various faces and mingled voices even in a noisy environment where it is hard to conclusively identify the emotions. Even so, there can be some drawbacks to every approach, which can be the stepping stone that can lay the foundation for the future works that can make use of texts shared in microblogging sites like twitter to analyze the group emotions of the students that follow the school/college's official twitter handle. This, in conjunction, with speech and face based emotion recognition methods can be a potent tool to accurately narrow down the emotions of the classroom representing the definite emotion or sentiment displayed by the students during the lecture.

REFERENCES

1. A. Sharma and V. Mansotra, "Deep Learning based Student Emotion Recognition from Facial Expressions in Classrooms." *International Journal of Engineering and Advanced Technology (IJEAT)*, vol. 8, no. 6, pp. 4691-99. 2019.
2. Chen, Sheng-Yeh, C. Hsu, C. Kuo, and L. Ku, "Emotionlines: An emotion corpus of multi-party conversations." *arXiv preprint arXiv:1802.08379*, 2018.
3. Giannopoulos, Panagiotis, I. Perikos, and I. Hatzilygeroudis. "Deep learning approaches for facial emotion recognition: A case study on FER-2013." In *Advances in Hybridization of Intelligent Methods*, pp. 1-16. Springer, Cham, 2018.
4. Brad, Florin, R. Iacob, I. Hosu, and T. Rebedea. "Dataset for a neural natural language interface for databases (NNLIDB)." *arXiv preprint arXiv:1707.03172*, 2017.
5. C. Ma, C. Sun, D. Song, X. Li and H. Xu, "A deep learning approach for online learning emotion recognition," *13th International Conference on Computer Science & Education (ICCSE)*, Colombo, 2018, pp. 1-5
6. Wang, Kunxia, N. An, B. Nan Li, Y. Zhang, and L. Li. "Speech emotion recognition using Fourier parameters." *IEEE Transactions on Affective Computing* 6, no. 1, 2015: 69-75.
7. Bhadra, Sweta, U. Sharma, and A. Choudhury. "Study on feature extraction of speech emotion recognition." *ADBU Journal of Engineering Technology* 4, 2016.
8. Jing, Xiao-Yuan, H. Wong, and D. Zhang. "Face recognition based on discriminant fractional Fourier feature extraction." *Pattern Recognition Letters* 27, no. 13, 2006: 1465-1471.
9. Abdulrahman, Muzammil, and A. Eleyan. "Facial expression recognition using support vector machines." In *2015 23nd Signal Processing and Communications Applications Conference (SIU)*, pp. 276-279. IEEE, 2015.
10. Rish and Irina. "An empirical study of the naive Bayes classifier." In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, no. 22, pp. 41-46. 2001.
11. Zhalehpour, Sara, Z. Akhtar, and C.E. Erdem. "Multimodal emotion recognition based on peak frame selection from video." *Signal, Image and Video Processing* 10, no. 5 2016: 827-834.
12. Birchfield, Stan, and C. Tomasi. "A pixel dissimilarity measure that is insensitive to image sampling." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, no. 4 1998: 401-406.
13. Imani, Maryam, and H. Ghassemian. "Band clustering-based feature extraction for classification of hyperspectral images using limited training samples." *IEEE Geoscience and remote sensing letters* 11, no. 8 2013: 1325-1329.
14. Liu, Gang, Y. Lei, and J. HL Hansen. "A novel feature extraction strategy for multi-stream robust emotion identification." In *Eleventh Annual Conference of the International Speech Communication Association*. 2010.
15. Hermansky, Hynek, N. Morgan, A. Bayya, and P. Kohn. "RASTA-PLP speech analysis technique." In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 121-124. IEEE, 1992.
16. Albanie, Samuel, A. Nagrani, A. Vedaldi, and A. Zisserman. "Emotion recognition in speech using cross-modal transfer in the wild." *arXiv preprint arXiv:1808.05561*, 2018.
17. Gupta, Saurabh, J. Hoffman, and J. Malik. "Cross modal distillation for supervision transfer." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2827-2836. 2016.
18. Nagrani, Arsha, J.S. Chung, and A. Zisserman. "Voxceleb: a large-scale speaker identification dataset." *arXiv preprint arXiv:1706.08612*, 2017.

19. Petridis, Stavros, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic. "End-to-end audiovisual speech recognition." In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6548-6552. IEEE, 2018.
20. Chao, Linlin, J. Tao, M. Yang, Y. Li, and Z. Wen. "Audio visual emotion recognition with temporal alignment and perception attention." *arXiv preprint arXiv:1603.08321*, 2016.
21. Avots, Egils, T. Sapiński, M. Bachmann, and D. Kamińska. "Audiovisual emotion recognition in wild." *Machine Vision and Applications* 30, no. 5 2019: 975-985.
22. Zhang, Shiqing, 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* 28, no. 10 2017: 3030-3043.
23. Tzirakis, Panagiotis, G. Digeorgios, M.A. Nicolaou, B.W. Schuller, and S. Zafeiriou. "End-to-end multimodal emotion recognition using deep neural networks." *IEEE Journal of Selected Topics in Signal Processing* 11, no. 8 2017: 1301-1309.
24. Seng, K. Phooi, L. Ang, and C.S. Ooi. "A combined rule-based & machine learning audio-visual emotion recognition approach." *IEEE Transactions on Affective Computing* 9, no. 1 2016: 3-1.
25. Ringeval, Fabien, A. Sonderegger, J. Sauer, and D. Lalanne. "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions." In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pp. 1-8. IEEE, 2013.
26. Zuo, Wangmeng, D. Zhang, and K. Wang. "Bidirectional PCA with assembled matrix distance metric for image recognition." *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 36, no. 4 2006: 863-872.
27. Liu, Li-Ping, Y. Jiang, and Z. Zhou. "Least square incremental linear discriminant analysis." In *2009 Ninth IEEE International Conference on Data Mining*, pp. 298-306. IEEE, 2009.
28. Woo, Kyoung-Ho, T. Yang, K. Park, and C. Lee. "Robust voice activity detection algorithm for estimating noise spectrum." *Electronics Letters* 36, no. 2 (2000): 180-181.
29. Kaminski, Wladyslaw, and P. Strumillo. "Kernel orthonormalization in radial basis function neural networks." *IEEE Transactions on Neural Networks* 8, no. 5 1997: 1177-1183.
30. Jensen and O. Helvig. "Implementing the Viola-Jones face detection algorithm." Master's thesis, Technical University of Denmark, DTU, DK-2800 Kgs. Lyngby, Denmark, 2008.
31. Lim, Wootaek, D. Jang, and T. Lee. "Speech emotion recognition using convolutional and recurrent neural networks." In *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pp. 1-4. IEEE, 2016.

AUTHORS PROFILE

Dr. Archana Sharma received Doctorate from Jodhpur National University, Jodhpur, MCA from University of Jammu, India. She is currently working as Assistant Professor, Department of Computer Science, Govt.M.A.M College, Cluster University of Jammu. She has 12 years teaching experience at university of Jammu. Her research area is data mining and artificial intelligence. She has published several papers in National & International Journals.

Dr. Vibhakar Mansotra received Doctorate in computer science from University of Jammu, M.Sc., M.Phil., (Physics), PGDCA, M.Tech.(IIT-Delhi), India. He is currently working as Professor, Former Head of Department of Computer Science and IT, Dean, Faculty of Mathematical Science & Director, CITES&M, Coordinator IGNOU (S.C-1201), University of Jammu and Chairperson Division-IV, Computer Society of India. He has 26 years teaching experience at university of Jammu. His research area is data mining, software engineering, artificial intelligence and information retrieval. He has published several papers in National & International Journals.